# Novelty Assessment Report

**Paper**: On Entropy Control in LLM-RL Algorithms
**PDF URL**: https://openreview.net/pdf?id=LqazVN5epT
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-30

## Abstract

For RL algorithms, appropriate entropy control is crucial to their effectiveness. To control the policy entropy, a commonly used method is entropy regularization, which is adopted in various popular RL algorithms including PPO, SAC and A3C. Although entropy regularization proves effective in robotic and games RL conventionally, studies found that it gives weak to no gains in LLM-RL training. In this work, we study the issues of entropy bonus in LLM-RL setting. Specifically, we first argue that the conventional entropy regularization suffers from the LLM's extremely large response space and the sparsity of the optimal outputs. As a remedy, we propose AEnt, an entropy control method that utilizes a new clamped entropy bonus with an automatically adjusted coefficient. The clamped entropy is evaluated with the re-normalized policy defined on certain smaller token space, which encourages exploration within a more compact response set. In addition, the algorithm automatically adjusts entropy coefficient according to the clamped entropy value, effectively controlling the entropy-induced bias while leveraging the entropy's benefits. AEnt is tested in math-reasoning tasks under different base models and datasets, and it is observed that AEnt outperforms the baselines consistently across multiple benchmarks.



## Core Task Landscape

This paper addresses: **Entropy Control in Large Language Model Reinforcement Learning**
A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

### Taxonomy Overview

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

- **Entropy Regularization Methods and Mechanisms**
- **Exploration-Driven Approaches**
- **Token-Level and Sample-Level Analysis**
- **Entropy Collapse and Training Stability**
- **Policy Optimization Algorithms and Frameworks**
- **Application Domains and Specialized Settings**

### Complete Taxonomy Tree

- Entropy Control in Large Language Model Reinforcement Learning Survey Taxonomy
- Entropy Regularization Methods and Mechanisms
  - Adaptive and Dynamic Entropy Regularization ★ (4 papers)
  - [0] On Entropy Control in LLM-RL Algorithms (Anon et al., 2026) View paper
  - [26] EntroPIC: Towards Stable Long-Term Training of LLMs via Entropy Stabilization with Proportional-Integral Control (Kai Yang, 2025) View paper
  - [43] Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models (Fan Jiajun, 2025) View paper
  - [45] Rediscovering Entropy Regularization: Adaptive Coefficient Unlocks Its Potential for LLM Reinforcement Learning (Zhang XiaoYun, 2025) View paper
  - Fixed-Coefficient and Clipping-Based Methods (4 papers)
  - [6] Entropy Ratio Clipping as a Soft Global Constraint for Stable Reinforcement Learning (Zhenpeng Su, 2025) View paper
  - [14] Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models (Kim Junsu, 2025) View paper
  - [16] CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning (Su, 2025) View paper
  - [32] Rethinking entropy interventions in rlvr: An entropy change perspective (Wang Hong, 2025) View paper
  - KL-Divergence Regularization Design (3 papers)
  - [11] Entropy-regularized process reward model (Zhang Hanning, 2024) View paper
  - [20] On the design of kl-regularized policy gradient algorithms for llm reasoning (Zhang Yifan, 2025) View paper
  - [42] APO: Enhancing Reasoning Ability of MLLMs via Asymmetric Policy Optimization (Hong Min-jie, 2025) View paper
  - Entropy Minimization Approaches (2 papers)
  - [21] The unreasonable effectiveness of entropy minimization in llm reasoning (Agarwal, 2025) View paper
  - [23] Entropy Regularizing Activation: Boosting Continuous Control, Large Language Models, and Image Classification with Activation as Entropy Constraints (Liao, 2025) View paper
- Exploration-Driven Approaches
  - Curiosity and Intrinsic Motivation (2 papers)
  - [13] Controlling large language model agents with entropic activation steering (D'Oro, 2024) View paper
  - [15] Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models (Song, 2025) View paper
  - Uncertainty-Guided Exploration (2 papers)

    - [24] Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents (Wang Jiawei, 2025) View paper
    - [30] Towards Agents That Know When They Don't Know: Uncertainty as a Control Signal for Structured Reasoning (Martell, 2025) View paper
    - Exploration-Exploitation Balance Frameworks (4 papers)
    - [2] Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism (Liu Jia, 2025) View paper
    - [7] Reasoning with exploration: An entropy perspective (Cheng, 2025) View paper
    - [18] Learn the Ropes, Then Trust the Wins: Self-imitation with Progressive Exploration for Agentic Reinforcement Learning (Qin, 2025) View paper
    - [33] Exploration vs Exploitation: Rethinking RLVR through Clipping, Entropy, and Spurious Reward (Peter Chen, 2025) View paper
- Token-Level and Sample-Level Analysis
    - High-Entropy Token Analysis (3 papers)
    - [1] Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning (Wang Shen-zhi, 2025) View paper
    - [22] First return, entropy-eliciting explore (Zheng Tianyu, 2025) View paper
    - [35] Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping (Chen Shuang, 2025) View paper
    - Low-Entropy Token and Segment Analysis (3 papers)
    - [28] Beyond High-Entropy Exploration: Correctness-Aware Low-Entropy Segment-Based Advantage Shaping for Reasoning LLMs (Xinzhu Chen, 2025) View paper
    - [29] Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward (Huang Guan-hua, 2025) View paper
    - [34] Stabilizing Knowledge, Promoting Reasoning: Dual-Token Constraints for RLVR (Wang Jiakang, 2025) View paper
    - Entropy-Based Sample Weighting (3 papers)
    - [19] Efficient Reinforcement Learning with Semantic and Token Entropy for LLM Reasoning (Hongye Cao, 2025) View paper
    - [27] Entropy-guided sequence weighting for efficient exploration in RL-based LLM fine-tuning (Vanlioglu, 2025) View paper
    - [36] ESPO: Entropy Importance Sampling Policy Optimization (Yuepeng Sheng, 2025) View paper
- Entropy Collapse and Training Stability
    - Entropy Collapse Diagnosis and Mechanisms (3 papers)
    - [4] The entropy mechanism of reinforcement learning for reasoning language models (Cui, 2025) View paper
    - [10] Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning (Deng Jia, 2025) View paper
    - [25] Rethinking Entropy Regularization in Large Reasoning Models (Jiang Yu-xian, 2025) View paper
    - Entropy Collapse Prevention Strategies (3 papers)
    - [3] Epo: Entropy-regularized policy optimization for llm agents reinforcement learning (Xu Wujiang, 2025) View paper
    - [12] Arbitrary Entropy Policy Optimization: Entropy Is Controllable in Reinforcement Fine-tuning (Wang Chen, 2025) View paper
    - [46] Arbitrary Entropy Policy Optimization Breaks The Exploration Bottleneck of Reinforcement Learning (Chen Wang, 2025) View paper
- Policy Optimization Algorithms and Frameworks
    - Sequence and Trajectory-Level Optimization (3 papers)
    - [17] Entropy-regularized token-level policy optimization for language agent reinforcement (Wen, 2024) View paper
    - [41] CTRLS: Chain-of-Thought Reasoning via Latent State-Transition (Wu, 2025) View paper
    - Advantage Estimation and Credit Assignment (2 papers)
    - [38] Quantile Advantage Estimation for Entropy-Safe Reasoning (WU Junkang, 2025) View paper
    - [44] Offline Reinforcement Learning for LLM Multi-Step Reasoning (Wang Huai-jie, 2024) View paper
- Application Domains and Specialized Settings
    - Mathematical Reasoning and Formal Tasks (2 papers)
    - [31] Confucius3-Math: A Lightweight High-Performance Reasoning LLM for Chinese K-12 Mathematics Learning (Wu Lixin, 2025) View paper
    - [37] Exploring RL-based LLM Training for Formal Language Tasks with Programmed Rewards (Soemers, 2024) View paper
    - Multimodal and Specialized Applications (1 papers)
    - [9] Entropy-reinforced planning with large language models for drug discovery (Liu Xue-feng, 2024) View paper
    - Offline and Continual Learning Settings (3 papers)
    - [8] Parseval Regularization for Continual Reinforcement Learning (Chung, 2024) View paper
    - [48] Generalizing Consistency Policy to Visual RL with Prioritized Proximal Experience Regularization (Li, 2024) View paper
    - [50] Statistical analysis of Inverse Entropy-regularized Reinforcement Learning (Denis Belomestny, 2025) View paper
    - Efficiency and Resource Optimization (1 papers)
    - [5] Enhancing efficiency and exploration in reinforcement learning for llms (Xi Xiangyu, 2025) View paper
    - Domain-Specific Applications (3 papers)
    - [39] Risk-aware operation of offshore multi-energy microgrids using large-language-model assisted distributed pareto-optimal reinforcement learning (C Liu, 2025) View paper
    - [40] Semi-Structured Interview System Based on Fine-Tuned Large Language Model and Reinforcement Learning from Human Feedback (Yanni Ma, 2025) View paper
    - [47] Adaptive Confidence-Weighted LLM Infusion for Financial Reinforcement Learning (Emran Y. Alturki, 2025) View paper

## Narrative

Core task: entropy control in large language model reinforcement learning. The field addresses how to manage the randomness and diversity of token-level decisions when training LLMs with RL, balancing exploration of novel responses against exploitation of high-reward behaviors. The taxonomy organizes research into several main branches: Entropy Regularization Methods and Mechanisms focuses on explicit penalties or bonuses that shape policy entropy, including fixed-coefficient schemes and adaptive strategies that adjust regularization strength during training. Exploration-Driven Approaches emphasize curiosity signals and uncertainty-based mechanisms to guide search in large action spaces. Token-Level and Sample-Level Analysis examines entropy at different granularities, from individual token distributions to full sequence variability. Entropy Collapse and Training Stability investigates pathologies where policies become overly deterministic or unstable, while Policy Optimization Algorithms and Frameworks covers broader algorithmic designs that incorporate entropy considerations. Application Domains and Specialized Settings explores how entropy control manifests in reasoning tasks, code generation, and other specialized contexts. Representative works such as EPO Entropy Regularized[3] and ETTRL Entropy

Mechanism[2] illustrate how regularization can be integrated into policy gradient methods, while Efficiency Exploration RL[5] and Reasoning Exploration Entropy[7] highlight exploration-centric designs.

A particularly active line of work centers on adaptive and dynamic entropy regularization, where the strength or form of entropy penalties evolves based on training signals or task characteristics. Entropy Control LLM-RL[0] sits squarely within this adaptive branch, proposing mechanisms that adjust regularization dynamically rather than relying on fixed hyperparameters. This contrasts with neighboring efforts like Adaptive Divergence Regularization[43], which modulates KL penalties between policy and reference distributions, and Adaptive Entropy Coefficient[45], which tunes a scalar entropy weight over time. The central trade-off across these methods is between maintaining sufficient exploration to discover high-quality solutions and preventing entropy collapse that leads to degenerate or repetitive outputs. Open questions include how to set or learn adaptation schedules, whether token-level or sequence-level entropy metrics are more informative, and how entropy control interacts with reward shaping and other training stabilizers in large-scale LLM settings.

## Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. EntroPIC: Towards Stable Long-Term Training of LLMs via Entropy Stabilization with Proportional-Integral Control

**Authors**: Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Long-term training of large language models (LLMs) requires maintaining stable exploration to prevent the model from collapsing into sub-optimal behaviors. Entropy is crucial in this context, as it controls exploration and helps avoid premature convergence to sub-optimal solutions. However, existing reinforcement learning methods struggle to maintain an appropriate level of entropy, as the training process involves a mix of positive and negative samples, each affecting entropy in different ways ...

#### Relationship Analysis

Both papers belong to the Adaptive and Dynamic Entropy Regularization category, focusing on automatically adjusting entropy control during LLM-RL training. They overlap in addressing entropy collapse issues and proposing dynamic coefficient adjustment mechanisms to balance exploration and exploitation. The key difference is that the original paper (AEnt) uses clamped entropy on a reduced token space with automatic coefficient adjustment based on entropy thresholds, while the candidate paper (EntroPIC) employs Proportional-Integral (PI) control theory to dynamically adjust weights of positive/negative samples based on deviation from target entropy, with theoretical convergence guarantees for both on-policy and off-policy settings.

---

### 2. Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models

**Authors**: Fan Jiajun, Wei Tong, Jiajun Fan, Cheng, Chaoran, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Balancing exploration and exploitation during reinforcement learning fine-tuning of generative models presents a critical challenge, as existing approaches rely on fixed divergence regularization that creates an inherent dilemma: strong regularization preserves model capabilities but limits reward optimization, while weak regularization enables greater alignment but risks instability or reward hacking. We introduce Adaptive Divergence Regularized Policy Optimization (ADRPO), which automatically ...

#### Relationship Analysis

Both papers belong to the Adaptive and Dynamic Entropy Regularization category, focusing on automatically adjusting entropy regularization during training. They overlap in addressing the challenge of balancing exploration and exploitation in LLM-RL through dynamic entropy control mechanisms. The key difference is that the original paper (AEnt) uses clamped entropy on a reduced token space with automatic coefficient adjustment based on entropy levels, while the candidate paper (ADRPO) adjusts regularization strength based on advantage estimates to modulate exploration-exploitation at the sample level, and extends beyond LLMs to flow matching models for text-to-image generation.

---

### 3. Rediscovering Entropy Regularization: Adaptive Coefficient Unlocks Its Potential for LLM Reinforcement Learning

**Authors**: Zhang XiaoYun, Yuan Xiao-jian, Xiaoyun Zhang, Huang Di, Xiaojian Yuan, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Reasoning ability has become a defining capability of Large Language Models (LLMs), with Reinforcement Learning with Verifiable Rewards (RLVR) emerging as a key paradigm to enhance it. However, RLVR training often suffers from policy entropy collapse, where the policy becomes overly deterministic, hindering exploration and limiting reasoning performance. While entropy regularization is a common remedy, its effectiveness is highly sensitive to the fixed coefficient, making it unstable across task...

#### Relationship Analysis

Both papers belong to the Adaptive and Dynamic Entropy Regularization category, focusing on automatically adjusting entropy regularization strength during LLM-RL training. They share overlapping concerns about entropy collapse in LLM-RL and both propose adaptive coefficient adjustment mechanisms to maintain policy entropy within desired ranges. The key difference is that the original paper (AEnt) introduces a clamped entropy bonus evaluated on a reduced token space with automatic coefficient adjustment based on clamped entropy values, while the candidate paper (AER) proposes difficulty-aware coefficient allocation at the sample level, initial-anchored target entropy determination, and dynamic global coefficient adjustment without token space reduction.

## Contributions Analysis

**Overall novelty summary.** The paper proposes AEnt, an adaptive entropy regularization method for LLM-RL that addresses issues arising from large response spaces and sparse optimal outputs. It resides in the 'Adaptive and Dynamic Entropy Regularization' leaf, which contains four papers including this one. This leaf sits within the broader 'Entropy Regularization Methods and Mechanisms' branch, indicating a moderately populated research direction focused on explicit entropy control techniques. The taxonomy shows fifty papers across the entire field, with this particular leaf representing one of several approaches to entropy management in LLM-RL training.

The taxonomy reveals neighboring leaves addressing related but distinct mechanisms: 'Fixed-Coefficient and Clipping-Based Methods' explores static regularization schemes, 'KL-Divergence Regularization Design' examines divergence-based constraints, and 'Entropy Minimization Approaches' focuses on concentration rather than exploration. The paper's adaptive coefficient adjustment connects it to the broader 'Exploration-Driven Approaches' branch, which includes curiosity-based and uncertainty-guided methods. The scope note for the paper's leaf explicitly excludes fixed-coefficient methods and clipping-based approaches, positioning AEnt as a dynamic alternative to static entropy control strategies.

Among thirty candidates examined, the analysis identified limited prior work overlap. The theoretical analysis of entropy regularization issues in LLM-RL showed no refutable candidates across ten examined papers. The clamped entropy bonus mechanism similarly found no overlapping prior work among ten candidates. The adaptive coefficient adjustment scheme encountered one refutable candidate among ten examined, suggesting some existing work on dynamic entropy tuning. The relatively small number of refutable findings across contributions indicates that, within the examined candidate set, the combination of clamped entropy and automatic coefficient adjustment appears less extensively explored.

Based on the limited search scope of thirty semantically similar papers, the work appears to occupy a moderately novel position within adaptive entropy regularization. The taxonomy structure suggests this is an active but not overcrowded research area, with the paper's specific combination of clamped entropy and automatic coefficient adjustment showing limited overlap in the examined candidate pool. The analysis does not cover exhaustive literature review or assess contributions outside the top-K semantic matches and their citations.

---

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

## Contribution 1: Theoretical analysis of entropy regularization issues in LLM-RL

**Description**: The authors provide a theoretical framework explaining why traditional entropy regularization fails in LLM-RL settings. They show that entropy collapse indicates learning stagnancy and that conventional entropy regularization suffers from bias due to LLM's large response space and sparse optimal actions, as formalized in Propositions 1 and 2.

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. Rethinking Entropy Regularization in Large Reasoning Models

**URL**: View paper

**Brief Assessment**

Rethinking Entropy Regularization[25] focuses on empirical analysis of entropy collapse in LRMs and proposes selective masking mechanisms, rather than providing formal theoretical propositions about entropy regularization bias as in the original paper's Propositions 1 and 2.

---

### 2. Entropy Regularizing Activation: Boosting Continuous Control, Large Language Models, and Image Classification with Activation as Entropy Constraints

**URL**: View paper

**Brief Assessment**

Entropy Regularizing Activation[23] focuses on architectural entropy constraints via activation functions rather than analyzing why traditional entropy regularization fails in LLM-RL. The candidate does not provide theoretical analysis of entropy regularization bias or collapse mechanisms in LLM settings.

---

### 3. Decoupling regularization from the action space

**URL**: View paper

**Brief Assessment**

Decoupling Regularization Action[51] addresses entropy regularization bias arising from varying action space sizes across states, not the specific LLM-RL context with extremely large vocabulary spaces and sparse optimal token sequences that the original paper analyzes.

---

### 4. Action redundancy in reinforcement learning

**URL**: View paper

**Brief Assessment**

Action Redundancy RL[53] focuses on action redundancy in general RL settings with large action spaces (e.g., Atari, MuJoCo), not specifically on LLM-RL training. The paper does not address the specific theoretical issues of entropy regularization in LLM settings such as sparse optimal actions in language generation or the bias formalized in the original paper's Propositions 1 and 2.

---

### 5. Sparse actor-critic: Sparse tsallis entropy regularized reinforcement learning in a continuous action space

**URL**: View paper

**Brief Assessment**

Sparse Tsallis Entropy[52] focuses on continuous control tasks with sparse Tsallis entropy regularization, not on LLM-RL settings or the specific theoretical issues of entropy collapse and bias in large discrete action spaces that characterize language models.

---

### 6. Efficient entropy for policy gradient with multidimensional action space

**URL**: View paper

**Brief Assessment**

Multidimensional Action Entropy[58] addresses computational challenges of entropy calculation in high-dimensional discrete action spaces (e.g., multi-agent coordination), not the bias issues arising from LLM's large vocabulary and sparse optimal responses that the original paper analyzes.

---

### 7. Offline reinforcement learning for learning to dispatch for job shop scheduling

**URL**: View paper

**Brief Assessment**

The candidate paper (Job Shop Scheduling[55]) focuses on offline reinforcement learning for job shop scheduling problems, not on entropy regularization in LLM-RL settings. The candidate addresses entropy regularization in the context of discrete action spaces for scheduling, which is a fundamentally different domain from language model training.

---

### 8. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm

**URL**: View paper

**Brief Assessment**

Neural Natural Actor-Critic[56] focuses on entropy-regularized RL with neural network approximation in general MDPs, not specifically on LLM-RL settings. The candidate does not address the specific issues of large response spaces and sparse optimal actions in LLM contexts that the original paper identifies.

---

### 9. Efficient Learning for Entropy-Regularized Markov Decision Processes via Multilevel Monte Carlo

**URL**: View paper

**Brief Assessment**

Multilevel Monte Carlo[54] focuses on entropy-regularized MDPs with Polish state and action spaces using Monte Carlo sampling methods, not on LLM-specific reinforcement learning settings or the bias issues arising from large response spaces in language models.

---

#### 10. Implicitly regularized rl with implicit q-values

**URL**: View paper

**Brief Assessment**

Implicitly Regularized RL[57] focuses on implicit Q-value parametrization for continuous action spaces in general RL, not on analyzing entropy regularization failures specific to LLM settings with large vocabulary spaces and sparse optimal tokens.

---

### Contribution 2: AEnt algorithm with clamped entropy bonus

**Description**: The authors introduce AEnt, a novel entropy regularization method that computes entropy on a re-normalized policy defined over a reduced token space (top probability tokens). This clamped entropy encourages exploration within a more compact response set, reducing the bias induced by the extremely large vocabulary in LLMs.

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

**URL**: View paper

**Brief Assessment**

Adaptive Joint Entropy[72] focuses on intrinsic reward mechanisms for exploration in general RL settings, not on entropy regularization methods for LLM policy optimization with token space reduction.

---

#### 2. Historical decision-making regularized maximum entropy reinforcement learning

**URL**: View paper

**Brief Assessment**

Historical Decision Regularization[70] focuses on historical decision-making regularization within maximum entropy RL for continuous control tasks, not on clamped entropy computed over reduced token spaces for LLM-RL settings.

---

#### 3. ESPO: Entropy Importance Sampling Policy Optimization

**URL**: View paper

**Brief Assessment**

ESPO Importance Sampling[36] focuses on entropy-based importance sampling for sequence decomposition and adaptive clipping in group-based policy optimization, not on clamped entropy regularization over reduced token spaces as proposed in the original paper's AEnt method.

---

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

**URL**: View paper

**Brief Assessment**

Adaptive Entropy Multi-Agent[61] focuses on multi-agent RL with adaptive entropy regularization across different agents, not on clamped entropy for LLM token spaces. The candidate addresses agent-level entropy allocation in cooperative tasks, while the original addresses token-space reduction for LLM vocabulary exploration.

---

#### 5. State entropy regularization for robust reinforcement learning

**URL**: View paper

**Brief Assessment**

State Entropy Regularization[68] focuses on state entropy regularization for robustness in standard RL settings, not on clamped entropy over reduced token spaces for LLM-RL. The candidate addresses robustness to transition/reward perturbations in MDPs, while the original addresses exploration bias from large vocabulary in language model training.

---

#### 6. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning

**URL**: View paper

**Brief Assessment**

Maximum Entropy Exploration[69] focuses on maximizing entropy of achieved goal distributions in multi-goal RL for exploration, not on clamped entropy bonuses for policy optimization in LLM-RL settings with vocabulary-based token spaces.

---

#### 7. Provably efficient maximum entropy exploration

**URL**: View paper

**Brief Assessment**

Provably Efficient Exploration[71] focuses on maximum entropy exploration in MDPs with state-visitation frequency objectives, using conditional gradient methods. The original paper addresses entropy regularization specifically for LLM-RL with vocabulary-level token space reduction, which is a fundamentally different problem domain and technical approach.

---

#### 8. Arbitrary Entropy Policy Optimization: Entropy Is Controllable in Reinforcement Fine-tuning

**URL**: View paper

**Brief Assessment**

Arbitrary Entropy Policy[12] focuses on temperature-based sampling and REINFORCE regularization to control entropy, rather than clamping the token space for entropy calculation. The candidate's mechanism differs fundamentally from the original's clamped entropy approach.

---

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

**URL**: View paper

**Brief Assessment**

CE-GPPO Coordinating Entropy[16] focuses on gradient-preserving mechanisms for clipped tokens in PPO-style algorithms, not on clamped entropy bonuses computed over reduced token spaces as in the original paper's AEnt method.

---

### 10. Reasoning with exploration: An entropy perspective

**URL**: View paper

**Brief Assessment**

Reasoning Exploration Entropy[7] uses entropy to shape advantages in RL training, while the original paper proposes clamped entropy regularization on a reduced token space. These are fundamentally different mechanisms: advantage shaping vs. entropy regularization on re-normalized policies.

---

## Contribution 3: Adaptive entropy coefficient adjustment scheme

**Description**: The authors propose an automatic adjustment mechanism for the entropy coefficient during training. The coefficient is dynamically updated to keep the clamped entropy within specified bounds, balancing the benefits of entropy regularization against its bias and preventing issues like entropy collapse or explosion.

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**

Adaptive Maximum Entropy[59] focuses on off-policy maximum entropy deep RL with meta-gradients for SAC temperature tuning. The original paper addresses on-policy LLM-RL with clamped entropy on reduced token spaces and automatic coefficient adjustment based on entropy bounds, representing a fundamentally different technical approach and application domain.

---

### 2. Relative entropy inverse reinforcement learning

**URL**: View paper

**Brief Assessment**

Relative Entropy IRL[60] focuses on inverse reinforcement learning with a fixed entropy regularization framework, not on adaptive coefficient adjustment during training as proposed in the original paper.

---

### 3. Rediscovering Entropy Regularization: Adaptive Coefficient Unlocks Its Potential for LLM Reinforcement Learning

**URL**: View paper

**Prior Art Analysis**

Adaptive Entropy Coefficient[45] demonstrates that adaptive entropy coefficient adjustment mechanisms existed prior to the original paper's submission. The candidate paper proposes a comprehensive adaptive entropy regularization framework that dynamically adjusts entropy coefficients during training to maintain policy entropy within target bounds. Both papers address the same core problem: preventing entropy collapse/explosion through automatic coefficient adjustment rather than fixed coefficients. The candidate's mechanism includes difficulty-aware allocation, initial-anchored target entropy, and dynamic global coefficient adjustment - all components that automatically modify the entropy coefficient based on training dynamics, which directly overlaps with the original paper's claimed novelty of 'automatic adjustment mechanism for the entropy coefficient during training.'

**Evidence**

Evidence 1 - **Rationale**: Both papers propose automatic/adaptive adjustment of entropy coefficients. The original claims novelty in 'automatically adjusted coefficient' while the candidate demonstrates a prior framework with 'dynamic global coefficient adjustment' - directly overlapping core mechanisms. - **Original**: we propose aent, an entropy control method that utilizes a new clamped entropy bonus with an automatically adjusted coefficient. the clamped entropy is evaluated with the re-normalized policy defined on certain smaller token space, which encourages exploration within a more compact response set. in ... - **Candidate**: we propose adaptive entropy regularization (aer) - a framework that dynamically balances exploration and exploitation via three components: difficulty-aware coefficient allocation, initial-anchored target entropy, and dynamic global coefficient adjustment. experiments on multiple mathematical reason...

Evidence 2 - **Rationale**: Both papers describe mechanisms that adjust coefficients to keep entropy within bounds - the original increases/decreases λ based on entropy limits, while the candidate maintains entropy near target values. This demonstrates the same fundamental approach to adaptive coefficient control. - **Original**: the algorithm will try to confine ˜h within [ ˜hlow, ˜hhigh] by increasing/decreasing λ when ˜h(πθ) is lower/higher than the limits. the intuition is that when entropy is high, the coefficient should be tuned down to reduce the entropy induced bias and shift weights to reward maximization, which in ... - **Candidate**: effective exploration may require maintaining policy entropy at a 'sweet spot' below its initial level to avoid entropy collapse and explosion. he et al. (2025) have similar empirical observations that they monitor policy entropy during training and preventing it from falling below a prespecified ta...

Evidence 3 - **Rationale**: Both papers implement step-by-step coefficient adjustment during training. The original adjusts at each global step, while the candidate adjusts dynamically - both demonstrating automatic coefficient modification mechanisms that existed prior to the original's claimed novelty. - **Original**: at the end of each global step, the entropy coefficient is adjusted according to scheme 4.1. the whole process is summarized in algorithm 1. - **Candidate**: aer estimates task difficulty with respect to the current policy and adaptively adjusts the entropy coefficient at the sample level. it further sets the target entropy as a fraction of the initial policy entropy and dynamically adjusts a global scaling factor for coefficients to prevent the policy e...

---

### 4. Survey of Unified Representation Technology of Multi-dimensional Information for Low Altitude Intelligent Network

**URL**: View paper

**Brief Assessment**

Multi-Dimensional Information Representation[65] focuses on unified representation technology for low altitude intelligent networks and mentions 'dynamic entropy regularization mechanism' only in passing. The candidate does not provide sufficient detail about adaptive entropy coefficient adjustment methods to challenge the original paper's novelty claim regarding automatic adjustment mechanisms that keep clamped entropy within specified bounds during training.

---

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

**URL**: View paper

**Brief Assessment**

Adaptive Entropy Multi-Agent[61] adapts entropy coefficients per agent in multi-agent settings to balance exploration across agents. The original adapts a single coefficient during LLM training to maintain clamped entropy bounds, addressing different technical challenges in different domains.

---

### 6. A novel dynamically adjusted entropy algorithm for collision avoidance in autonomous ships based on deep reinforcement learning

**URL**: View paper

**Brief Assessment**

Collision Avoidance Ships[66] uses a quadratically decreasing entropy method for ship navigation tasks, which is a predetermined decay schedule rather than an automatic adjustment mechanism that responds to clamped entropy bounds as in the original paper's LLM-RL context.

---

### 7. Off-policy deep reinforcement learning with automatic entropy adjustment for adaptive online grid emergency control

**URL**: View paper

**Brief Assessment**

Automatic Entropy Grid[63] focuses on grid emergency control applications in power systems, not general RL frameworks or LLM training. The technical domain and application context differ fundamentally from the original paper's LLM-RL setting.

---

### 8. State-dependent maximum entropy reinforcement learning for robot long-horizon task learning

**URL**: View paper

**Brief Assessment**

State-Dependent Maximum Entropy[62] focuses on adjusting entropy weights at specific critical states (up-stage, stuck, ahead) in long-horizon robotic tasks, not on automatic coefficient adjustment to prevent entropy collapse/explosion during training as in the original paper's LLM-RL context.

---

### 9. Deep reinforcement learning in maximum entropy framework with automatic adjustment of mixed temperature parameters for path planning

**URL**: View paper

**Brief Assessment**

Mixed Temperature Path[67] focuses on path planning with mixed temperature parameters for state-dependent entropy control, while the original paper addresses LLM-RL training with clamped entropy and automatic coefficient adjustment to handle large token spaces.

---

### 10. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor

**URL**: View paper

**Brief Assessment**

Soft Actor-Critic[64] uses a fixed temperature parameter α (or reward scaling) that requires manual tuning, not an adaptive adjustment mechanism. The paper explicitly states 'we found reward scale to be the only hyperparameter that requires tuning' and discusses sensitivity to this fixed parameter rather than proposing automatic adjustment during training.

---

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References


- [0] On Entropy Control in LLM-RL Algorithms View paper
- [1] Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning View paper
- [2] Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism View paper
- [3] Epo: Entropy-regularized policy optimization for llm agents reinforcement learning View paper
- [4] The entropy mechanism of reinforcement learning for reasoning language models View paper
- [5] Enhancing efficiency and exploration in reinforcement learning for llms View paper
- [6] Entropy Ratio Clipping as a Soft Global Constraint for Stable Reinforcement Learning View paper
- [7] Reasoning with exploration: An entropy perspective View paper
- [8] Parseval Regularization for Continual Reinforcement Learning View paper
- [9] Entropy-reinforced planning with large language models for drug discovery View paper
- [10] Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning View paper
- [11] Entropy-regularized process reward model View paper
- [12] Arbitrary Entropy Policy Optimization: Entropy Is Controllable in Reinforcement Fine-tuning View paper
- [13] Controlling large language model agents with entropic activation steering View paper
- [14] Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models View paper
- [15] Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models View paper
- [16] CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning View paper
- [17] Entropy-regularized token-level policy optimization for language agent reinforcement View paper
- [18] Learn the Ropes, Then Trust the Wins: Self-imitation with Progressive Exploration for Agentic Reinforcement Learning View paper
- [19] Efficient Reinforcement Learning with Semantic and Token Entropy for LLM Reasoning View paper
- [20] On the design of kl-regularized policy gradient algorithms for llm reasoning View paper
- [21] The unreasonable effectiveness of entropy minimization in llm reasoning View paper
- [22] First return, entropy-eliciting explore View paper
- [23] Entropy Regularizing Activation: Boosting Continuous Control, Large Language Models, and Image Classification with Activation as Entropy Constraints View paper
- [24] Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents View paper
- [25] Rethinking Entropy Regularization in Large Reasoning Models View paper
- [26] EntroPIC: Towards Stable Long-Term Training of LLMs via Entropy Stabilization with Proportional-Integral Control View paper
- [27] Entropy-guided sequence weighting for efficient exploration in RL-based LLM fine-tuning View paper
- [28] Beyond High-Entropy Exploration: Correctness-Aware Low-Entropy Segment-Based Advantage Shaping for Reasoning LLMs View paper
- [29] Low-probability Tokens Sustain Exploration in Reinforcement Learning with Verifiable Reward View paper
- [30] Towards Agents That Know When They Don't Know: Uncertainty as a Control Signal for Structured Reasoning View paper

- [31] Confucius3-Math: A Lightweight High-Performance Reasoning LLM for Chinese K-12 Mathematics Learning View paper
- [32] Rethinking entropy interventions in rlvr: An entropy change perspective View paper
- [33] Exploration vs Exploitation: Rethinking RLVR through Clipping, Entropy, and Spurious Reward View paper
- [34] Stabilizing Knowledge, Promoting Reasoning: Dual-Token Constraints for RLVR View paper
- [35] Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping View paper
- [36] ESPO: Entropy Importance Sampling Policy Optimization View paper
- [37] Exploring RL-based LLM Training for Formal Language Tasks with Programmed Rewards View paper
- [38] Quantile Advantage Estimation for Entropy-Safe Reasoning View paper
- [39] Risk-aware operation of offshore multi-energy microgrids using large-language-model assisted distributed pareto-optimal reinforcement learning View paper
- [40] Semi-Structured Interview System Based on Fine-Tuned Large Language Model and Reinforcement Learning from Human Feedback View paper
- [41] CTRLS: Chain-of-Thought Reasoning via Latent State-Transition View paper
- [42] APO: Enhancing Reasoning Ability of MLLMs via Asymmetric Policy Optimization View paper
- [43] Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models View paper
- [44] Offline Reinforcement Learning for LLM Multi-Step Reasoning View paper
- [45] Rediscovering Entropy Regularization: Adaptive Coefficient Unlocks Its Potential for LLM Reinforcement Learning View paper
- [46] Arbitrary Entropy Policy Optimization Breaks The Exploration Bottleneck of Reinforcement Learning View paper
- [47] Adaptive Confidence-Weighted LLM Infusion for Financial Reinforcement Learning View paper
- [48] Generalizing Consistency Policy to Visual RL with Prioritized Proximal Experience Regularization View paper
- [49] Entropy-Regularized Token-Level Policy Optimization for Large Language Models View paper
- [50] Statistical analysis of Inverse Entropy-regularized Reinforcement Learning View paper
- [51] Decoupling regularization from the action space View paper
- [52] Sparse actor-critic: Sparse tsallis entropy regularized reinforcement learning in a continuous action space View paper
- [53] Action redundancy in reinforcement learning View paper
- [54] Efficient Learning for Entropy-Regularized Markov Decision Processes via Multilevel Monte Carlo View paper
- [55] Offline reinforcement learning for learning to dispatch for job shop scheduling View paper
- [56] Finite-time analysis of entropy-regularized neural natural actor-critic algorithm View paper
- [57] Implicitly regularized rl with implicit q-values View paper
- [58] Efficient entropy for policy gradient with multidimensional action space View paper
- [59] Off-policy asymptotic and adaptive maximum entropy deep reinforcement learning View paper
- [60] Relative entropy inverse reinforcement learning View paper
- [61] An adaptive entropy-regularization framework for multi-agent reinforcement learning View paper
- [62] State-dependent maximum entropy reinforcement learning for robot long-horizon task learning View paper
- [63] Off-policy deep reinforcement learning with automatic entropy adjustment for adaptive online grid emergency control View paper
- [64] Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor View paper
- [65] Survey of Unified Representation Technology of Multi-dimensional Information for Low Altitude Intelligent Network View paper
- [66] A novel dynamically adjusted entropy algorithm for collision avoidance in autonomous ships based on deep reinforcement learning View paper
- [67] Deep reinforcement learning in maximum entropy framework with automatic adjustment of mixed temperature parameters for path planning View paper
- [68] State entropy regularization for robust reinforcement learning View paper
- [69] Maximum entropy gain exploration for long horizon multi-goal reinforcement learning View paper
- [70] Historical decision-making regularized maximum entropy reinforcement learning View paper
- [71] Provably efficient maximum entropy exploration View paper
- [72] Adaptive joint entropy reward: a mechanism to efficient exploration in reinforcement learning View paper