Novelty Assessment Report

Paper: RESTRAIN: From Spurious Votes to Signals — Self-Training RL with Self-Penalization
PDF URL: https://openreview.net/pdf?id=87ySF7viys
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Reinforcement learning with human-annotated data has boosted chain-of-thought reasoning in large reasoning models, but these gains come at high costs in labeled data while faltering on harder tasks. A natural next step is experience-driven learning, where models improve without curated labels by adapting to unlabeled data. We introduce REinforcement learning with Self-resTRAINt training (RESTRAIN), a self-penalizing RL framework that converts the absence of gold labels into a useful learning signal. Instead of overcommitting to spurious majority votes, RESTRAIN exploits signals from the model's entire answer distribution: penalizing overconfident rollouts and low-consistency examples while preserving promising reasoning chains. This self-penalization mechanism integrates seamlessly into policy optimization methods such as GRPO, enabling continual self-improvement without supervision. On challenging reasoning benchmarks, RESTRAIN delivers large gains using only unlabeled data. With Qwen3-4B-Base and OctoThinker Hybrid-8B-Base, it boosts Pass@1 by up to +140.7% on AIME25, +36.2% on MMLU STEM, and +19.6% on GPQA-Diamond, nearly matching gold-label training while using no gold labels. These results demonstrate that RESTRAIN establishes a scalable path toward stronger reasoning without gold labels.

Core Task Landscape

This paper addresses: **Self-improving reinforcement learning without gold labels**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Self-Reward Generation and Preference Learning**
- **Majority Voting and Distribution-Based Learning**
- **Curriculum and Online Learning Frameworks**
- **Theoretical Foundations and Cognitive Frameworks**
- **Domain-Specific Applications**
- **Robotic and Embodied RL**
- **Specialized Learning Paradigms**
- **Evaluation and Meta-Learning**
- **Specialized Domains and Applications**

Complete Taxonomy Tree

- Self-improving reinforcement learning without gold labels Survey Taxonomy
- Self-Reward Generation and Preference Learning
  - Self-Feedback and Confidence-Based Rewards (7 papers)
  - [1] TTRL: Test-Time Reinforcement Learning (Zuo Yuxin, 2025) View paper
  - [4] Language Model Self-improvement by Reinforcement Learning Contemplation (Pang, 2023) View paper
  - [7] Post-Training Large Language Models via Reinforcement Learning from Self-Feedback (van Niekerk, 2025) View paper
  - [8] Co-Reward: Self-supervised Reinforcement Learning for Large Language Model Reasoning via Contrastive Agreement (Zhang, 2025) View paper
  - [16] Self-Evolved Reward Learning for LLMs (Huang Cheng-hua, 2024) View paper
  - [17] Self Rewarding Self Improving (T Simonds, 2025) View paper
  - [20] R-Zero: Self-Evolving Reasoning LLM from Zero Data (Huang, 2025) View paper
  - AI-Generated Feedback and Synthetic Preferences (3 papers)
  - [2] RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback (Lee, 2023) View paper
  - [35] West-of-N: Synthetic Preferences for Self-Improving Reward Models (Pace, 2024) View paper
  - [42] Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection (Lee Kyung-Jae, 2024) View paper
  - Self-Correction and Iterative Refinement (3 papers)
  - [11] Training Language Models to Self-Correct via Reinforcement Learning (Kumar, 2024) View paper
  - [21] SERL: Self-Examining Reinforcement Learning on Open-Domain (Weixuan Ou, 2025) View paper
  - [33] Self-improvement in language models: The sharpening mechanism (Huang Audrey, 2024) View paper
- Majority Voting and Distribution-Based Learning
  - Majority Voting with Self-Penalization ★ (2 papers)
  - [0] RESTRAIN: From Spurious Votes to Signals — Self-Training RL with Self-Penalization (Anon et al., 2026) View paper
  - [31] RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization (Yu, 2025) View paper
  - Evolutionary and Consistency-Based Selection (2 papers)

    - [23] Can Large Reasoning Models Self-Train? (Tajwar, 2025) View paper
    - [38] Evolving Language Models without Labels: Majority Drives Selection, Novelty Promotes Variation (Zhou Yujun, 2025) View paper
- Curriculum and Online Learning Frameworks
    - Self-Play and Adversarial Curriculum (2 papers)
    - [32] SPELL: Self-Play Reinforcement Learning for evolving Long-Context Language Models (Yang Ziyi, 2025) View paper
    - [36] SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn Reinforcement Learning (Liu Bo, 2025) View paper
    - Online Adaptation and Bilevel Optimization (3 papers)
    - [12] SAIL: Self-Improving Efficient Online Alignment of Large Language Models (Ding, 2024) View paper
    - [19] WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning (Liu, 2024) View paper
    - [24] EvoTest: Evolutionary Test-Time Learning for Self-Improving Agentic Systems (He Yufei, 2025) View paper
- Theoretical Foundations and Cognitive Frameworks
    - Theoretical Analysis of Self-Improvement (1 papers)
    - [14] RL-STaR: Theoretical Analysis of Reinforcement Learning Frameworks for Self-Taught Reasoner (Chang, 2024) View paper
    - Cognitive and Metacognitive Learning (2 papers)
    - [3] Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs (Gandhi, 2025) View paper
    - [34] Truly Self-Improving Agents Require Intrinsic Metacognitive Learning (Liu, 2025) View paper
- Domain-Specific Applications
    - Code Generation and Repository-Level Tasks (2 papers)
    - [10] RLCoder: Reinforcement Learning for Repository-Level Code Completion (Yan-Lin Wang, 2024) View paper
    - [39] ReST meets ReAct: Self-Improvement for Multi-Step Reasoning LLM Agent (Aksitov, 2023) View paper
    - Mathematical and Geometric Reasoning (2 papers)
    - [13] GeoDRL: A Self-Learning Framework for Geometry Problem Solving using Reinforcement Learning in Deductive Reasoning (Shuai Peng, 2023) View paper
    - [27] Bridging supervised learning and reinforcement learning in math reasoning (Chen Huayu, 2025) View paper
    - Multimodal and Web-Based Agents (1 papers)
    - [15] UniRL: Self-Improving Unified Multimodal Models via Supervised and Reinforcement Learning (Mao, 2025) View paper
- Robotic and Embodied RL
    - Autonomous Navigation and Exploration (2 papers)
    - [28] Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation (Kahn, 2018) View paper
    - [37] Badgr: An autonomous self-supervised learning-based navigation system (Kahn, 2021) View paper
    - Locomotion and Continuous Improvement (1 papers)
    - [22] Grow your limits: Continuous improvement with real-world rl for robotic locomotion (Laura Smith, 2024) View paper
    - Social Navigation and Human Interaction (2 papers)
    - [6] SELFI: Autonomous Self-Improvement with Reinforcement Learning for Social Navigation (Hirose, 2024) View paper
    - [9] Autonomous Robotic Reinforcement Learning with Asynchronous Human Feedback (Max Balsells, 2023) View paper
- Specialized Learning Paradigms
    - Semi-Supervised and Label-Efficient Preference Learning (3 papers)
    - [25] STRAPPER: Preference-based Reinforcement Learning via Self-training Augmentation and Peer Regularization (Kang, 2023) View paper
    - [30] Preference VLM: Leveraging VLMs for Scalable Preference-Based Reinforcement Learning (Ghosh, 2025) View paper
    - [41] Reinforcement Learning for Autonomous Self-Improving Robotic Systems (Sharma, 2024) View paper
    - Multi-View and Shared Autonomy (2 papers)
    - [44] Multi-view Disentanglement for Reinforcement Learning with Multiple Cameras (Dunion, 2024) View paper
    - [46] Shared autonomy via deep reinforcement learning (Siddharth Reddy, 2018) View paper
    - Autonomous Goal Detection and Self-Supervised Exploration (3 papers)
    - [40] Autonomous Goal Detection and Cessation in Reinforcement Learning: A Case Study on Source Term Estimation (Shi Yi-wei, 2024) View paper
    - [47] S4rl: Surprisingly simple self-supervision for offline reinforcement learning in robotics (S Sinha, 2022) View paper
    - [50] Augmenting unsupervised reinforcement learning with self-reference (Zhao, 2023) View paper
- Evaluation and Meta-Learning
    - Progress Evaluation and Regression Detection (1 papers)
    - [45] Progress or regress? self-improvement reversal in post-training (Wu Ting, 2024) View paper
    - Meta-Learning and Algorithm Discovery (2 papers)
    - [29] Meta-Learning Reinforcement Learning for Crypto-Return Prediction (Wang Jun-Qiao, 2025) View paper
    - [43] Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Recursive Self-Improvement through Reinforcement Learning (Ishibashi Yoichi, 2024) View paper
- Specialized Domains and Applications (5 papers)
    - [5] SSRL: Self-Search Reinforcement Learning (Fan Yuchen, 2025) View paper
    - [18] Reinforcement Learning of Self-enhancing Camera Image and Signal Processing (Chandrajit L. Bajaj, 2023) View paper
    - [26] Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach (Xiangkun He, 2022) View paper
    - [48] The ingredients of real-world robotic reinforcement learning (Henry Zhu, 2020) View paper
    - [49] Removing the need for ground truth UWB data collection: self-supervised ranging error correction using deep reinforcement learning (Dieter Coppens, 2024) View paper

Narrative

Core task: Self-improving reinforcement learning without gold labels. This field explores how agents can bootstrap their own learning signals in the absence of external supervision, relying instead on self-generated rewards, peer comparisons, or iterative refinement. The taxonomy reveals several major branches: Self-Reward Generation and Preference Learning focuses on methods that let models judge their own outputs or learn from synthetic preferences (e.g., Self Rewarding[17], RLAIF vs RLHF[2]); Majority Voting and Distribution-Based Learning emphasizes aggregating multiple candidate solutions to identify high-quality trajectories (e.g., RESTRAIN[0], RESTRAIN

Self-Driven[31]); Curriculum and Online Learning Frameworks address how to schedule tasks or adapt policies over time (e.g., TTRL[1], SSRL[5]); Theoretical Foundations and Cognitive Frameworks provide conceptual underpinnings (e.g., Cognitive Behaviors STaRs[3]); and Domain-Specific Applications, Robotic and Embodied RL, and Specialized Learning Paradigms tackle concrete settings ranging from code generation (RLCoder[10]) to robotic manipulation (SERL[21]) and web navigation (WebRL[19]).

A particularly active line of work centers on majority voting and self-penalization strategies, where agents sample multiple rollouts and use agreement or distributional properties to filter or reweight training data. RESTRAIN[0] sits squarely in this cluster, proposing a self-penalization mechanism that discourages overconfident majority votes and encourages exploration of diverse high-quality solutions. This contrasts with simpler majority-voting schemes and aligns closely with RESTRAIN Self-Driven[31], which extends the idea to fully autonomous settings. Meanwhile, works like SSRL[5] and Cognitive Behaviors STaRs[3] explore complementary angles—SSRL[5] emphasizes staged self-improvement with curriculum design, while Cognitive Behaviors STaRs[3] integrates cognitive reasoning steps into the self-training loop. Across these branches, a central tension emerges between exploiting strong majority signals and maintaining sufficient exploration to avoid reward hacking or premature convergence, a challenge that RESTRAIN[0] addresses through its penalization framework.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization

Authors: Yu, Zhaoning, Zhaoning Yu, Tao, Leitian, et al. (27 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Reinforcement learning with human-annotated data has boosted chain-of-thought reasoning in large reasoning models, but these gains come at high costs in labeled data while faltering on harder tasks. A natural next step is experience-driven learning, where models improve without curated labels by adapting to unlabeled data. We introduce RESTRAIN (REinforcement learning with Self-restraint), a self-penalizing RL framework that converts the absence of gold labels into a useful learning signal. Inst...

⚠ Similarity Notice

This paper is highly similar to the original paper; it may be a variant or near-duplicate. Please manually verify.

Contributions Analysis

Overall novelty summary. The paper introduces RESTRAIN, a self-penalizing reinforcement learning framework that enables models to improve on unlabeled data by exploiting signals from the entire answer distribution rather than relying solely on majority votes. It resides in the 'Majority Voting with Self-Penalization' leaf of the taxonomy, which contains only two papers total. This leaf sits within the broader 'Majority Voting and Distribution-Based Learning' branch, indicating a relatively sparse but emerging research direction focused on avoiding spurious convergence in self-improvement settings.

The taxonomy reveals that RESTRAIN's immediate neighbors include evolutionary and consistency-based selection methods in a sibling leaf, as well as broader self-reward generation approaches (e.g., confidence-based rewards, AI-generated feedback) in adjacent branches. While the field contains substantial work on self-feedback and synthetic preference generation (seven papers in Self-Feedback alone), the specific combination of majority voting with explicit penalization mechanisms remains less explored. The taxonomy's scope notes clarify that RESTRAIN's penalization focus distinguishes it from pure majority voting or evolutionary novelty promotion methods.

Among the 24 candidates examined across three contributions, none were found to clearly refute any aspect of RESTRAIN. The core framework (10 candidates examined, 0 refutable), pseudo-label weighting scheme (5 candidates, 0 refutable), and negative rollout penalization (9 candidates, 0 refutable) all appear to lack direct prior work within this limited search scope. This suggests that while the broader field of self-improving RL is active, the specific integration of distribution-based penalization with policy optimization methods like GRPO represents a relatively unexplored combination.

Based on the top-24 semantic matches examined, RESTRAIN appears to occupy a novel position at the intersection of majority voting and self-penalization. However, this assessment is constrained by the limited search scope and the taxonomy's focus on self-improving RL without gold labels. The analysis does not cover exhaustive prior work in supervised learning, traditional RL with external rewards, or related fields where similar penalization ideas might exist under different framing.

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

Contribution 1: RESTRAIN framework for self-driven RL with self-penalization

Description: The authors propose RESTRAIN, a reinforcement learning framework that enables models to self-improve on unlabeled data by penalizing overconfident rollouts and low-consistency examples while preserving promising reasoning chains, without requiring gold labels or external supervision.

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. RLSR: Reinforcement Learning from Self Reward

URL: View paper

Brief Assessment

RLSR[71] focuses on self-judging and self-rewarding mechanisms where models evaluate their own solutions without ground truth, whereas RESTRAIN specifically addresses self-penalization through pseudo-label weighting, negative rollout penalization, and prompt-level weighting based on answer distribution consistency. The technical approaches differ fundamentally in their reward signal generation methods.

---

2. Self-Supervised, Active Learning Seismic Full Waveform Inversion

URL: View paper

Brief Assessment

Seismic Waveform Inversion[66] addresses seismic full waveform inversion using active learning and reinforcement learning concepts for geophysical parameter estimation, not general language model reasoning or self-improvement without gold labels in NLP tasks.

---

3. A Lifetime Extended Energy Management Strategy for Fuel Cell Hybrid Electric Vehicles via Self-Learning Fuzzy Reinforcement Learning

URL: View paper

Brief Assessment

Fuel Cell Energy[68] focuses on fuzzy reinforcement learning for energy management in fuel cell vehicles, not on self-driven RL with self-penalization for reasoning tasks without gold labels. The domains and technical approaches are fundamentally different.

---

4. BEAR: Reinforcement Learning for Throughput Aware Borrowing in Energy Harvesting Systems

URL: View paper

Brief Assessment

BEAR[70] focuses on energy harvesting systems with adaptive penalty rewards for power allocation in wireless transmission, not on self-improvement of language models through self-penalization on unlabeled reasoning data.

---

5. E2CL: Exploration-based Error Correction Learning for Embodied Agents

URL: View paper

Brief Assessment

E2CL[64] focuses on embodied agents learning from exploration-induced errors in interactive environments with environmental feedback, not on self-penalizing RL for reasoning tasks without gold labels. The candidate addresses environment alignment through error correction, while the original addresses self-improvement on unlabeled reasoning data through self-penalization mechanisms.

---

6. RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization

URL: View paper

Brief Assessment

RESTRAIN Self-Driven[31] and the original paper appear to be the same work. The candidate's abstract describes an identical framework with the same name, methodology, and experimental results, indicating this is the original paper itself rather than prior work that could refute its novelty.

---

7. Road Detection for Reinforcement Learning Based Autonomous Car

URL: View paper

Brief Assessment

Road Detection[69] focuses on road detection for autonomous driving using supervised learning to support PPO, not on self-driven RL frameworks that enable models to self-improve on unlabeled data through self-penalization mechanisms.

---

8. Human-compatible driving partners through data-regularized self-play reinforcement learning

URL: View paper

Brief Assessment

Data-Regularized Self-Play[65] focuses on multi-agent autonomous driving coordination through self-play with human reference policies, not on self-improvement for reasoning tasks without gold labels. The technical domains and objectives are fundamentally different.

---

9. A Novel Route Planning Approach Based on Energy-Based Action Sample Reinforcement Learning

URL: View paper

Brief Assessment

Energy-Based Route Planning[67] focuses on route planning for autonomous vehicles using Q-learning with energy-based action sampling, not on self-improvement of language models through self-penalization mechanisms without gold labels.

---

10. Self-supervised boundary offline reinforcement learning

URL: View paper

Brief Assessment

Boundary Offline RL[72] addresses offline RL with synthetic negative samples from uncertain regions using GANs, focusing on distribution mismatch in offline settings. RESTRAIN targets online self-improvement on unlabeled data through self-penalization of overconfident rollouts and low-consistency examples, operating in a fundamentally different paradigm without requiring synthetic data generation or GANs.

Contribution 2: Pseudo-label weighting scheme based on answer frequency

Description: The authors develop a weighting mechanism that assigns weights to pseudo-labels based on their frequency across multiple rollouts, using a monotonic shaping function to down-weight spurious low-frequency answers while avoiding the brittleness of strict majority voting.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Semi-Supervised Clustering Framework for Fine-grained Scene Graph Generation

URL: View paper

Brief Assessment

Scene Graph Clustering[61] addresses scene graph generation with pseudo-labeling for visual relationships, not reinforcement learning for reasoning tasks. The frequency-based weighting in Scene Graph Clustering[61] is applied to relationship categories in computer vision, fundamentally different from the ORIGINAL paper's weighting of answer frequencies across multiple rollouts in RL-based reasoning.

---

2. Semi-Supervised Learning using Pseudo-Labels: A Case Study in Northern SÃ¡mi ASR

URL: View paper

Brief Assessment

Pseudo-Labels ASR[62] applies pseudo-labeling to automatic speech recognition for Northern Sámi, focusing on filtering pseudo-labels through WER-based agreement between teacher-student models. This is fundamentally different from the original paper's frequency-based weighting scheme for multiple rollouts in reinforcement learning for reasoning tasks.

---

3. SuperST: Superficial Self-Training for Few-Shot Text Classification

URL: View paper

Brief Assessment

SuperST[59] focuses on text classification with binary confidence thresholding (high vs. low confidence pseudo-labels), not frequency-based weighting across multiple rollouts. The original paper's monotonic shaping function approach to weight pseudo-labels by their frequency is not present in SuperST[59].

---

4. A bearing fault detection method using pseudo-labeling CNN models and multiple frequency analysis

URL: View paper

Brief Assessment

Bearing Fault Detection[60] applies pseudo-labeling in a completely different domain (bearing fault detection using CNNs and frequency analysis) rather than reinforcement learning or language model reasoning tasks.

---

5. Pseudo-Labeling Based Domain Adaptation for Personality Mining

URL: View paper

Brief Assessment

Personality Mining[63] focuses on domain adaptation for personality mining tasks, not reinforcement learning or reasoning model training. The candidate's context is too limited to assess any pseudo-label weighting mechanism.

Contribution 3: Negative rollout penalization mechanism

Description: The authors introduce a penalization mechanism that explicitly penalizes all rollouts when majority consensus is very low, encouraging the model to explore alternative reasoning paths in unreliable supervision scenarios where no answer can be confidently trusted.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Distributed Data-Driven Inverse Reinforcement Learning for Multi-Agent Systems

URL: View paper

Brief Assessment

Distributed Inverse RL[53] focuses on inverse reinforcement learning for multi-agent systems using state trajectories and penalty terms in value functions, which is fundamentally different from the ORIGINAL paper's negative rollout penalization mechanism for low-consensus scenarios in language model reasoning.

---

2. Success-Rate Targeted Reinforcement Learning by Disorientation Penalty

URL: View paper

Brief Assessment

Disorientation Penalty[58] focuses on penalizing cycling behaviors in value function learning for undiscounted returns in traditional RL settings, not on penalizing rollouts in low-consensus scenarios during language model training with majority voting.

---

3. Reward Penalties on Augmented States for Solving Richly Constrained RL Effectively

URL: View paper

Brief Assessment

Augmented States Penalties[55] focuses on trajectory-level reward penalties in constrained RL with cost thresholds (e.g., safety constraints), not on penalizing rollouts based on low majority consensus in self-supervised reasoning tasks. The mechanisms serve fundamentally different purposes in different problem domains.

---

4. Actor-critic objective penalty function method: an adaptive strategy for trajectory tracking in autonomous driving

URL: View paper

Brief Assessment

Actor-Critic Penalty[51] addresses trajectory tracking in autonomous driving using penalty functions for constraint handling in MPC-based control, not reinforcement learning with rollout penalization for low-consensus scenarios in language model reasoning.

---

5. Enforcing the consensus between Trajectory Optimization and Policy Learning for precise robot control

URL: View paper

Brief Assessment

Trajectory Policy Consensus[56] focuses on robot control through consensus between trajectory optimization and policy learning, not on penalization mechanisms for low-consensus scenarios in reinforcement learning. The candidate addresses a fundamentally different problem domain (robotics optimal control) with different technical approaches (augmented Lagrangian, multiple shooting) than the original paper's self-penalization framework for reasoning models.

---

6. P2BPO: Permeable Penalty Barrier-Based Policy Optimization for Safe RL

URL: View paper

Brief Assessment

P2BPO[57] addresses safe reinforcement learning with constraint satisfaction using penalty barriers, not negative rollout penalization for low-consensus scenarios in reasoning tasks. The candidate focuses on safety constraints in MDPs, while the original addresses unreliable supervision in chain-of-thought reasoning.

---

7. RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization

URL: View paper

Brief Assessment

RESTRAIN Self-Driven[31] describes the same negative rollout penalization mechanism as the original paper. The candidate text mentions 'penalizing overconfident rollouts and low-consistency examples' which matches the original's description, confirming this is the same work rather than refuting prior art.

---

8. Constrained Reinforcement Learning-Enabled Policies With Augmented Lagrangian for Cooperative Intersection Management

URL: View paper

Brief Assessment

Intersection Management[54] focuses on constrained RL for autonomous vehicle traffic control at intersections, not on penalization mechanisms for low-consensus scenarios in language model reasoning tasks.

---

9. Wasserstein-Barycenter Consensus for Cooperative Multi-Agent Reinforcement Learning

URL: View paper

Brief Assessment

Wasserstein-Barycenter Consensus[52] addresses multi-agent coordination through optimal transport theory and barycenter computation, not rollout penalization in single-agent RL with low-consensus scenarios. The candidate focuses on cooperative MARL consensus mechanisms rather than penalization strategies for unreliable supervision.

Appendix: Text Similarity Detection

Textual similarity detection checked 23 papers and found 3 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. RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_3

⚠ **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] RESTRAIN: From Spurious Votes to Signals — Self-Training RL with Self-Penalization View paper
- [1] TTRL: Test-Time Reinforcement Learning View paper
- [2] RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback View paper
- [3] Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs View paper
- [4] Language Model Self-improvement by Reinforcement Learning Contemplation View paper
- [5] SSRL: Self-Search Reinforcement Learning View paper
- [6] SELFI: Autonomous Self-Improvement with Reinforcement Learning for Social Navigation View paper
- [7] Post-Training Large Language Models via Reinforcement Learning from Self-Feedback View paper
- [8] Co-Reward: Self-supervised Reinforcement Learning for Large Language Model Reasoning via Contrastive Agreement View paper
- [9] Autonomous Robotic Reinforcement Learning with Asynchronous Human Feedback View paper
- [10] RLCoder: Reinforcement Learning for Repository-Level Code Completion View paper
- [11] Training Language Models to Self-Correct via Reinforcement Learning View paper
- [12] SAIL: Self-Improving Efficient Online Alignment of Large Language Models View paper
- [13] GeoDRL: A Self-Learning Framework for Geometry Problem Solving using Reinforcement Learning in Deductive Reasoning View paper
- [14] RL-STaR: Theoretical Analysis of Reinforcement Learning Frameworks for Self-Taught Reasoner View paper
- [15] UniRL: Self-Improving Unified Multimodal Models via Supervised and Reinforcement Learning View paper
- [16] Self-Evolved Reward Learning for LLMs View paper
- [17] Self Rewarding Self Improving View paper
- [18] Reinforcement Learning of Self-enhancing Camera Image and Signal Processing View paper
- [19] WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning View paper
- [20] R-Zero: Self-Evolving Reasoning LLM from Zero Data View paper
- [21] SERL: Self-Examining Reinforcement Learning on Open-Domain View paper
- [22] Grow your limits: Continuous improvement with real-world rl for robotic locomotion View paper
- [23] Can Large Reasoning Models Self-Train? View paper
- [24] EvoTest: Evolutionary Test-Time Learning for Self-Improving Agentic Systems View paper
- [25] STRAPPER: Preference-based Reinforcement Learning via Self-training Augmentation and Peer Regularization View paper
- [26] Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach View paper
- [27] Bridging supervised learning and reinforcement learning in math reasoning View paper
- [28] Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation View paper
- [29] Meta-Learning Reinforcement Learning for Crypto-Return Prediction View paper
- [30] Preference VLM: Leveraging VLMs for Scalable Preference-Based Reinforcement Learning View paper
- [31] RESTRAIN: From Spurious Votes to Signals - Self-Driven RL with Self-Penalization View paper
- [32] SPELL: Self-Play Reinforcement Learning for evolving Long-Context Language Models View paper
- [33] Self-improvement in language models: The sharpening mechanism View paper
- [34] Truly Self-Improving Agents Require Intrinsic Metacognitive Learning View paper
- [35] West-of-N: Synthetic Preferences for Self-Improving Reward Models View paper
- [36] SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn Reinforcement Learning View paper
- [37] Badgr: An autonomous self-supervised learning-based navigation system View paper
- [38] Evolving Language Models without Labels: Majority Drives Selection, Novelty Promotes Variation View paper
- [39] ReST meets ReAct: Self-Improvement for Multi-Step Reasoning LLM Agent View paper
- [40] Autonomous Goal Detection and Cessation in Reinforcement Learning: A Case Study on Source Term Estimation View paper
- [41] Reinforcement Learning for Autonomous Self-Improving Robotic Systems View paper
- [42] Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection View paper
- [43] Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Recursive Self-Improvement through Reinforcement Learning View paper
- [44] Multi-view Disentanglement for Reinforcement Learning with Multiple Cameras View paper
- [45] Progress or regress? self-improvement reversal in post-training View paper
- [46] Shared autonomy via deep reinforcement learning View paper
- [47] S4rl: Surprisingly simple self-supervision for offline reinforcement learning in robotics View paper
- [48] The ingredients of real-world robotic reinforcement learning View paper
- [49] Removing the need for ground truth UWB data collection: self-supervised ranging error correction using deep reinforcement learning View paper
- [50] Augmenting unsupervised reinforcement learning with self-reference View paper
- [51] Actor-critic objective penalty function method: an adaptive strategy for trajectory tracking in autonomous driving View paper

- [52] Wasserstein-Barycenter Consensus for Cooperative Multi-Agent Reinforcement Learning View paper
- [53] Distributed Data-Driven Inverse Reinforcement Learning for Multi-Agent Systems View paper
- [54] Constrained Reinforcement Learning-Enabled Policies With Augmented Lagrangian for Cooperative Intersection Management View paper
- [55] Reward Penalties on Augmented States for Solving Richly Constrained RL Effectively View paper
- [56] Enforcing the consensus between Trajectory Optimization and Policy Learning for precise robot control View paper
- [57] P2BPO: Permeable Penalty Barrier-Based Policy Optimization for Safe RL View paper
- [58] Success-Rate Targeted Reinforcement Learning by Disorientation Penalty View paper
- [59] SuperST: Superficial Self-Training for Few-Shot Text Classification View paper
- [60] A bearing fault detection method using pseudo-labeling CNN models and multiple frequency analysis View paper
- [61] Semi-Supervised Clustering Framework for Fine-grained Scene Graph Generation View paper
- [62] Semi-Supervised Learning using Pseudo-Labels: A Case Study in Northern SÃ¡mi ASR View paper
- [63] Pseudo-Labeling Based Domain Adaptation for Personality Mining View paper
- [64] E2CL: Exploration-based Error Correction Learning for Embodied Agents View paper
- [65] Human-compatible driving partners through data-regularized self-play reinforcement learning View paper
- [66] Self-Supervised, Active Learning Seismic Full Waveform Inversion View paper
- [67] A Novel Route Planning Approach Based on Energy-Based Action Sample Reinforcement Learning View paper
- [68] A Lifetime Extended Energy Management Strategy for Fuel Cell Hybrid Electric Vehicles via Self-Learning Fuzzy Reinforcement Learning View paper
- [69] Road Detection for Reinforcement Learning Based Autonomous Car View paper
- [70] BEAR: Reinforcement Learning for Throughput Aware Borrowing in Energy Harvesting Systems View paper
- [71] RLSR: Reinforcement Learning from Self Reward View paper
- [72] Self-supervised boundary offline reinforcement learning View paper