Novelty Assessment Report

Paper: Online Rubrics Elicitation from Pairwise Comparisons
PDF URL: https://openreview.net/pdf?id=ebgsbC4x5W
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

Rubrics provide a flexible way to train LLMs on open-ended long-form answers where verifiable rewards are not applicable and human preferences provide coarse signals. Prior work shows that reinforcement learning with rubric-based rewards leads to consistent gains in LLM post-training. Most existing approaches rely on rubrics that remain static over the course of training. Such static rubrics, however, are vulnerable to reward-hacking type behaviors and fail to capture emergent desiderata that arise during training. We introduce Online Rubrics Elicitation (OnlineRubrics), a method that dynamically curates evaluation criteria in an online manner through pairwise comparisons of responses from current and reference policies. This online process enables continuous identification and mitigation of errors as training proceeds. Empirically, this approach yields consistent improvements of up to 8% over training exclusively with static rubrics across AlpacaEval, GPQA, ArenaHard as well as the validation sets of expert questions and rubrics. We qualitatively analyze the elicited criteria and identify prominent themes such as transparency, practicality, organization, and reasoning.

Core Task Landscape

This paper addresses: **Dynamic Rubric Elicitation from Pairwise Comparisons during Reinforcement Learning**

A total of **30 papers** were analyzed and organized into a taxonomy with **13 categories**.

Taxonomy Overview

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

- **Pairwise Preference-Based Reward Modeling**
- **Dynamic and Adaptive Reward Modeling**
- **Generative and Rule-Based Reward Modeling**
- **Feedback-Driven Optimization and Interaction**
- **Vision-Language and Multimodal Reinforcement Learning**
- **Specialized Domain Applications of Preference-Based RL**

Complete Taxonomy Tree

- Dynamic Rubric Elicitation from Pairwise Comparisons during Reinforcement Learning Survey Taxonomy
- Pairwise Preference-Based Reward Modeling
  - Pairwise Reward Model Training for RLHF
  - General Language Model Alignment via Pairwise Preferences (3 papers)
    - [2] Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (Bai Yuntao, 2022) View paper
    - [7] A Unified Pairwise Framework for RLHF: Bridging Generative Reward Modeling and Policy Optimization (Xu, 2025) View paper
    - [26] AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback (Dubois, 2023) View paper
  - Domain-Specific Pairwise Preference Applications (6 papers)
    - [1] Pref-GRPO: Pairwise Preference Reward-based GRPO for Stable Text-to-Image Reinforcement Learning (Wang Yi-bin, 2025) View paper
    - [6] LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning (Zhang Di, 2024) View paper
    - [12] LLaMA-Berry: Pairwise Optimization for Olympiad-level Mathematical Reasoning via O1-like Monte Carlo Tree Search (Di Zhang, 2025) View paper
    - [13] Posterior-GRPO: Rewarding Reasoning Processes in Code Generation (Zhang, 2025) View paper
    - [19] PaCo-RL: Advancing Reinforcement Learning for Consistent Image Generation with Pairwise Reward Modeling (Bowen Ping, 2025) View paper
    - [21] One Model to Critique Them All: Rewarding Agentic Tool-Use via Efficient Reasoning (Li Renhao, 2025) View paper
  - Efficient Pairwise Preference Elicitation (3 papers)
  - [16] Efficient Evaluation of LLMs via Branching Preference Learning (H Zhang, 2024) View paper
  - [22] Efficiently Acquiring Human Feedback with Bayesian Deep Learning (Fang, 2024) View paper
  - [30] ResponseRank: Data-Efficient Reward Modeling through Preference Strength Learning (T Kaufmann, n.d.) View paper
  - Pairwise Preference Modeling in Non-RL Contexts (4 papers)
  - [18] Scoring from pairwise winning indices (Arcidiacono, 2023) View paper
  - [24] Two of a kind or the ratings game? Adaptive pairwise preferences and latent factor models (Suhrid Balakrishnan, 2012) View paper
  - [25] Representation Learning and Pairwise Ranking for Implicit Feedback in Recommendation Systems (Sidana, 2022) View paper
  - [27] Pairwise comparison locomotion scoring for dairy cattle. (John Gardenier, 2021) View paper

- Dynamic and Adaptive Reward Modeling
  - Online Rubric and Criteria Elicitation ★ (2 papers)
  - [0] Online Rubrics Elicitation from Pairwise Comparisons (Anon et al., 2026) View paper
  - [29] Dynamic Evaluation of Reward Models via Pairwise Maximum Discrepancy Competition (S Luo, n.d.) View paper
  - Adaptive Preference Learning and Model Refinement (2 papers)
  - [23] Ext2Gen: Alignment through Unified Extraction and Generation for Robust Retrieval-Augmented Generation (Song, 2025) View paper
  - [28] Improving Reward Model Generalization from Adversarial Process Enhanced Preferences (Z Zhang, n.d.) View paper
- Generative and Rule-Based Reward Modeling
  - Generative Reward Models with Synthetic Reasoning (2 papers)
  - [4] Generative reward modeling via synthetic criteria preference learning (Xiaobo Liang, 2025) View paper
  - [17] PaTaRM: Bridging Pairwise and Pointwise Signals via Preference-Aware Task-Adaptive Reward Modeling (Jian Ai, 2025) View paper
  - Static Rubric and Rule-Based Reward Generation (2 papers)
  - [9] AutoRule: Reasoning Chain-of-thought Extracted Rule-based Rewards Improve Preference Learning (Xiong, 2025) View paper
  - [11] OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment (Liu Tian-ci, 2025) View paper
- Feedback-Driven Optimization and Interaction
  - Structured Feedback for Artifact Optimization (2 papers)
  - [3] Improving the Validity of Automatically Generated Feedback via Reinforcement Learning (Alexander Scarlatos, 2024) View paper
  - [14] Feedback descent: Open-ended text optimization via pairwise comparison (Yoonho Lee, 2025) View paper
  - Proactive Engagement and Clarification Seeking (1 papers)
  - [15] MACAROON: Training Vision-Language Models To Be Your Engaged Partners (Shujin Wu, 2024) View paper
- Vision-Language and Multimodal Reinforcement Learning
  - Vision-Language Foundation Models for Reward Feedback (1 papers)
  - [10] RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback (Wang Yu-Fei, 2024) View paper
  - Multimodal Cold-Start and Decoupling Strategies (1 papers)
  - [20] Metis-SPECS: Decoupling Multimodal Learning via Self-distilled Preference-based Cold Start (Chen Kun, 2025) View paper
- Specialized Domain Applications of Preference-Based RL (2 papers)
  - [5] Learning Interpretable Models of Aircraft Handling Behaviour by Reinforcement Learning from Human Feedback (Tom Bewley, 2023) View paper
  - [8] DeepMesh: Auto-Regressive Artist-mesh Creation with Reinforcement Learning (Zhao, 2025) View paper

Narrative

Core task: dynamic rubric elicitation from pairwise comparisons during reinforcement learning. The field has organized itself around several complementary directions. Pairwise Preference-Based Reward Modeling focuses on extracting reward signals directly from human or automated comparisons, often using Bradley-Terry models or ranking-based objectives (e.g., Helpful Harmless Assistant[2], AlpacaFarm[26]). Dynamic and Adaptive Reward Modeling emphasizes methods that update or refine reward structures online, including approaches that elicit criteria or rubrics interactively (e.g., OpenRubrics[11], Valid Feedback RL[3]). Generative and Rule-Based Reward Modeling explores using language models to produce interpretable reward functions or rules (e.g., Generative Reward Modeling[4], AutoRule[9]). Feedback-Driven Optimization and Interaction investigates iterative loops where agent behavior and human feedback co-evolve (e.g., Feedback Descent[14], LLaMA-Berry[6]). Vision-Language and Multimodal Reinforcement Learning extends preference learning to settings with visual or cross-modal inputs (e.g., RL-VLM-F[10]). Finally, Specialized Domain Applications of Preference-Based RL applies these techniques to robotics, aircraft handling, and other domains (e.g., Aircraft Handling RLHF[5]).

A particularly active line of work centers on making reward modeling more transparent and adaptive. Generative approaches like Generative Reward Modeling[4] and AutoRule[9] aim to produce human-readable criteria, while online elicitation methods such as OpenRubrics[11] and Valid Feedback RL[3] refine rubrics during training. Online Rubrics Elicitation[0] sits squarely within this dynamic and adaptive branch, emphasizing the interactive discovery of evaluation criteria from pairwise comparisons. Compared to OpenRubrics[11], which also targets rubric generation, Online Rubrics Elicitation[0] focuses more explicitly on the temporal dynamics of elicitation during the learning loop. Meanwhile, Valid Feedback RL[3] shares the goal of incorporating evolving human input but does not necessarily structure feedback as explicit rubrics. These contrasts highlight an open question: how to balance interpretability, adaptability, and sample efficiency when criteria themselves must be learned alongside policies.

Related Works in Same Category

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

1. Dynamic Evaluation of Reward Models via Pairwise Maximum Discrepancy Competition

Authors: S Luo, P Cao, Z Zhu, K Feng, Z Wang, et al. (6 authors total) | **URL**: View paper

Abstract

â¦ datasets with pre-annotated preference pairs, which often fail â¦ harmless assistant with reinforcement learning from human â¦ distinct preferences and evaluation criteria. These cases â¦

Relationship Analysis

Both papers belong to the Online Rubric and Criteria Elicitation category, focusing on dynamically generating evaluation criteria during training using pairwise comparisons. While the original paper (Online Rubrics Elicitation) directly elicits new rubric criteria from pairwise comparisons of policy rollouts to augment static rubrics during reinforcement learning, the candidate paper (PMDC) uses pairwise maximum discrepancy to evaluate and rank existing reward models rather than to generate new criteria for training. The key difference is that the original paper generates criteria online to improve policy training, whereas the candidate paper uses discrepancy-based selection to evaluate pre-existing reward models.

Contributions Analysis

Overall novelty summary. The paper introduces OnlineRubrics, a method for dynamically eliciting evaluation criteria during reinforcement learning by analyzing pairwise comparisons between current and reference policy outputs. It resides in the 'Online Rubric and Criteria Elicitation' leaf of the taxonomy, which contains only two papers total. This places the work in a relatively sparse research direction within the broader field of dynamic and adaptive reward modeling, suggesting that online rubric generation during training remains an underexplored area compared to static reward modeling approaches.

The taxonomy reveals that the paper sits within 'Dynamic and Adaptive Reward Modeling,' which contrasts with neighboring branches like 'Pairwise Preference-Based Reward Modeling' (static Bradley-Terry models) and 'Generative and Rule-Based Reward Modeling' (fixed rule extraction). The sibling paper in the same leaf, OpenRubrics, also targets rubric generation but differs in temporal dynamics emphasis. Adjacent leaves include 'Adaptive Preference Learning and Model Refinement' (iterative reward updates) and 'Static Rubric and Rule-Based Reward Generation' (fixed criteria extraction), highlighting the paper's focus on continuous, online elicitation rather than one-shot or static approaches.

Among 23 candidates examined across three contributions, none were found to clearly refute the proposed work. The core OnlineRubrics method examined 10 candidates with zero refutable overlaps, the dataset contribution examined 3 candidates with no refutations, and the formal gradient variance motivation examined 10 candidates with no refutations. This limited search scope suggests that within the top semantic matches and citation expansions analyzed, no prior work directly anticipates the combination of online rubric elicitation with pairwise comparison-driven criteria discovery during RL training.

Based on the limited literature search of 23 candidates, the work appears to occupy a novel position at the intersection of dynamic reward modeling and interpretable criteria generation. The sparse taxonomy leaf and absence of refutable prior work within the examined scope indicate potential originality, though a broader search beyond top-K semantic matches might reveal additional related efforts in adaptive evaluation or online preference learning.

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

Contribution 1: Online Rubrics Elicitation (OnlineRubrics) method

Description: A framework that dynamically elicits new evaluation criteria during reinforcement learning by comparing responses from the current policy and a control policy. This enables continuous identification and mitigation of errors as training proceeds, addressing limitations of static rubrics that are vulnerable to reward-hacking and fail to capture emergent desiderata.

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. Carrot and stick: Eliciting comparison data and beyond

URL: View paper

Brief Assessment

Carrot and Stick[39] focuses on truthfully eliciting pairwise comparison data through peer prediction mechanisms using bonus-penalty payments, not on dynamically eliciting evaluation criteria during reinforcement learning for LLM training.

---

2. T2I-Eval-R1: Reinforcement Learning-Driven Reasoning for Interpretable Text-to-Image Evaluation

URL: View paper

Brief Assessment

T2I-Eval-R1[40] focuses on text-to-image evaluation using reinforcement learning with coarse-grained quality scores, not on dynamically eliciting evaluation criteria through pairwise comparisons during RL training as in the original paper's OnlineRubrics method.

---

3. Relatively rational: Learning utilities and rationalities jointly from pairwise preferences

URL: View paper

Brief Assessment

Relatively Rational[33] focuses on jointly learning utilities and rationality coefficients from pairwise preferences in ranking and action-advice settings, not on dynamically eliciting evaluation criteria during reinforcement learning as OnlineRubrics does.

---

4. A Large Language Model-Driven Reward Design Framework via Dynamic Feedback for Reinforcement Learning

URL: View paper

Brief Assessment

LLM Reward Design[34] focuses on generating reward function code for RL tasks through iterative code generation and trajectory preference evaluation, not on dynamically eliciting evaluation criteria through pairwise comparisons of policy responses during training.

---

5. Writing-RL: Advancing Long-form Writing via Adaptive Curriculum Reinforcement Learning

URL: View paper

Brief Assessment

Writing-RL[35] focuses on adaptive curriculum RL for long-form writing with pairwise comparison rewards, not on dynamically eliciting evaluation criteria through pairwise comparisons during RL as described in the original paper's OnlineRubrics framework.

---

6. Improving reinforcement learning from human feedback using contrastive rewards

URL: View paper

Brief Assessment

Contrastive Rewards[38] focuses on improving reward model robustness in RLHF through contrastive penalty terms calculated from baseline responses, not on dynamically eliciting evaluation criteria through pairwise comparisons during training.

---

7. Reward learning from multiple feedback types

URL: View paper

Brief Assessment

Multiple Feedback Types[36] focuses on learning reward functions from diverse feedback types (ratings, comparisons, demonstrations, corrections) in RL environments, not on dynamically eliciting evaluation criteria through pairwise comparisons during training as OnlineRubrics does.

---

8. A survey of reinforcement learning from human feedback

URL: View paper

Brief Assessment

RLHF Survey[32] provides a broad overview of reinforcement learning from human feedback methods but does not describe dynamic criteria elicitation during training through pairwise comparisons as proposed in OnlineRubrics.

---

9. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons

URL: View paper

Brief Assessment

K-wise Comparisons[31] focuses on theoretical foundations for learning reward models from pairwise/k-wise comparisons in RLHF, not on dynamically eliciting evaluation criteria during training. The paper addresses reward model convergence and policy optimization, which is a different problem from online rubric elicitation.

---

10. Pref-GUIDE: Continual Policy Learning from Real-Time Human Feedback via Preference-Based Learning

URL: View paper

Brief Assessment

Pref-GUIDE[37] focuses on converting real-time scalar human feedback into preference-based data for RL policy learning, not on dynamically eliciting evaluation criteria through pairwise comparisons during training as OnlineRubrics does.

Contribution 2: Two curated datasets for expert and generalist domains

Description: Two rubric-based datasets (Generalist Rubrics and Expert Rubrics) containing prompts with human-authored, prompt-specific rubrics composed of weighted, binary-checkable criteria. These datasets enable training and evaluation of rubric-based reinforcement learning methods across different domains.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Rubric-Guided Lightweight Large Language Model Annotation of Patient Medication Reviews: Ordinal Agreement, Uncertainty, and Downstream Learnability

URL: View paper

Brief Assessment

Rubric-Guided Annotation[43] focuses on using LLMs to annotate patient medication reviews with rubrics for evaluation purposes, not on creating rubric-based datasets for training reinforcement learning methods across different domains.

---

2. A Scalable Framework for Evaluating Health Language Models

URL: View paper

Brief Assessment

Health LLM Framework[41] focuses on evaluation methodology using rubrics for health-related queries, not on creating training datasets with rubrics for reinforcement learning across expert and generalist domains.

---

3. A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains

URL: View paper

Brief Assessment

Medical LLM Benchmark[42] focuses on clinical safety-effectiveness evaluation with binary/graded scoring criteria for medical LLMs, not on rubric-based reinforcement learning training datasets. The datasets serve different purposes: evaluation vs. RL training.

Contribution 3: Formal motivation showing gradient variance reduction

Description: A theoretical result (Proposition 1) demonstrating that augmenting rubrics to better approximate the true criterion set reduces the variance term in policy gradient updates, leading to improved stability and sample efficiency during training by tightening the upper bound on unmodeled criteria mass.

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. BiVWAC: Improving deep reinforcement learning algorithms using Bias-Variance Weighted Actor-Critic

URL: View paper

Brief Assessment

BiVWAC[53] focuses on bias-variance weighting in critic losses for actor-critic algorithms in continuous control tasks, not on gradient variance reduction through augmenting evaluation criteria in policy gradient methods as described in the original paper's rubric-based framework.

---

2. PPO-BR: Dual-Signal Entropy-Reward Adaptation for Trust Region Policy Optimization

URL: View paper

Brief Assessment

PPO-BR[49] focuses on adaptive trust region mechanisms through entropy-reward signals in PPO, not on gradient variance reduction through augmenting evaluation criteria in rubric-based policy gradient methods.

---

3. Reimagining Exploration: Theoretical Insights and Practical Advancements in Policy Gradient Methods

URL: View paper

Brief Assessment

Exploration Policy Gradient[46] focuses on policy stochasticity and exploration bonuses in general RL settings, not on rubric-based reward modeling or variance reduction through augmenting evaluation criteria as in the original paper.

---

4. ISTANBUL TECHNICAL UNIVERSITY å» GRADUATE SCHOOL

URL: View paper

Brief Assessment

Istanbul Technical University[52] focuses on value gradient methods on Stiefel manifolds for robotic control, not on rubric-based reward modeling or gradient variance reduction through augmenting evaluation criteria in policy gradient methods.

---

5. KIPPO: Koopman-Inspired Proximal Policy Optimization

URL: View paper

Brief Assessment

KIPPO[48] focuses on reducing gradient variance through Koopman-inspired latent space linearization in policy optimization, while the original paper addresses variance reduction through augmenting evaluation rubrics in rubric-based reinforcement learning. These are fundamentally different technical approaches to variance reduction in distinct RL settings.

---

6. AlgaeDICE: Policy Gradient from Arbitrary Experience

URL: View paper

Brief Assessment

AlgaeDICE[44] focuses on off-policy policy gradient methods without importance weighting, not on variance reduction through augmenting evaluation criteria in rubric-based systems.

---

7. Scalable Robot Learning

URL: View paper

Brief Assessment

Scalable Robot Learning[51] mentions variance reduction in the context of importance sampling and auxiliary objectives for robot learning, which is a different technical domain and mechanism than the original paper's theoretical result about gradient variance reduction through augmenting rubrics in policy gradient methods for LLM training.

---

8. Enhancing Agentic RL with Progressive Reward Shaping and Value-based Sampling Policy Optimization

URL: View paper

Brief Assessment

Progressive Reward Shaping[50] focuses on curriculum-based reward design for tool-integrated reasoning tasks, not on theoretical analysis of gradient variance reduction in policy gradient methods through augmenting evaluation criteria.

---

9. Variance aware reward smoothing for deep reinforcement learning

URL: View paper

Brief Assessment

Variance Aware Smoothing[45] focuses on reward smoothing techniques to reduce variance in rewards themselves, whereas the original paper's Proposition 1 addresses variance reduction in policy gradient updates through augmenting evaluation criteria (rubrics). These are fundamentally different mechanisms operating at different stages of the RL pipeline.

---

10. Learning to Balance Lead Bias in News Summarization

URL: View paper

Brief Assessment

Lead Bias Summarization[47] focuses on news summarization and lead bias issues, not on policy gradient methods or gradient variance reduction through augmenting evaluation criteria in reinforcement learning contexts.

Appendix: Text Similarity Detection

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

References

- [0] Online Rubrics Elicitation from Pairwise Comparisons View paper
- [1] Pref-GRPO: Pairwise Preference Reward-based GRPO for Stable Text-to-Image Reinforcement Learning View paper
- [2] Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback View paper
- [3] Improving the Validity of Automatically Generated Feedback via Reinforcement Learning View paper
- [4] Generative reward modeling via synthetic criteria preference learning View paper
- [5] Learning Interpretable Models of Aircraft Handling Behaviour by Reinforcement Learning from Human Feedback View paper
- [6] LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning View paper
- [7] A Unified Pairwise Framework for RLHF: Bridging Generative Reward Modeling and Policy Optimization View paper
- [8] DeepMesh: Auto-Regressive Artist-mesh Creation with Reinforcement Learning View paper
- [9] AutoRule: Reasoning Chain-of-thought Extracted Rule-based Rewards Improve Preference Learning View paper
- [10] RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback View paper
- [11] OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment View paper
- [12] LLaMA-Berry: Pairwise Optimization for Olympiad-level Mathematical Reasoning via O1-like Monte Carlo Tree Search View paper
- [13] Posterior-GRPO: Rewarding Reasoning Processes in Code Generation View paper
- [14] Feedback descent: Open-ended text optimization via pairwise comparison View paper
- [15] MACAROON: Training Vision-Language Models To Be Your Engaged Partners View paper
- [16] Efficient Evaluation of LLMs via Branching Preference Learning View paper
- [17] PaTaRM: Bridging Pairwise and Pointwise Signals via Preference-Aware Task-Adaptive Reward Modeling View paper
- [18] Scoring from pairwise winning indices View paper
- [19] PaCo-RL: Advancing Reinforcement Learning for Consistent Image Generation with Pairwise Reward Modeling View paper
- [20] Metis-SPECS: Decoupling Multimodal Learning via Self-distilled Preference-based Cold Start View paper
- [21] One Model to Critique Them All: Rewarding Agentic Tool-Use via Efficient Reasoning View paper
- [22] Efficiently Acquiring Human Feedback with Bayesian Deep Learning View paper
- [23] Ext2Gen: Alignment through Unified Extraction and Generation for Robust Retrieval-Augmented Generation View paper
- [24] Two of a kind or the ratings game? Adaptive pairwise preferences and latent factor models View paper
- [25] Representation Learning and Pairwise Ranking for Implicit Feedback in Recommendation Systems View paper
- [26] AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback View paper
- [27] Pairwise comparison locomotion scoring for dairy cattle. View paper
- [28] Improving Reward Model Generalization from Adversarial Process Enhanced Preferences View paper
- [29] Dynamic Evaluation of Reward Models via Pairwise Maximum Discrepancy Competition View paper
- [30] ResponseRank: Data-Efficient Reward Modeling through Preference Strength Learning View paper
- [31] Principled reinforcement learning with human feedback from pairwise or k-wise comparisons View paper
- [32] A survey of reinforcement learning from human feedback View paper
- [33] Relatively rational: Learning utilities and rationalities jointly from pairwise preferences View paper
- [34] A Large Language Model-Driven Reward Design Framework via Dynamic Feedback for Reinforcement Learning View paper
- [35] Writing-RL: Advancing Long-form Writing via Adaptive Curriculum Reinforcement Learning View paper
- [36] Reward learning from multiple feedback types View paper
- [37] Pref-GUIDE: Continual Policy Learning from Real-Time Human Feedback via Preference-Based Learning View paper
- [38] Improving reinforcement learning from human feedback using contrastive rewards View paper

- [39] Carrot and stick: Eliciting comparison data and beyond View paper
- [40] T2I-Eval-R1: Reinforcement Learning-Driven Reasoning for Interpretable Text-to-Image Evaluation View paper
- [41] A Scalable Framework for Evaluating Health Language Models View paper
- [42] A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains View paper
- [43] Rubric-Guided Lightweight Large Language Model Annotation of Patient Medication Reviews: Ordinal Agreement, Uncertainty, and Downstream Learnability View paper
- [44] AlgaeDICE: Policy Gradient from Arbitrary Experience View paper
- [45] Variance aware reward smoothing for deep reinforcement learning View paper
- [46] Reimagining Exploration: Theoretical Insights and Practical Advancements in Policy Gradient Methods View paper
- [47] Learning to Balance Lead Bias in News Summarization View paper
- [48] KIPPO: Koopman-Inspired Proximal Policy Optimization View paper
- [49] PPO-BR: Dual-Signal Entropy-Reward Adaptation for Trust Region Policy Optimization View paper
- [50] Enhancing Agentic RL with Progressive Reward Shaping and Value-based Sampling Policy Optimization View paper
- [51] Scalable Robot Learning View paper
- [52] ISTANBUL TECHNICAL UNIVERSITY å» GRADUATE SCHOOL View paper
- [53] BiVWAC: Improving deep reinforcement learning algorithms using Bias-Variance Weighted Actor-Critic View paper