Novelty Assessment Report

Paper: Hybrid Reinforcement: when reward is sparse, better to be dense

PDF URL: https://openreview.net/pdf?id=0CajQNVKyB

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

Post-training for reasoning in large language models has increasingly relied on verifiable rewards: deterministic checkers that provide 0–1 correctness signals. While reliable, such binary feedback is brittle—many tasks admit partially correct or alternative answers that verifiers under-credit, and the resulting all-or-nothing supervision limits learning. Reward models offer richer, continuous feedback, which can serve as a complementary supervisory signal to verifiers. We introduce HERO (Hybrid Ensemble Reward Optimization), a reinforcement learning framework that integrates sparse verifier signals with dense reward model scores in a structured way. HERO employs stratified normalization to bound reward-model scores within verifier-defined groups, preserving correctness while refining quality distinctions, and variance-aware weighting to emphasize challenging prompts where dense signals matter most. Across diverse mathematical reasoning benchmarks, HERO consistently outperforms reward model-only and verifier-only baselines, with strong gains on both verifiable and hard-to-verify tasks. Our results show that hybrid reward design retains the stability of verifiers while leveraging the nuance of reward models to advance reasoning.

Disclaimer

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

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

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

Core Task Landscape

This paper addresses: **Integrating Sparse Verifiable Rewards with Dense Reward Model Signals for Reasoning**

A total of **43 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Hybrid Reward Integration Frameworks**
- **Dense Reward Model Design and Training**
- **Sparse Verifiable Reward Optimization**
- **Reward Model Evaluation and Design Principles**
- **Domain-Specific Reward Integration Applications**
- **Advanced RL Optimization Techniques**
- **Auxiliary Topics and Methodological Studies**

Complete Taxonomy Tree

- Integrating Sparse Verifiable Rewards with Dense Reward Model Signals for Reasoning Survey Taxonomy
- Hybrid Reward Integration Frameworks
  - Stratified Normalization and Weighting Schemes ★ (2 papers)
  - [0] Hybrid Reinforcement: when reward is sparse, better to be dense (Anon et al., 2026) View paper
  - [22] Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense (Tao, 2025) View paper
  - Multi-Stage Dense-to-Sparse Reward Transitions (2 papers)
  - [30] From Sparse to Dense: Toddler-inspired Reward Transition in Goal-Oriented Reinforcement Learning (Junseok Park, 2024) View paper
  - [34] D2SR: Transferring Dense Reward Function to Sparse by Network Resetting (Yongle Luo, 2023) View paper
- Dense Reward Model Design and Training
  - Process-Level Reward Models (4 papers)
  - [15] Your Reward Function for RL is Your Best PRM for Search: Unifying RL and Search-Based TTS (Jin Can, 2025) View paper
  - [16] Process supervision-guided policy optimization for code generation (Dai Ning, 2024) View paper
  - [25] PROPA: Toward Process-level Optimization in Visual Reasoning via Reinforcement Learning (Yanbei Jiang, 2025) View paper
  - [28] Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Reward Design (Wei Quan, 2025) View paper
  - Generative and Self-Supervised Dense Reward Learning (3 papers)
  - [14] MedGR: Breaking the Data Barrier for Medical Reasoning via Generative Reward Learning (W Zhi, 2025) View paper
  - [21] DrS: Learning Reusable Dense Rewards for Multi-Stage Tasks (Mu, 2024) View paper
  - [38] Self-Supervised Online Reward Shaping in Sparse-Reward Environments (Memarian, 2021) View paper
  - Heuristic-Based Dense Reward Generation (2 papers)
  - [19] Reinforcement Learning for Classical Planning: Viewing Heuristics as Dense Reward Generators (Gehring, 2021) View paper
  - [20] Driving Beyond Privilege: Distilling Dense-Reward Knowledge into Sparse-Reward Policies (Feeza Khan Khanzada, 2025) View paper
  - Vision-Language Dense Reward Models (2 papers)
  - [8] Self-rewarding vision-language model via reasoning decomposition (Li, 2025) View paper
  - [12] A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning (Zhai Shaopeng, 2025) View paper
- Sparse Verifiable Reward Optimization
  - Outcome-Based Reinforcement Learning (2 papers)

  - [1] Exploring the limit of outcome reward for learning mathematical reasoning (Lyu, 2025) View paper
  - [2] Teaching Large Language Models to Reason with Reinforcement Learning (Havrilla, 2024) View paper
  - Sparse Reward Exploration and Stability (3 papers)
  - [3] RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning (Feng Sicheng, 2025) View paper
  - [13] Don't Just Fine-tune the Agent, Tune the Environment (Lu, 2025) View paper
  - [24] A Practitioner's Guide to Multi-turn Agentic Reinforcement Learning (Wang Rui-yi, 2025) View paper
- Reward Model Evaluation and Design Principles (3 papers)
  - [5] On designing effective rl reward at training time for llm reasoning (Gao, 2024) View paper
  - [32] Deep Model-Based Reinforcement Learning for Predictive Control of Robotic Systems with Dense and Sparse Rewards (Luke Antonyshyn, 2024) View paper
  - [33] A Study on Dense and Sparse (Visual) Rewards in Robot Policy Learning (Abdalkarim Mohtasib, 2021) View paper
- Domain-Specific Reward Integration Applications
  - Embodied and Robotic Control (3 papers)
  - [6] Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning (Chen Han-yang, 2025) View paper
  - [7] Conrft: A reinforced fine-tuning method for vla models via consistency policy (Yu-Hui Chen, 2025) View paper
  - [29] RobotKeyframing: Learning Locomotion with High-Level Objectives via Mixture of Dense and Sparse Rewards (Zargarbashi, 2024) View paper
  - Multi-Agent and Multi-Task Systems (2 papers)
  - [9] Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning (B Liu, 2023) View paper
  - [26] GRADE: Personalized Multi-Task Fusion via Group-relative Reinforcement Learning with Adaptive Dirichlet Exploration (Hong, 2025) View paper
  - Specialized Reasoning Domains (3 papers)
  - [10] MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models (Binghui Guo, 2025) View paper
  - [23] TaoSR-SHE: Stepwise Hybrid Examination Reinforcement Learning Framework for E-commerce Search Relevance (Jin Yi-ming, 2025) View paper
  - [41] Process-Verified Reinforcement Learning for Theorem Proving via Lean (M Kim, n.d.) View paper
  - Search and Retrieval Agents (1 papers)
  - [11] Repurposing Synthetic Data for Fine-grained Search Agent Supervision (Zhao, 2025) View paper
- Advanced RL Optimization Techniques
  - Policy Gradient and GRPO Enhancements (2 papers)
  - [4] Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search (K. Laurent, 2024) View paper
  - [18] Enhancing Agentic RL with Progressive Reward Shaping and Value-based Sampling Policy Optimization (Zhuoran Zhuang, 2025) View paper
- Auxiliary Topics and Methodological Studies (10 papers)
  - [17] Dense Rewards and Continual Reinforcement Learning for Task-oriented Dialogue Policies (Geishauser, 2024) View paper
  - [27] Recovery of partly sparse and dense signals (Izuru Miyazaki, 2023) View paper
  - [31] Generative and Pseudo-Relevant Feedback for Sparse, Dense and Learned Sparse Retrieval (Mackie, 2023) View paper
  - [35] Curiosity and exploration in mazes (dense vs. sparse rewards) (Gopnik, 2021) View paper
  - [36] Automated Ligand Design in Simulated Molecular Docking (Rob Maccallum, 2022) View paper
  - [37] Sparse and Dense Approaches for the Full-rank Retrieval of Responses for Dialogues (Penha, 2022) View paper
  - [39] Multi-Slice Dense-Sparse Learning for Efficient Liver and Tumor Segmentation. (Zhao, 2021) View paper
  - [40] Reinforcement Learning from Human Feedback to Fine-Tune Vision Models (Sompura, n.d.) View paper
  - [42] SMARTER NOT HARDER: GENERATIVE PROCESS EVALUATION WITH INTRINSIC-SIGNAL DRIVING AND ABILITY-ADAPTIVE REWARD SHAPING (SHAPING, n.d.) View paper
  - [43] REINFORCEMENT LEARNING (Ann Russell, n.d.) View paper

Narrative

Core task: Integrating sparse verifiable rewards with dense reward model signals for reasoning. This field addresses a fundamental challenge in training reasoning agents: how to combine infrequent but reliable outcome signals (such as final answer correctness) with continuous learned feedback that guides intermediate steps. The taxonomy reveals several complementary research directions. Hybrid Reward Integration Frameworks explore architectural strategies for blending sparse and dense signals, including normalization schemes and weighting methods exemplified by works like Hybrid Reinforcement[0] and Hybrid Reinforcement Dense[22]. Dense Reward Model Design focuses on training process reward models and step-level critics, as seen in Teaching LLMs Reason[2] and Dense Reward MCTS[4]. Sparse Verifiable Reward Optimization investigates outcome supervision and verification strategies, while Reward Model Evaluation examines design principles such as those discussed in Designing RL Reward[5]. Domain-Specific Applications demonstrate these techniques across reasoning tasks, dialogue systems, and embodied agents, with Advanced RL Optimization covering policy gradient methods and search-based approaches.

A central tension emerges between relying on dense learned rewards—which provide rich training signals but may introduce bias or reward hacking—and sparse verifiable outcomes that are trustworthy but offer limited guidance during exploration. Recent work explores various middle grounds: some studies like RewardMap[3] investigate mapping strategies between reward types, while others such as Outcome Reward Limit[1] examine the boundaries of outcome-only supervision. Hybrid Reinforcement[0] sits within the stratified normalization branch, focusing on principled methods to balance and weight heterogeneous reward sources during training. This approach contrasts with purely dense methods like Dense Reward MCTS[4] that emphasize continuous guidance, and differs from outcome-focused strategies by explicitly addressing how to normalize and combine signals of different sparsity levels. The positioning reflects ongoing efforts to retain the reliability of verifiable rewards while leveraging dense models to accelerate learning and improve sample efficiency in complex reasoning domains.

Related Works in Same Category

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

1. Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense

Authors: Tao, Leitian, Kulikov, Ilia, Leitian Tao, et al. (20 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Post-training for reasoning of large language models (LLMs) increasingly relies on verifiable rewards: deterministic checkers that provide 0-1 correctness signals. While reliable, such binary feedback is brittle--many tasks admit partially correct or alternative answers that verifiers under-credit, and the resulting all-or-nothing supervision limits learning. Reward models offer richer, continuous feedback, which can serve as a complementary supervisory signal to verifiers. We introduce HERO (Hy...

⚠ Similarity Notice

These papers share nearly identical titles, abstracts, methodology (HERO framework with stratified normalization and variance-aware weighting), experimental setups, and results tables. The content, structure, and technical contributions are essentially the same, indicating they are likely the same paper or very close variants (e.g., different submission versions).

Contributions Analysis

Overall novelty summary. The paper introduces HERO, a framework that combines sparse verifier signals with dense reward model scores through stratified normalization and variance-aware weighting. Within the taxonomy, it resides in the 'Stratified Normalization and Weighting Schemes' leaf under 'Hybrid Reward Integration Frameworks'. This leaf contains only two papers, indicating a relatively sparse research direction focused specifically on structured integration mechanisms that normalize rewards within verifier-defined groups. The positioning suggests the work addresses a targeted gap in hybrid reward design rather than entering a crowded subfield.

The taxonomy reveals that HERO's parent branch—Hybrid Reward Integration Frameworks—sits alongside Dense Reward Model Design (with four subtopics including process-level and generative approaches) and Sparse Verifiable Reward Optimization (covering outcome-based RL and exploration challenges). Neighboring leaves include 'Multi-Stage Dense-to-Sparse Reward Transitions', which explores temporal curriculum strategies rather than static integration. The taxonomy's scope notes clarify that HERO's structured normalization distinguishes it from general hybrid methods and from purely dense or sparse approaches, positioning it at the intersection of reliability-focused verification and richness-focused learned feedback.

Among the three contributions analyzed, the HERO framework and stratified normalization show no clear refutation across ten and two candidates examined respectively. However, the variance-aware weighting mechanism encountered four refutable candidates among ten examined, suggesting this component has more substantial prior exploration. The analysis examined twenty-two total candidates from top-K semantic search, a limited scope that captures nearby work but does not constitute exhaustive coverage. The statistics indicate that while the overall framework appears novel within this search scope, the weighting mechanism builds on more established techniques for emphasizing challenging examples in RL training.

Based on the limited search scope of twenty-two candidates, the work appears to occupy a relatively underexplored niche within hybrid reward integration. The stratified normalization approach shows stronger novelty signals than the weighting mechanism, which has more documented precedents. The taxonomy structure confirms that structured integration methods remain less densely populated than pure dense or sparse reward approaches, though the analysis cannot rule out relevant work outside the top-K semantic matches examined.

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

Contribution 1: HERO framework for hybrid reward optimization

Description: The authors propose a reinforcement learning framework that combines binary verifier signals with continuous reward model scores through stratified normalization and variance-aware weighting. This approach preserves correctness guarantees from verifiers while exploiting nuanced quality distinctions from reward models.

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. Discriminative reward co-training

URL: View paper

Brief Assessment

Discriminative Reward Co-training[45] focuses on self-imitation learning using a discriminator to distinguish between current policy trajectories and beneficial past trajectories stored in a buffer. In contrast, HERO combines binary verifier signals with continuous reward model scores through stratified normalization for mathematical reasoning tasks. The candidate does not address verifier-reward model integration or mathematical reasoning domains.

2. Learning to Explore in Diverse Reward Settings via Temporal-Difference-Error Maximization

URL: View paper

Brief Assessment

TD-Error Maximization[48] focuses on exploration strategies through temporal-difference error maximization in diverse reward settings (dense, sparse, exploration-adverse), not on combining verifier signals with reward models for reasoning tasks.

3. A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning

URL: View paper

Brief Assessment

Vision Language Action Critic[12] focuses on robotic manipulation with vision-language-action models and process rewards for real-world RL, not on combining sparse verifier signals with dense reward models for mathematical reasoning in LLMs.

4. Process supervision-guided policy optimization for code generation

URL: View paper

Brief Assessment

Process Supervision Policy[16] focuses on integrating process reward models (PRMs) into RL for code generation, providing line-level feedback during code generation. The ORIGINAL paper addresses combining sparse verifier signals with dense reward model scores for mathematical reasoning through stratified normalization and variance-aware weighting. These are distinct technical approaches in different domains (code vs. math reasoning).

5. Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search

URL: View paper

Brief Assessment

Dense Reward MCTS[4] focuses on Monte Carlo Tree Search with dense rewards for reasoning, not on combining binary verifier signals with continuous reward models through stratified normalization as HERO does.

6. Reward Generation via Large Vision-Language Model in Offline Reinforcement Learning

URL: View paper

Brief Assessment

Reward via Vision-Language[44] focuses on automated reward generation using vision-language models for offline RL in visual domains, not on combining binary verifiers with continuous reward models in language model reasoning tasks.

7. Teaching Large Language Models to Reason with Reinforcement Learning

URL: View paper

Brief Assessment

Teaching LLMs Reason[2] investigates multiple RL algorithms (expert iteration, PPO, return-conditioned RL) with both sparse and dense rewards but does not propose a structured framework for combining verifier signals with reward models through stratified normalization and variance-aware weighting as HERO does.

8. Rubrics as rewards: Reinforcement learning beyond verifiable domains

URL: View paper

Brief Assessment

Rubrics as Rewards[47] focuses on using instance-specific rubrics as reward signals for unstructured real-world reasoning tasks (medical, science domains), while HERO addresses combining binary verifiers with continuous reward models specifically for mathematical reasoning with verifiable outcomes. The technical approaches and target domains differ fundamentally.

9. Enhancing RLHF with Human Gaze Modeling

URL: View paper

Brief Assessment

RLHF Gaze Modeling[46] focuses on integrating human gaze patterns into RLHF through dense reward distribution and gaze-informed reward models, not on combining binary verifier signals with continuous reward model scores for mathematical reasoning tasks.

10. RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning

URL: View paper

Brief Assessment

RewardMap[3] focuses on visual reasoning tasks (transit maps) using detail rewards for partial correctness in multi-stage RL, not on combining binary verifier signals with continuous reward models for mathematical reasoning as in the original paper.

Contribution 2: Stratified normalization for reward integration

Description: A technique that rescales continuous reward model scores within correctness groups defined by binary verifiers. This ensures dense feedback refines learning only within verified correct responses, maintaining correctness semantics while adding gradations.

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

1. Neural Signatures Within and Between Chess Puzzle Solving and Standard Cognitive Tasks for Brain-Computer Interfaces: A Low-Cost Electroencephalography â¦

URL: View paper

Brief Assessment

Chess Puzzle BCI[58] focuses on EEG-based brain-computer interfaces for cognitive workload classification during chess puzzle solving and standard cognitive tasks. This work does not address reinforcement learning, reward modeling, or the integration of binary correctness signals with continuous reward scores for language model training.

2. Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense

URL: View paper

Brief Assessment

Hybrid Reinforcement Dense[22] is the same paper as the original submission. Both describe identical stratified normalization techniques with the same mathematical formulation (Equation 3) and implementation details. This is not a prior work but the same work.

Contribution 3: Variance-aware weighting mechanism

Description: An adaptive reweighting scheme that adjusts the contribution of different prompts during training based on reward-model score variance. It emphasizes harder prompts with high variance while down-weighting easy prompts with uniform responses.

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. Reward-Guided Prompt Evolving in Reinforcement Learning for LLMs

URL: View paper

Brief Assessment

Reward-Guided Prompt[57] uses variance of reward-model scores across candidate responses to weight prompts during training, but this is applied to prompt selection/evolution rather than the ORIGINAL paper's context of hybrid reward design that combines verifier and reward-model signals within correctness groups.

2. Comparing Comparisons: Informative and Easy Human Feedback with Distinguishability Queries

URL: View paper

Brief Assessment

Distinguishability Queries[53] uses variance-based informativeness for query selection in RLHF, not for adaptive reweighting of prompts during training. The variance measure (equation 5) selects which queries to present to humans, whereas the original paper's mechanism adjusts training contributions of different prompts based on reward-model score variance across candidate responses.

3. Foundations for Efficient and Near-Optimal Reinforcement Learning

URL: View paper

Brief Assessment

Efficient Near-Optimal RL[52] employs variance-aware weighting for function approximation in large state-action spaces to achieve optimal worst-case regret, whereas the original paper uses variance-aware weighting to emphasize challenging prompts during LLM training based on reward-model score variance across responses.

4. Reinforce-ada: An adaptive sampling framework for reinforce-style llm training

URL: View paper

Prior Art Analysis

Reinforce-ada[49] demonstrates prior work on variance-aware weighting mechanisms that adjust contributions based on prompt difficulty. The candidate paper introduces a variance-aware weighting scheme that emphasizes prompts with high reward-model score variance, which is conceptually similar to the original paper's approach of emphasizing harder prompts with high variance while down-weighting easy prompts. Both papers use variance as a signal for prompt difficulty and adaptively weight training contributions accordingly. The candidate's framework explicitly allocates sampling budgets based on pass rate variance (difficulty estimation), implementing weights like 1/√p_i that prioritize difficult prompts—the same principle underlying the original's variance-aware mechanism.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly describe down-weighting easy prompts and emphasizing harder prompts based on variance/difficulty metrics to improve learning signal quality. - **Original**: easy prompts, where most responses are uniformly correct or incorrect, contribute little additional learning signal and are down-weighted. in contrast, harder prompts-where candidate responses vary widely and reward-model scores provide valuable discrimination-are emphasized. - **Candidate**: to effectively balance cost and signal discovery, we must prioritize prompts based on their pass rates. this motivates us to consider a non-linear objectivej f(θ) =e x[f(pθ(x))], such as the log-likelihoodf(p) = logp. the gradient of this objective naturally acquires a prompt-dependent weight, ∇jf =...

5. Dual-Weighted Reinforcement Learning for Generative Preference Modeling

URL: View paper

Brief Assessment

Dual-Weighted RL[54] uses variance-aware weighting based on reward-model score variance to emphasize challenging prompts. However, this is applied in a different context (generative preference modeling with Bradley-Terry objectives) rather than the ORIGINAL paper's hybrid verifiable-reward framework combining rule-based verifiers with reward models for mathematical reasoning.

6. AWPO: Enhancing Tool-Use of Large Language Models through Explicit Integration of Reasoning Rewards

URL: View paper

Prior Art Analysis

AWPO[56] demonstrates prior work on variance-aware weighting mechanisms that adaptively adjust the contribution of different prompts during training based on reward variance. Both papers employ variance-based weighting schemes to emphasize challenging prompts with high variance while down-weighting easy prompts with uniform responses. The candidate paper's 'variance-aware gating' and 'difficulty-aware weighting' that 'adaptively modulate advantages from reasoning signals based on group-relative statistics' directly parallels the original paper's 'variance-aware weighting mechanism that adaptively adjusts the contribution of different prompts during training' where 'higher values suggest greater disagreement and thus a richer training signal.'

Evidence

Evidence 1 - **Rationale**: Both papers describe the same core mechanism: down-weighting easy/uniform prompts and emphasizing harder prompts with high variance. The candidate's 'difficulty-aware weighting' directly corresponds to the original's emphasis on 'harder prompts' where 'candidate responses vary widely.' - **Original**: easy prompts, where most responses are uniformly correct or incorrect, contribute little additional learning signal and are down-weighted. in contrast, harder prompts-where candidate responses vary widely and reward-model scores provide valuable discrimination-are emphasized. - **Candidate**: awpo incorporates variance-aware gating and difficulty-aware weighting to adaptively modulate advantages from reasoning signals based on group-relative statistics, alongside a tailored clipping mechanism for stable optimization.

Evidence 2 - **Rationale**: The original paper explicitly introduces a 'variance-aware weighting scheme' that uses standard deviation across candidate responses to reflect uncertainty and disagreement. The candidate's 'variance-aware gating' and use of 'group-relative statistics' represents the same technical approach to identifying and emphasizing informative prompts. - **Original**: to realign training effort with informativeness, we introduce avariance-aware weightingscheme. for each prompt, let σu denote the standard deviation of reward-model scores across candidate responses, with¯σas a running mean. this variance reflects uncertainty: higher values suggest greater disagreem... - **Candidate**: awpo incorporates variance-aware gating and difficulty-aware weighting to adaptively modulate advantages from reasoning signals based on group-relative statistics

7. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources

URL: View paper

Prior Art Analysis

Mmr1[51] demonstrates prior work on variance-aware mechanisms for emphasizing challenging prompts in reinforcement learning. The candidate paper introduces a 'variance-aware weighting mechanism' that adaptively adjusts the contribution of different prompts during training based on reward variance, emphasizing harder prompts with high variance while down-weighting easy prompts. This directly parallels the original paper's contribution of using variance-aware weighting to emphasize challenging prompts where dense signals matter most. Both papers employ variance metrics to identify informative prompts and reweight them during training, with the candidate providing a detailed formulation using variance promotion scores (VPS) that combine outcome variance and trajectory diversity.

Evidence

Evidence 1 - **Rationale**: Both papers use variance metrics computed from model responses to identify informative prompts. The original uses standard deviation of reward-model scores to reflect uncertainty, while the candidate computes outcome variance score from pass rates. Both approaches quantify variance to guide prompt selection. - **Original**: To realign training effort with informativeness, we introduce avariance-aware weightingscheme. For each prompt, let σu denote the standard deviation of reward-model scores across candidate responses, with¯σas a running mean. This variance reflects uncertainty: higher values suggest greater disagreem... - **Candidate**: let x be a prompt with ground-truth answer ¯yfrom the training set. in grpo framework, the model generates n responses {yi}n i=1 for each x. A task-specific verifier v(x, yi,¯y)∈ {0,1}evaluates each response, returning1ify i matches¯yand0otherwise. thepass rateforxis then defined as: p(x) = 1 n nx i...

Evidence 2 - **Rationale**: Both papers formalize variance-aware weighting through mathematical functions. The original defines a bounded monotone weighting function based on variance that up-weights difficult prompts, while the candidate combines outcome variance and trajectory diversity scores. Both approaches translate variance metrics into sample weights for training. - **Original**: we define a bounded monotone weighting function: wdifficulty(σu) =w min + (wmax -w min)· 1 1 + exp -k(σ u -¯σ) ,(4) wherew min andw max set the minimum and maximum weights, andkcontrols the slope of the transition. in practice, we treat these as tunable hyperparameters; unless otherwise stated, we... - **Candidate**: the overall variance promotion score (vps) is computed as a weighted combination of ovs and tds: vps(x) =α·ovs(x) +β·tds(x), where α, β >0balance their contributions. ovs increases the expected reward variance, while tds provides a lower bound by encouraging trajectory diversity. together, they are ...

8. Task Specific Sharpness Aware O-RAN Resource Management using Multi Agent Reinforcement Learning

URL: View paper

Brief Assessment

Task Specific Sharpness[55] applies variance-aware weighting in a multi-agent RL framework for O-RAN resource management, not for prompt-based language model training. The technical domains and applications differ fundamentally.

9. Optimizing Chain-of-Thought Reasoners via Gradient Variance Minimization in Rejection Sampling and RL

URL: View paper

Prior Art Analysis

Gradient Variance Minimization[50] demonstrates prior work on variance-aware weighting mechanisms that adaptively adjust contributions based on prompt difficulty. The candidate paper proposes 'gvm-raft, a prompt-specific dynamic sample allocation strategy designed to minimize stochastic gradient variance' and explicitly states that 'we propose a dynamic sample budget allocation strategy that adaptively assigns computational resources across prompts based on theoretical insights.' This directly addresses the same core concept as the original paper's variance-aware weighting mechanism, which 'adaptively adjusts the contribution of different prompts during training' and 'emphasizes harder prompts with high variance while down-weighting easy prompts.' Both papers identify variance in reward/gradient signals as a key indicator of prompt difficulty and use this to reweight training contributions.

Evidence

Evidence 1 - **Rationale**: Both papers propose adaptive weighting mechanisms that allocate resources based on prompt-specific characteristics. The candidate explicitly frames this as 'dynamic sample budget allocation' based on 'gradient variance minimization,' while the original frames it as 'variance-aware weighting' based on reward-model score variance. - **Original**: variance-aware weighting mechanism that adaptively adjusts the contribution of different prompts during training. easy prompts, where most responses are uniformly correct or incorrect, contribute little additional learning signal and are down-weighted. in contrast, harder prompts-where candidate res... - **Candidate**: we propose a dynamic sample budget allocation strategy that adaptively assigns computational resources across prompts based on theoretical insights. this leads to a more efficient monte carlo estimation of the elbo gradient. our resulting algorithm, a refined raft variant with dynamic inference budg...

Evidence 2 - **Rationale**: Both papers use variance as the key metric for determining prompt difficulty and allocating resources. The original uses 'standard deviation of reward-model scores' while the candidate uses 'stochastic gradient norms' and 'prompt acceptance rates,' but both aim to minimize variance in their respective training signals. - **Original**: for each prompt, let σu denote the standard deviation of reward-model scores across candidate responses, with¯σas a running mean. this variance reflects uncertainty: higher values suggest greater disagreement and thus a richer training signal. we define a bounded monotone weighting function: wdiffic... - **Candidate**: we propose gvm-raft, a prompt-specific dynamic sample allocation strategy designed to minimize stochastic gradient variance under a computational budget constraint. the method dynamically allocates computational resources by monitoring prompt acceptance rates and stochastic gradient norms, ensuring ...

Evidence 3 - **Rationale**: Both papers identify the same core insight: that uniform treatment of prompts is inefficient, and that emphasizing high-variance (difficult) prompts while de-emphasizing low-variance (easy) prompts improves training efficiency and performance. - **Original**: this design operationalizes our intuition: ambiguous, high-variance prompts are emphasized because they reveal more about model weaknesses and reward-model sensitivity, while trivial, lowvariance prompts are downweighted to avoid wasting capacity. - **Candidate**: this work identifies the main bottleneck in cot training as inefficient stochastic gradient estimation due to static sampling strategies. we propose gvm-raft, a prompt-specific dynamic sample allocation strategy designed to minimize stochastic gradient variance under a computational budget constrain...

10. Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense

URL: View paper

Brief Assessment

Hybrid Reinforcement Dense[22] is the same paper as the original submission. Both describe identical variance-aware weighting mechanisms with the same mathematical formulation (Equation 4) and hyperparameters. This is not a prior work but the same work.

Appendix: Text Similarity Detection

Textual similarity detection checked 21 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. Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense

Detected in: Core Task (sibling), Contribution: contribution_2, 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] Hybrid Reinforcement: when reward is sparse, better to be dense View paper
- [1] Exploring the limit of outcome reward for learning mathematical reasoning View paper
- [2] Teaching Large Language Models to Reason with Reinforcement Learning View paper
- [3] RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning View paper
- [4] Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search View paper
- [5] On designing effective rl reward at training time for llm reasoning View paper
- [6] Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning View paper
- [7] Conrft: A reinforced fine-tuning method for vla models via consistency policy View paper
- [8] Self-rewarding vision-language model via reasoning decomposition View paper
- [9] Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning View paper
- [10] MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models View paper
- [11] Repurposing Synthetic Data for Fine-grained Search Agent Supervision View paper
- [12] A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning View paper
- [13] Don't Just Fine-tune the Agent, Tune the Environment View paper

- [14] MedGR: Breaking the Data Barrier for Medical Reasoning via Generative Reward Learning View paper
- [15] Your Reward Function for RL is Your Best PRM for Search: Unifying RL and Search-Based TTS View paper
- [16] Process supervision-guided policy optimization for code generation View paper
- [17] Dense Rewards and Continual Reinforcement Learning for Task-oriented Dialogue Policies View paper
- [18] Enhancing Agentic RL with Progressive Reward Shaping and Value-based Sampling Policy Optimization View paper
- [19] Reinforcement Learning for Classical Planning: Viewing Heuristics as Dense Reward Generators View paper
- [20] Driving Beyond Privilege: Distilling Dense-Reward Knowledge into Sparse-Reward Policies View paper
- [21] DrS: Learning Reusable Dense Rewards for Multi-Stage Tasks View paper
- [22] Hybrid Reinforcement: When Reward Is Sparse, It's Better to Be Dense View paper
- [23] TaoSR-SHE: Stepwise Hybrid Examination Reinforcement Learning Framework for E-commerce Search Relevance View paper
- [24] A Practitioner's Guide to Multi-turn Agentic Reinforcement Learning View paper
- [25] PROPA: Toward Process-level Optimization in Visual Reasoning via Reinforcement Learning View paper
- [26] GRADE: Personalized Multi-Task Fusion via Group-relative Reinforcement Learning with Adaptive Dirichlet Exploration View paper
- [27] Recovery of partly sparse and dense signals View paper
- [28] Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Reward Design View paper
- [29] RobotKeyframing: Learning Locomotion with High-Level Objectives via Mixture of Dense and Sparse Rewards View paper
- [30] From Sparse to Dense: Toddler-inspired Reward Transition in Goal-Oriented Reinforcement Learning View paper
- [31] Generative and Pseudo-Relevant Feedback for Sparse, Dense and Learned Sparse Retrieval View paper
- [32] Deep Model-Based Reinforcement Learning for Predictive Control of Robotic Systems with Dense and Sparse Rewards View paper
- [33] A Study on Dense and Sparse (Visual) Rewards in Robot Policy Learning View paper
- [34] D2SR: Transferring Dense Reward Function to Sparse by Network Resetting View paper
- [35] Curiosity and exploration in mazes (dense vs. sparse rewards) View paper
- [36] Automated Ligand Design in Simulated Molecular Docking View paper
- [37] Sparse and Dense Approaches for the Full-rank Retrieval of Responses for Dialogues View paper
- [38] Self-Supervised Online Reward Shaping in Sparse-Reward Environments View paper
- [39] Multi-Slice Dense-Sparse Learning for Efficient Liver and Tumor Segmentation. View paper
- [40] Reinforcement Learning from Human Feedback to Fine-Tune Vision Models View paper
- [41] Process-Verified Reinforcement Learning for Theorem Proving via Lean View paper
- [42] SMARTER NOT HARDER: GENERATIVE PROCESS EVALUATION WITH INTRINSIC-SIGNAL DRIVING AND ABILITY-ADAPTIVE REWARD SHAPING View paper
- [43] REINFORCEMENT LEARNING View paper
- [44] Reward Generation via Large Vision-Language Model in Offline Reinforcement Learning View paper
- [45] Discriminative reward co-training View paper
- [46] Enhancing RLHF with Human Gaze Modeling View paper
- [47] Rubrics as rewards: Reinforcement learning beyond verifiable domains View paper
- [48] Learning to Explore in Diverse Reward Settings via Temporal-Difference-Error Maximization View paper
- [49] Reinforce-ada: An adaptive sampling framework for reinforce-style llm training View paper
- [50] Optimizing Chain-of-Thought Reasoners via Gradient Variance Minimization in Rejection Sampling and RL View paper
- [51] Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources View paper
- [52] Foundations for Efficient and Near-Optimal Reinforcement Learning View paper
- [53] Comparing Comparisons: Informative and Easy Human Feedback with Distinguishability Queries View paper
- [54] Dual-Weighted Reinforcement Learning for Generative Preference Modeling View paper
- [55] Task Specific Sharpness Aware O-RAN Resource Management using Multi Agent Reinforcement Learning View paper
- [56] AWPO: Enhancing Tool-Use of Large Language Models through Explicit Integration of Reasoning Rewards View paper
- [57] Reward-Guided Prompt Evolving in Reinforcement Learning for LLMs View paper
- [58] Neural Signatures Within and Between Chess Puzzle Solving and Standard Cognitive Tasks for Brain-Computer Interfaces: A Low-Cost Electroencephalography â¦ View paper