Novelty Assessment Report

Paper: OPPO: Accelerating PPO-based RLHF via Pipeline Overlap
PDF URL: https://openreview.net/pdf?id=31Mr6wLBeF
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Proximal Policy Optimization (PPO)-based reinforcement learning from human feedback (RLHF) is a widely adopted paradigm for aligning large language models (LLMs) with human preferences. However, its training pipeline suffers from substantial inefficiencies due to sequential multi-model dependencies (e.g., reward model depends on actor outputs) and long-tail response lengths, where a few long responses straggle the stage completion. We present OPPO, a novel, lightweight, and model-agnostic PPO-based RLHF framework that improves training efficiency by overlapping pipeline execution. OPPO introduces two novel techniques: (1) Intra-step overlap, which streams upstream model outputs (e.g., actor model) in right-sized chunks, enabling the downstream model (e.g., reward) to begin prefill while the upstream continues decoding; and (2) Inter-step overlap, which adaptively overcommits a few prompts and defers long generations to future steps, mitigating tail latency without discarding partial work. OPPO integrates easily with existing PPO implementations with a lightweight wrapper. Extensive evaluations show that OPPO accelerates PPO-based RLHF training by $1.8\times$--$2.8\times$ and improves GPU utilization by $1.4\times$--$2.1\times$ without compromising training convergence.

Core Task Landscape

This paper addresses: **Accelerating PPO-based RLHF Training**

A total of **38 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **System-Level Acceleration and Pipeline Optimization**
- **Algorithmic Improvements to PPO and RLHF**
- **Reward Model Improvements**
- **Data and Feedback Optimization**
- **Domain-Specific RLHF Applications**
- **Empirical Analysis and Benchmarking**

Complete Taxonomy Tree

- Accelerating PPO-based RLHF Training Survey Taxonomy
- System-Level Acceleration and Pipeline Optimization
  - Pipeline Overlap and Streaming Execution ★ (2 papers)
  - [0] OPPO: Accelerating PPO-based RLHF via Pipeline Overlap (Anon et al., 2026) View paper
  - [12] Faster, more efficient RLHF through off-policy asynchronous learning (M Noukhovitch, 2025) View paper
  - Memory Efficiency and Parameter-Efficient Training (3 papers)
  - [7] PERL: Parameter Efficient Reinforcement Learning from Human Feedback (Sidahmed, 2024) View paper
  - [18] Efficient RLHF: Reducing the Memory Usage of PPO (Santacroce, 2023) View paper
  - [36] Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of RLHF (Sun, 2023) View paper
- Algorithmic Improvements to PPO and RLHF
  - Alternative RL Algorithms and PPO Replacements (3 papers)
  - [2] Superhf: Supervised iterative learning from human feedback (Gabriel Mukobi, 2023) View paper
  - [4] Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models (Li, 2023) View paper
  - [32] SLiC-HF: Sequence Likelihood Calibration with Human Feedback (Zhao, 2023) View paper
  - PPO Enhancements and Stabilization (5 papers)
  - [10] Symmetric Reinforcement Learning Loss for Robust Learning on Diverse Tasks and Model Scales (Byun, 2024) View paper
  - [14] Pairwise Proximal Policy Optimization: Harnessing Relative Feedback for LLM Alignment (Wu, 2023) View paper
  - [26] Group Policy Gradient (Chen Jun-hua, 2025) View paper
  - [27] DPO Meets PPO: Reinforced Token Optimization for RLHF (Zhong Han, 2024) View paper
  - [29] Advancing AI Agents Through Reinforcement Learning: Stabilizing Actor-Critic On-policy Algorithms, Enabling Smooth Policy Transitions, and Enhancing Chain-of â¦ (Byun, 2024) View paper
  - Unified Frameworks and Theoretical Foundations (2 papers)
  - [16] One Framework to Rule Them All: Unifying RL-Based and RL-Free Methods in RLHF (Cai Xin, 2025) View paper
  - [17] UNA: unifying alignments of RLHF/PPO, DPO and KTO by a generalized implicit reward function (Wang Zhichao, 2024) View paper
- Reward Model Improvements
  - Reward Model Robustness and Noise Mitigation (2 papers)

    - [3] Improving Reinforcement Learning from Human Feedback Using Contrastive Rewards (Shen Wei, 2024) View paper
    - [5] Improving Reinforcement Learning from Human Feedback with Efficient Reward Model Ensemble (Zhang Shun, 2024) View paper
    - Value Function and Advantage Estimation (1 papers)
    - [15] VRPO: Rethinking Value Modeling for Robust RL Training under Noisy Supervision (Zhu Ding-wei, 2025) View paper
    - Hybrid Reward Systems and Multi-Dimensional Evaluation (2 papers)
    - [11] Exploring Data Scaling Trends and Effects in Reinforcement Learning from Human Feedback (Shen Wei, 2025) View paper
    - [20] PPO-based Reinforcement Learning with Human Feedback with Hybrid Oversight and Predictive Reward Evaluation for AGI (Atul Sharma, 2025) View paper
- Data and Feedback Optimization
    - Simulation and Synthetic Feedback (1 papers)
    - [1] Alpacafarm: A simulation framework for methods that learn from human feedback (Dubois, 2023) View paper
    - Human Feedback Mechanisms and Interaction Design (4 papers)
    - [6] Human-inspired framework to accelerate reinforcement learning: A. Beikmohammadi, S. MagnÃºsson (A Beikmohammadi, 2025) View paper
    - [25] Improving the Precision of Hidden DangerÂ Recognition in Power Dispatch Duty Logs through RLHF Multi-round Human Feedback Mechanism (Siwu Yu, 2025) View paper
    - [30] Towards Faithful and Controllable Personalization via Critique-Post-Edit Reinforcement Learning (Zhu, 2025) View paper
    - [31] TEMPO: Timestep Explanations for Modeling Preferences in Online Preference-Based RL (Jakob Karlaus, 2025) View paper
- Domain-Specific RLHF Applications
    - Conversational and Multi-Turn Optimization (2 papers)
    - [28] Proximal Policy Optimization Actual Combat: Manipulating Output Tokenizer Length (Fan Miao, 2023) View paper
    - [34] Aligning LLMs Toward Multi-Turn Conversational Outcomes Using Iterative PPO (Daniel R. Jiang, 2025) View paper
    - Specialized Task Domains (4 papers)
    - [8] Enhancing LLMs for Physics Problem-Solving using Reinforcement Learning with Human-AI Feedback (Anand, 2024) View paper
    - [21] Reinforcement learning for question answering in programming domain using public community scoring as a human feedback (Gorbatovski, 2024) View paper
    - [23] Reinforcement Learning for Safe LLM Code Generation (Huang, 2025) View paper
    - [24] Optimizing the Human-machine Collaborative Mechanism in English Oral Teaching: A Feedback Method Based on Reinforcement Learning (Chang, 2025) View paper
    - Robotics and Physical Interaction (5 papers)
    - [9] Research on reinforcement learning based on PPO algorithm for human-machine intervention in autonomous driving. (Gaosong Shi, 2024) View paper
    - [19] Accelerating Virtual Fixture Estimation for Robot Manipulation using Reinforcement Learning and Human Demonstrations (Diego Fernandez Prado, 2024) View paper
    - [22] Learning Real-World Acrobatic Flight from Human Preferences (Geles, 2025) View paper
    - [35] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice (Plisnier, 2018) View paper
    - [38] Bayesian Proximal Policy Optimization with Adaptive Learning and Episodic Memory for Social Robot Navigation (Faria, n.d.) View paper
    - Non-LLM Applications (2 papers)
    - [13] Human-Guided Data Augmentation via Diffusion Model for Surface Defect Recognition Under Limited Data (Tiyu Fang, 2025) View paper
    - [33] Hur ska jag lÃ¥ta? Att hitta en rÃ¶st fÃ¶r Blossom-roboten (Qiaolan, 2025) View paper
- Empirical Analysis and Benchmarking (1 papers)
    - [37] Secrets of RLHF in Large Language Models Part I: PPO (Zheng Rui, 2023) View paper

Narrative

Core task: Accelerating PPO-based reinforcement learning from human feedback training. The field of RLHF acceleration has evolved into a multi-faceted landscape, with the taxonomy revealing six major branches that address complementary bottlenecks. System-Level Acceleration and Pipeline Optimization focuses on engineering solutions such as pipeline overlap, streaming execution, and memory-efficient implementations (e.g., Efficient RLHF Memory[18]) to reduce wall-clock time. Algorithmic Improvements to PPO and RLHF explore modifications to the core learning dynamics, including variance reduction techniques (VRPO[15]), alternative policy gradient formulations (Directed Policy Gradient[35]), and hybrid methods that blend PPO with other paradigms (DPO Meets PPO[27]). Reward Model Improvements tackle the quality and robustness of learned preferences through ensemble methods (Reward Model Ensemble[5]) and contrastive formulations (Contrastive Rewards[3]). Data and Feedback Optimization investigates how to scale and augment training signals (Data Scaling RLHF[11], Diffusion Data Augmentation[13]), while Domain-Specific RLHF Applications demonstrate tailored deployments in areas like autonomous driving (Autonomous Driving Intervention[9]) and code generation (Safe Code Generation[23]). Finally, Empirical Analysis and Benchmarking provides controlled testbeds (Alpacafarm[1]) and systematic studies (Secrets of RLHF[37]) to guide practitioners.

Within the system-level branch, a particularly active line of work addresses pipeline overlap and asynchronous execution to hide latency across the multi-model RLHF workflow. OPPO[0] exemplifies this direction by introducing overlapped scheduling that interleaves actor rollouts, critic evaluations, and reward model queries, achieving substantial speedups without sacrificing sample efficiency. This approach contrasts with Off Policy Async[12], which relaxes on-policy constraints to enable fully asynchronous updates, trading some alignment stability for throughput gains. Meanwhile, works like Superhf[2] and TEMPO[31] explore complementary angles—Superhf[2] optimizes distributed communication patterns, while TEMPO[31] focuses on temporal credit assignment within the pipeline. OPPO[0] sits squarely in the pipeline overlap cluster, sharing the goal of minimizing idle GPU time with Off Policy Async[12] but maintaining tighter synchronization to preserve PPO's on-policy guarantees, thus offering a middle ground between pure synchronous training and fully decoupled asynchronous schemes.

Related Works in Same Category

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

1. Faster, more efficient RLHF through off-policy asynchronous learning

Authors: M Noukhovitch, S Huang, S Xhonneux | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ RLHF. This enables asynchronous generation of new samples while simultaneously training on old samples, leading to faster training â¦ popular methods, namely PPO, RLOO, and Online â¦

Relationship Analysis

Both papers belong to the Pipeline Overlap and Streaming Execution category, addressing the inefficiency of sequential multi-model dependencies in PPO-based RLHF training. While OPPO focuses on intra-step overlap by streaming actor outputs in chunks to enable concurrent prefilling in downstream models and inter-step overlap through adaptive prompt overcommitment, the candidate paper (Faster, more efficient RLHF) takes a different approach by separating generation and training onto different GPUs to enable asynchronous execution, accepting one-step staleness (off-policy learning) as a trade-off. The key distinction is that OPPO maintains synchronous, on-policy training while overlapping execution stages within the same step, whereas the candidate paper explicitly embraces asynchronous, off-policy training across steps to leverage specialized inference libraries like vLLM.

Contributions Analysis

Overall novelty summary. The paper proposes OPPO, a framework that accelerates PPO-based RLHF training through intra-step and inter-step overlap techniques, achieving 1.8×–2.8× speedups. It resides in the 'Pipeline Overlap and Streaming Execution' leaf, which contains only one sibling paper (Off Policy Async). This leaf sits within the broader 'System-Level Acceleration and Pipeline Optimization' branch, indicating a relatively sparse research direction focused specifically on overlapping multi-model pipeline stages. The taxonomy reveals that system-level acceleration is one of six major branches, suggesting that pipeline overlap represents a targeted but underexplored approach compared to algorithmic or reward model improvements.

The taxonomy shows that OPPO's leaf is adjacent to 'Memory Efficiency and Parameter-Efficient Training' (containing three papers on LoRA and model integration) within the same parent branch. Neighboring branches include 'Algorithmic Improvements to PPO and RLHF' (with 10 papers across three leaves) and 'Reward Model Improvements' (with 5 papers across three leaves). The scope notes clarify that OPPO's pipeline overlap focus excludes algorithmic modifications to PPO itself and memory reduction techniques without streaming execution. This positioning suggests the work addresses a distinct bottleneck—sequential multi-model dependencies—rather than competing directly with algorithmic or reward modeling innovations.

Among 17 candidates examined across three contributions, none were found to clearly refute OPPO's novelty. The intra-step overlap technique examined 10 candidates with no refutable matches, while inter-step overlap examined 2 candidates and the overall OPPO framework examined 5 candidates, both with zero refutations. The single sibling paper (Off Policy Async) explores asynchronous updates by relaxing on-policy constraints, whereas OPPO maintains tighter synchronization through streaming and adaptive overcommitment. This limited search scope suggests that within the examined top-17 semantic matches, OPPO's specific combination of intra-step streaming and inter-step tail-latency mitigation appears distinct from prior pipeline optimization strategies.

Based on the top-17 candidates examined, OPPO appears to occupy a relatively novel position within the sparse pipeline overlap research direction. The analysis does not cover exhaustive literature search or broader system optimization techniques outside the semantic neighborhood. The taxonomy structure indicates that while system-level acceleration is an active area overall, the specific focus on overlapping multi-model RLHF pipelines through streaming and adaptive scheduling remains underexplored compared to algorithmic or memory-centric approaches.

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

Contribution 1: Intra-step overlap technique for PPO-based RLHF

Description: A technique that streams actor model outputs in adaptive chunks to downstream models (e.g., reward model), enabling the reward model to begin prefilling while the actor continues decoding. This overlaps generation and scoring stages within a single PPO step, hiding prefilling latency and reducing execution bubbles without altering the generated responses or PPO update semantics.

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. Moshi: a speech-text foundation model for real-time dialogue

URL: View paper

Brief Assessment

Moshi[39] focuses on real-time speech-text dialogue modeling with multi-stream audio generation, not on PPO-based RLHF training optimization or streaming model outputs for reward scoring.

---

2. Dynamic Chunk Convolution for Unified Streaming and Non-Streaming Conformer ASR

URL: View paper

Brief Assessment

Dynamic Chunk Convolution[45] addresses streaming ASR by chunking audio inputs for speech recognition models, not RL training pipelines. The candidate focuses on acoustic model inference with convolution operations, whereas the original contribution concerns overlapping generation and scoring stages in PPO-based RLHF training.

---

3. RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation

URL: View paper

Brief Assessment

RAGCache[41] focuses on optimizing RAG systems by overlapping retrieval with LLM generation, not on PPO-based RLHF training pipelines or streaming actor model outputs to reward models.

---

4. Massively Parallel Open Source Encoding for Adaptive Streaming

URL: View paper

Brief Assessment

Parallel Adaptive Streaming[44] focuses on distributed video encoding by splitting videos into non-overlapping chunks processed in parallel across machines. This is fundamentally different from streaming model outputs in chunks during PPO-based RLHF training to overlap generation and scoring stages.

---

5. Parallel and streaming generation of ghost data for structured grids.

URL: View paper

Brief Assessment

Ghost Data Streaming[46] addresses parallel generation of ghost data for structured grids in scientific computing, which is unrelated to streaming model outputs in machine learning pipelines or PPO-based RLHF training.

---

6. Expander Chunked Codes

URL: View paper

Brief Assessment

Expander Chunked Codes[47] addresses network coding for data dissemination with chunked transmission schemes, not reinforcement learning or model training pipelines. The streaming mechanism in Expander Chunked Codes[47] relates to network packet transmission, not overlapping generation and scoring stages in PPO-based RLHF.

---

7. Omniflatten: An end-to-end gpt model for seamless voice conversation

URL: View paper

Brief Assessment

Omniflatten[40] focuses on full-duplex spoken dialogue systems using speech tokenization and chunking for real-time voice conversation. It does not address PPO-based RLHF training pipelines, reward model scoring, or the overlap of generation and scoring stages in reinforcement learning contexts.

---

8. Segment streaming for the three-phase execution model: Design and implementation

URL: View paper

Brief Assessment

Segment Streaming[43] addresses real-time task scheduling with memory-computation overlap in embedded systems using DMA and scratchpad memory, not LLM training pipelines or PPO-based RLHF workflows.

---

9. Network Codes with Overlapping Chunks over Line Networks: A Case for Linear-Time Codes

URL: View paper

Brief Assessment

Overlapping Chunks[48] addresses network coding over packet networks with chunked data transmission, not reinforcement learning or PPO training pipelines. The candidate focuses on overlapping data chunks for error correction in network transmission, while the original contribution concerns overlapping execution stages (generation and scoring) in RLHF training.

---

10. gLLM: Global Balanced Pipeline Parallelism Systems for Distributed LLMs Serving with Token Throttling

URL: View paper

Brief Assessment

gLLM[42] focuses on pipeline parallelism for LLM serving with token throttling to balance prefill and decode stages. The original paper addresses PPO-based RLHF training by streaming actor outputs to reward models, which is a different domain and use case.

Contribution 2: Inter-step overlap technique for PPO-based RLHF

Description: A technique that adaptively overcommits a small number of prompts per batch and defers long-response generations to future iterations. This mitigates tail latency caused by heterogeneous response lengths while preserving partial generation work and maintaining batch size, with dynamic adjustment of the overcommitment level to balance throughput gains against statistical deviations.

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. Integrating Edge Computing and Machine Learning for Low-Latency Decision Making in Next-Generation Intelligent Transportation Infrastructures

URL: View paper

Brief Assessment

Edge Computing Transportation[49] focuses on edge computing architectures for intelligent transportation systems with machine learning for traffic management and vehicle coordination. It does not address PPO-based RLHF training pipelines, prompt overcommitment strategies, or techniques for mitigating tail latency in language model training.

---

2. A Unified Data and Machine Learning Framework for Cross-Device, Cross-Channel Identity Resolution for Consistent Personalization in B2C Digital Sales

URL: View paper

Brief Assessment

Cross Device Identity[50] addresses cross-device identity resolution for B2C personalization, not reinforcement learning or PPO training optimization. The candidate focuses on linking user identifiers across devices and channels for consistent customer experiences, which is entirely unrelated to the original paper's contribution of overlapping prompts in RLHF training pipelines.

Contribution 3: OPPO framework for accelerating PPO-based RLHF

Description: A lightweight and model-agnostic framework that accelerates PPO-based RLHF training by overlapping pipeline execution through intra-step and inter-step techniques. OPPO integrates easily with existing PPO implementations via a lightweight wrapper and achieves substantial speedups and GPU utilization improvements without compromising training convergence.

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. An Adaptive Placement and Parallelism Framework for Accelerating RLHF Training

URL: View paper

Brief Assessment

Adaptive Placement Framework[53] focuses on model placement strategies (co-located, interleaving, disaggregated) across devices to reduce memory redundancy and communication costs, not on overlapping pipeline execution within stages as OPPO does.

---

2. OpenRLHF: A Ray-based Easy-to-use, Scalable and High-performance RLHF Framework

URL: View paper

Brief Assessment

OpenRLHF[54] focuses on system architecture using Ray-based distributed computing, vLLM inference optimization, and DeepSpeed for parallelism. It does not address pipeline overlap techniques (intra-step or inter-step) that are central to OPPO's contribution.

---

3. Rlinf-vla: A unified and efficient framework for vla+ rl training

URL: View paper

Brief Assessment

Rlinf VLA[52] focuses on vision-language-action (VLA) model training for robotics with RL, not on accelerating PPO-based RLHF for language models. The candidate addresses GPU allocation strategies for integrating rendering, training, and inference in robotic simulators, which is a fundamentally different problem domain than overlapping pipeline execution in multi-model RLHF training.

---

4. BQSched: A Non-intrusive Scheduler for Batch Concurrent Queries via Reinforcement Learning

URL: View paper

Brief Assessment

BQSched[51] focuses on batch query scheduling in database systems using RL, not on accelerating PPO-based RLHF training for language models. The candidate addresses a completely different domain (database query optimization) with different technical challenges.

---

5. HEPPO: Hardware-Efficient Proximal Policy Optimization a Universal Pipelined Architecture for Generalized Advantage Estimation

URL: View paper

Brief Assessment

HEPPO[55] focuses on FPGA-based hardware acceleration for the GAE computation stage in PPO, not on overlapping pipeline execution for multi-model RLHF training. The candidate addresses hardware-level optimization of a single computational component, while the original addresses system-level pipeline orchestration across four models (actor, critic, reference, reward) in RLHF workflows.

Appendix: Text Similarity Detection

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

References

- [0] OPPO: Accelerating PPO-based RLHF via Pipeline Overlap View paper
- [1] Alpacafarm: A simulation framework for methods that learn from human feedback View paper
- [2] Superhf: Supervised iterative learning from human feedback View paper
- [3] Improving Reinforcement Learning from Human Feedback Using Contrastive Rewards View paper
- [4] Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models View paper
- [5] Improving Reinforcement Learning from Human Feedback with Efficient Reward Model Ensemble View paper
- [6] Human-inspired framework to accelerate reinforcement learning: A. Beikmohammadi, S. MagnÃºsson View paper
- [7] PERL: Parameter Efficient Reinforcement Learning from Human Feedback View paper
- [8] Enhancing LLMs for Physics Problem-Solving using Reinforcement Learning with Human-AI Feedback View paper
- [9] Research on reinforcement learning based on PPO algorithm for human-machine intervention in autonomous driving. View paper
- [10] Symmetric Reinforcement Learning Loss for Robust Learning on Diverse Tasks and Model Scales View paper
- [11] Exploring Data Scaling Trends and Effects in Reinforcement Learning from Human Feedback View paper
- [12] Faster, more efficient RLHF through off-policy asynchronous learning View paper
- [13] Human-Guided Data Augmentation via Diffusion Model for Surface Defect Recognition Under Limited Data View paper
- [14] Pairwise Proximal Policy Optimization: Harnessing Relative Feedback for LLM Alignment View paper
- [15] VRPO: Rethinking Value Modeling for Robust RL Training under Noisy Supervision View paper
- [16] One Framework to Rule Them All: Unifying RL-Based and RL-Free Methods in RLHF View paper
- [17] UNA: unifying alignments of RLHF/PPO, DPO and KTO by a generalized implicit reward function View paper
- [18] Efficient RLHF: Reducing the Memory Usage of PPO View paper
- [19] Accelerating Virtual Fixture Estimation for Robot Manipulation using Reinforcement Learning and Human Demonstrations View paper
- [20] PPO-based Reinforcement Learning with Human Feedback with Hybrid Oversight and Predictive Reward Evaluation for AGI View paper
- [21] Reinforcement learning for question answering in programming domain using public community scoring as a human feedback View paper
- [22] Learning Real-World Acrobatic Flight from Human Preferences View paper
- [23] Reinforcement Learning for Safe LLM Code Generation View paper
- [24] Optimizing the Human-machine Collaborative Mechanism in English Oral Teaching: A Feedback Method Based on Reinforcement Learning View paper
- [25] Improving the Precision of Hidden DangerÂ Recognition in Power Dispatch Duty Logs through RLHF Multi-round Human Feedback Mechanism View paper
- [26] Group Policy Gradient View paper
- [27] DPO Meets PPO: Reinforced Token Optimization for RLHF View paper
- [28] Proximal Policy Optimization Actual Combat: Manipulating Output Tokenizer Length View paper
- [29] Advancing AI Agents Through Reinforcement Learning: Stabilizing Actor-Critic On-policy Algorithms, Enabling Smooth Policy Transitions, and Enhancing Chain-of â¦ View paper
- [30] Towards Faithful and Controllable Personalization via Critique-Post-Edit Reinforcement Learning View paper
- [31] TEMPO: Timestep Explanations for Modeling Preferences in Online Preference-Based RL View paper
- [32] SLiC-HF: Sequence Likelihood Calibration with Human Feedback View paper
- [33] Hur ska jag lÃ¥ta? Att hitta en rÃ¶st fÃ¶r Blossom-roboten View paper
- [34] Aligning LLMs Toward Multi-Turn Conversational Outcomes Using Iterative PPO View paper
- [35] Directed Policy Gradient for Safe Reinforcement Learning with Human Advice View paper
- [36] Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of RLHF View paper
- [37] Secrets of RLHF in Large Language Models Part I: PPO View paper
- [38] Bayesian Proximal Policy Optimization with Adaptive Learning and Episodic Memory for Social Robot Navigation View paper
- [39] Moshi: a speech-text foundation model for real-time dialogue View paper
- [40] Omniflatten: An end-to-end gpt model for seamless voice conversation View paper
- [41] RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation View paper
- [42] gLLM: Global Balanced Pipeline Parallelism Systems for Distributed LLMs Serving with Token Throttling View paper
- [43] Segment streaming for the three-phase execution model: Design and implementation View paper

- [44] Massively Parallel Open Source Encoding for Adaptive Streaming View paper
- [45] Dynamic Chunk Convolution for Unified Streaming and Non-Streaming Conformer ASR View paper
- [46] Parallel and streaming generation of ghost data for structured grids. View paper
- [47] Expander Chunked Codes View paper
- [48] Network Codes with Overlapping Chunks over Line Networks: A Case for Linear-Time Codes View paper
- [49] Integrating Edge Computing and Machine Learning for Low-Latency Decision Making in Next-Generation Intelligent Transportation Infrastructures View paper
- [50] A Unified Data and Machine Learning Framework for Cross-Device, Cross-Channel Identity Resolution for Consistent Personalization in B2C Digital Sales View paper
- [51] BQSched: A Non-intrusive Scheduler for Batch Concurrent Queries via Reinforcement Learning View paper
- [52] Rlinf-vla: A unified and efficient framework for vla+ rl training View paper
- [53] An Adaptive Placement and Parallelism Framework for Accelerating RLHF Training View paper
- [54] OpenRLHF: A Ray-based Easy-to-use, Scalable and High-performance RLHF Framework View paper
- [55] HEPPO: Hardware-Efficient Proximal Policy Optimization a Universal Pipelined Architecture for Generalized Advantage Estimation View paper