Novelty Assessment Report

Paper: Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game
PDF URL: https://openreview.net/pdf?id=vc9Tj11LNE
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We introduce Stackelberg Learning from Human Feedback (SLHF), a new framework for preference optimization. SLHF frames the alignment problem as a sequential-move game between two policies: a Leader, which commits to an action, and a Follower, which responds conditionally on the Leader's action. This approach decomposes preference optimization into a refinement problem for the Follower and an optimization problem against an adversary for the Leader. Unlike Reinforcement Learning from Human Feedback (RLHF), which assigns scalar rewards to actions, or Nash Learning from Human Feedback (NLHF), which seeks a simultaneous-move equilibrium, SLHF leverages the asymmetry of sequential play to capture richer preference structures. The sequential design of SLHF naturally enables inference-time refinement, as the Follower learns to improve the Leader's actions, and these refinements can be leveraged through iterative sampling. We compare the solution concepts of SLHF, RLHF, and NLHF, and lay out key advantages in consistency, data sensitivity, and robustness to intransitive preferences. Experiments on large language models demonstrate that SLHF achieves strong alignment across diverse preference datasets, scales from 0.5B to 8B parameters, and yields inference-time refinements that transfer across model families without further fine-tuning.

Core Task Landscape

This paper addresses: **Preference Optimization from Pairwise Human Feedback using Sequential Game Theory**

A total of **26 papers** were analyzed and organized into a taxonomy with **13 categories**.

Taxonomy Overview

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

- **Sequential Game-Theoretic Frameworks for Preference Learning**
- **Nash Equilibrium Approaches to Preference Optimization**
- **Bayesian and Bandit Approaches to Pairwise Preference Learning**
- **Experimental Design and Statistical Methods for Preference Studies**
- **Applications and Extensions of Game-Theoretic Preference Models**

Complete Taxonomy Tree

- Preference Optimization from Pairwise Human Feedback using Sequential Game Theory Survey Taxonomy
- Sequential Game-Theoretic Frameworks for Preference Learning
  - Stackelberg Equilibrium-Based Preference Optimization ★ (3 papers)
  - [0] Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game (Anon et al., 2026) View paper
  - [4] Sta-rlhf: Stackelberg aligned reinforcement learning with human feedback (J Makar-Limanov, 2024) View paper
  - [16] Bandits with Preference Feedback: A Stackelberg Game Perspective (Parnian Kassraie, 2024) View paper
  - Multi-Step Sequential Decision Processes with Preference Feedback (2 papers)
  - [3] Sequential Preference Ranking for Efficient Reinforcement Learning from Human Feedback (Minyoung Hwang, 2023) View paper
  - [26] Multi-Step Preference Optimization via Two-Player Markov Games (Y Wu, n.d.) View paper
  - Sequential Voting and Agenda-Based Preference Aggregation (2 papers)
  - [12] 5. Sequential pairwise voting (Osborne, 2025) View paper
  - [17] Revealed preferences of individual players in sequential games (Hiroki Nishimura, 2021) View paper
- Nash Equilibrium Approaches to Preference Optimization
  - Nash Learning from Human Feedback with Non-Transitive Preferences (2 papers)
  - [1] Nash learning from human feedback (Munos, 2024) View paper
  - [6] Extragradient Preference Optimization (EGPO): Beyond Last-Iterate Convergence for Nash Learning from Human Feedback (Zhou, 2025) View paper
  - Minimax and Self-Play Preference Optimization (2 papers)
  - [2] A Minimaximalist Approach to Reinforcement Learning from Human Feedback (Swamy, 2024) View paper
  - [5] Online Iterative Reinforcement Learning from Human Feedback with General Preference Model (Hanze Dong, 2024) View paper
  - Theoretical Foundations of Game-Theoretic Alignment (2 papers)
  - [10] Fundamental Limits of Game-Theoretic LLM Alignment: Smith Consistency and Preference Matching (Zhekun Shi, 2025) View paper
  - [11] Preference learning along multiple criteria: A game-theoretic perspective (Bhatia, 2020) View paper
- Bayesian and Bandit Approaches to Pairwise Preference Learning
  - Bayesian Optimization with Pairwise Comparisons (2 papers)
  - [14] A parametric approach to Bayesian optimization with pairwise comparisons (Marco Cox, 2017) View paper
  - [15] Learning to maximize the social welfare from preference feedback (A Bergerault, 2024) View paper

- Convergence and Efficiency in Preference-Based Learning (2 papers)
- [13] Cooperative intent: an exploration of computational learning in a discrete preference space (Stanton, 2024) View paper
- [21] Finite-time convergence to an -efficient Nash equilibrium in potential games (A Maddux, 2024) View paper

• Experimental Design and Statistical Methods for Preference Studies
- Optimal Designs for Pairwise Treatment Comparisons (2 papers)
- [18] Group sequential two-stage preference designs (Liu, 2023) View paper
- [22] Optimal designs for testing pairwise differences: A graphâ□□based game theoretic approach (Arpan Singh, 2024) View paper
- Dynamic and Adaptive Preference Elicitation Procedures (1 papers)
- [7] Dynamically optimized sequential experimentation (DOSE) for estimating economic preference parameters (Jonathan Chapman, 2024) View paper

• Applications and Extensions of Game-Theoretic Preference Models
- Strategic Manipulation and Robustness in Group Preference Aggregation (1 papers)
- [24] Strategic Manipulation in Group Decisions with Pairwise Comparisons: A Game Theoretical Perspective (Yasuo Sasaki, 2022) View paper
- Game-Theoretic Clustering and Classification with Pairwise Data (2 papers)
- [19] A Game Theoretic Based K-Nearest Neighbor Approach for Binary Classification (Rodica Ioana Lung, 2023) View paper
- [23] A game-theoretic approach to pairwise clustering and matching (M. Pelillo, 2013) View paper
- Behavioral and Applied Preference Modeling (4 papers)
- [8] Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach (Wei Lu, 2025) View paper
- [9] Modelling the dynamic vaccination game with evolutionary feedback: exploring pairwise interactions and vaccine strategies (Khondoker Nazmoon Nabi, 2024) View paper
- [20] Intransitivity cycles and their transformations: How dynamically adapting systems function (Alexander Poddiakov, 2013) View paper
- [25] Deterrence and Risk Preferences in Sequential Attacker-Defender Games with Continuous Efforts. (Vineet M. Payyappalli, 2018) View paper

Narrative

Core task: Preference optimization from pairwise human feedback using sequential game theory. The field structures itself around several complementary perspectives on how to extract and optimize preferences from comparative judgments. Sequential game-theoretic frameworks treat preference learning as a multi-stage interaction where one agent (e.g., a learner or policy) anticipates the responses of another (e.g., a human evaluator or reward model), leading to Stackelberg equilibrium formulations such as Stackelberg Learning[0] and Stackelberg Aligned RLHF[4]. Nash equilibrium approaches like Nash Learning[1] and Minimaximalist RLHF[2] instead model simultaneous best-response dynamics, often emphasizing robustness and worst-case guarantees. Bayesian and bandit methods (e.g., Bayesian Pairwise Comparisons[14], Stackelberg Bandits[16]) focus on uncertainty quantification and exploration-exploitation trade-offs when feedback is noisy or scarce. Experimental design branches address how to select informative pairs efficiently (Pairwise Testing Designs[22], Group Sequential Designs[18]), while application-oriented work extends these ideas to domains such as social welfare aggregation (Social Welfare Learning[15]) and strategic manipulation (Strategic Pairwise Manipulation[24]).

A particularly active line of inquiry contrasts leader-follower (Stackelberg) versus simultaneous-move (Nash) solution concepts, exploring how anticipatory reasoning affects convergence, sample efficiency, and alignment quality. Stackelberg Learning[0] sits squarely within this debate, emphasizing the benefits of sequential commitment when the learner can credibly shape the evaluator's behavior. This contrasts with Nash Learning[1], which assumes neither party commits first, and with Online Iterative RLHF[5], which iteratively refines policies without explicit game-theoretic equilibrium guarantees. Nearby works like Stackelberg Aligned RLHF[4] share the sequential equilibrium perspective but may differ in algorithmic implementation or application domain, while Extragradient Preference Optimization[6] explores gradient-based dynamics that can approximate equilibrium solutions. Open questions remain around computational tractability, the realism of equilibrium assumptions when human feedback is inconsistent, and how to integrate sequential game reasoning with modern large-scale preference datasets.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Sta-rlhf: Stackelberg aligned reinforcement learning with human feedback

Authors: J Makar-Limanov, A Prakash, D Goktas | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â□¦ model, which aligns its behavior with human preferences by optimizing the reward model, a â□¦ game can be better understood as a sequential game, whose solution is a Stackelberg, â□¦

Relationship Analysis

Both papers belong to the Stackelberg Equilibrium-Based Preference Optimization category, using leader-follower sequential game structures for preference learning. They share the core approach of formulating preference optimization as a Stackelberg game with a Leader policy and a Follower policy, both addressing the limitations of RLHF's scalar reward assumptions. The key difference is that the original paper (SLHF) frames the Leader as the initial policy generator and the Follower as a refinement policy that conditions on the Leader's realized action, enabling inference-time refinement, whereas the candidate paper (STA-RLHF) treats the language model as the Leader and the reward model as the Follower, focusing on joint optimization of policy and reward model through nested gradient descent.

2. Bandits with Preference Feedback: A Stackelberg Game Perspective

Authors: Parnian Kassraie, Andreas Krause, Barna Pasztor | **Year/Venue**: 2024 | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the Stackelberg Equilibrium-Based Preference Optimization category, applying leader-follower sequential game structures to preference learning problems. The original paper (SLHF) focuses on aligning large language models with human feedback by framing preference optimization as a sequential game between two policies (Leader and Follower), enabling inference-time refinement through iterative sampling. The candidate paper addresses bandits with preference feedback from a Stackelberg game perspective, which likely focuses on sequential decision-making in bandit settings rather than LLM alignment, representing a different application domain within the same game-theoretic framework.

Contributions Analysis

Overall novelty summary. The paper introduces Stackelberg Learning from Human Feedback (SLHF), framing preference optimization as a leader-follower sequential game where one policy commits first and another responds conditionally. Within the taxonomy, it resides in the 'Stackelberg Equilibrium-Based Preference Optimization' leaf, which contains only three papers total. This is a notably sparse research direction compared to the broader field of 26 papers across multiple equilibrium concepts, suggesting the sequential game-theoretic perspective on preference learning remains relatively underexplored despite active interest in Nash-based and Bayesian alternatives.

The taxonomy reveals that SLHF's closest conceptual neighbors are Nash equilibrium approaches (e.g., Nash Learning from Human Feedback, Minimax methods) and multi-step sequential decision processes. While Nash methods model simultaneous best-response dynamics, SLHF explicitly leverages temporal asymmetry and commitment. The taxonomy's scope notes clarify that sequential-move formulations like SLHF are excluded from Nash categories, positioning this work at a boundary between game-theoretic equilibrium concepts. Nearby leaves address Bayesian optimization and experimental design, but these focus on uncertainty quantification or data collection efficiency rather than strategic equilibrium structures.

Among the 30 candidates examined through semantic search and citation expansion, none were found to clearly refute any of the three core contributions: the SLHF framework itself, the STACKELBERGGDA algorithm, or the inference-time refinement capability. Each contribution was assessed against 10 candidates, with zero refutable overlaps identified. This suggests that within the limited search scope, the specific combination of Stackelberg equilibrium modeling, the proposed algorithmic approach, and the refinement mechanism appears distinct from prior work. However, the analysis explicitly acknowledges this is not an exhaustive literature review.

Given the sparse population of the Stackelberg leaf and the absence of refuting candidates in the top-30 semantic matches, the work appears to occupy a relatively novel position within the preference optimization landscape. The limited search scope means potentially relevant work outside the top-30 candidates or in adjacent fields may exist but was not captured. The taxonomy structure indicates this is an emerging rather than saturated research direction, though definitive novelty claims would require broader coverage.

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

Contribution 1: Stackelberg Learning from Human Feedback (SLHF) framework

Description: The authors propose SLHF, a novel preference optimization framework that models alignment as a two-player sequential game between a Leader policy and a Follower policy. Unlike RLHF and NLHF, SLHF leverages sequential play to capture richer preference structures and enables inference-time refinement without requiring scalar 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. Adversarial Preference Learning for Robust LLM Alignment

URL: View paper

Brief Assessment

Adversarial Preference Learning[53] focuses on adversarial robustness through iterative adversarial training with a conditional generative attacker, not on sequential game-theoretic frameworks for preference optimization. The candidate does not address sequential-move games, leader-follower dynamics, or Stackelberg equilibria.

---

2. Adversarial preference optimization: Enhancing your alignment via rm-llm game

URL: View paper

Brief Assessment

Adversarial Preference Optimization[49] focuses on an adversarial game between a reward model and LLM to address distribution shift, not on sequential-move game structures between leader-follower policies for preference optimization and inference-time refinement.

---

3. Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach

URL: View paper

Brief Assessment

Rational Moral Alignment[8] focuses on supervised fine-tuning for rational and moral preferences, not on sequential game-theoretic frameworks for preference optimization or inference-time refinement mechanisms.

---

4. LLM Driven Processes to Foster Explainable AI

URL: View paper

Brief Assessment

LLM Explainable AI[50] focuses on decision-support pipelines using sequential game frameworks for logistics and strategy analysis, not on language model preference optimization or alignment from human feedback.

---

5. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment

URL: View paper

Brief Assessment

Magnetic Preference Optimization[47] focuses on achieving last-iterate convergence to Nash equilibrium in a simultaneous-move two-player game, whereas SLHF proposes a sequential-move Stackelberg game framework with leader-follower dynamics. These represent fundamentally different game-theoretic solution concepts and algorithmic approaches to preference optimization.

---

6. Learning strategic language agents in the werewolf game with iterative latent space policy optimization

URL: View paper

Brief Assessment

Werewolf Latent Policy[52] focuses on strategic language games using game-theoretic methods (CFR) combined with LLM fine-tuning in a latent strategy space, not on preference optimization or human feedback alignment frameworks.

---

7. Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Models Alignment

URL: View paper

Brief Assessment

Magnetic Preference Optimization[48] focuses on achieving last-iterate convergence to Nash equilibrium in self-play settings, while SLHF proposes a fundamentally different sequential-move Stackelberg game framework with leader-follower dynamics rather than simultaneous-move Nash equilibrium.

---

8. Extragradient Preference Optimization (EGPO): Beyond Last-Iterate Convergence for Nash Learning from Human Feedback

URL: View paper

Brief Assessment

Extragradient Preference Optimization[6] focuses on Nash equilibrium computation for simultaneous-move games with non-transitive preferences, while SLHF proposes a fundamentally different sequential-move game framework with leader-follower dynamics. These are distinct solution concepts addressing different game-theoretic formulations.

---

9. Chasing Moving Targets with Online Self-Play Reinforcement Learning for Safer Language Models

URL: View paper

Brief Assessment

Online Self-Play Safety[51] focuses on adversarial safety alignment through attacker-defender co-evolution in a zero-sum game, not preference optimization. The candidate addresses a different problem domain (safety red-teaming) with different objectives (adversarial robustness vs. preference learning).

---

10. Fundamental Limits of Game-Theoretic LLM Alignment: Smith Consistency and Preference Matching

URL: View paper

Brief Assessment

Game-Theoretic LLM Limits[10] focuses on theoretical properties (Condorcet consistency, Smith consistency, preference matching impossibility) of game-theoretic alignment frameworks in general, not on proposing a specific sequential-move Stackelberg framework like SLHF.

Contribution 2: STACKELBERGGDA algorithm

Description: The authors introduce STACKELBERGGDA, a two-timescale gradient descent-ascent algorithm designed to efficiently approximate the unique Stackelberg equilibrium in the SLHF framework. The algorithm performs simultaneous gradient updates on Leader and Follower policies and scales to large language model fine-tuning without requiring explicit 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. A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance

URL: View paper

Brief Assessment

Two-Timescale Primal-Dual[36] focuses on regularized MDPs with occupation measures in standard RL settings, not preference optimization or Stackelberg games between leader-follower policies in LLM alignment.

---

2. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning

URL: View paper

Brief Assessment

Decentralized Policy Evaluation[32] focuses on policy evaluation in cooperative multi-agent reinforcement learning using gradient-tracking methods for decentralized optimization. The candidate addresses a fundamentally different problem (multi-agent policy evaluation) than the original paper (single-agent preference optimization from human feedback), uses different algorithmic frameworks (gradient-tracking for consensus vs. leader-follower sequential games), and targets different objectives (value function estimation vs. preference alignment).

---

3. An online actorâ□□critic algorithm with function approximation for constrained markov decision processes

URL: View paper

Brief Assessment

Constrained Actor-Critic[33] focuses on constrained MDPs with safety constraints using multi-timescale stochastic approximation for actor-critic methods. The ORIGINAL paper's STACKELBERGGDA addresses preference optimization in a Stackelberg game framework for LLM alignment, which is a fundamentally different problem domain and solution concept.

---

4. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation

URL: View paper

Brief Assessment

Gradient Descent-Ascent Convergence[34] focuses on theoretical convergence properties of two-timescale GDA in non-convex, non-concave zero-sum games with finite timescale separation. The original paper's STACKELBERGGDA is designed specifically for preference optimization in LLM alignment using a Stackelberg game formulation with learned preference models, not general game-theoretic equilibria.

---

5. Independent policy gradient methods for competitive reinforcement learning

URL: View paper

Brief Assessment

Independent Policy Gradient[27] addresses two-player zero-sum stochastic games with two-timescale gradient descent-ascent for Nash equilibrium computation, not Stackelberg equilibrium approximation in preference optimization frameworks. The candidate focuses on competitive RL settings where players observe only their own actions/rewards, while SLHF models sequential leader-follower dynamics for LLM alignment with human preferences.

---

6. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems

URL: View paper

Brief Assessment

Alternating Gradient Analysis[31] focuses on bilevel optimization problems with a different mathematical structure (nested optimization with lower-level constraints). While both use alternating gradient updates, the candidate addresses general stochastic bilevel problems rather than the specific Stackelberg game formulation for RLHF policy optimization that STACKELBERGGDA targets.

---

7. Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games

URL: View paper

Brief Assessment

Nash Equilibria Learning[30] focuses on two-timescale algorithms for computing Nash equilibria in general-sum stochastic games, not Stackelberg equilibria in preference optimization. The candidate addresses a fundamentally different game-theoretic solution concept (Nash vs. Stackelberg) and application domain (multi-agent RL in stochastic games vs. LLM alignment from human feedback).

---

8. Convergence Guarantees for Gradient-Based Learning in Continuous Games.

URL: View paper

Brief Assessment

Continuous Games Convergence[35] focuses on convergence guarantees for gradient-based learning in continuous games with Nash equilibria, not on Stackelberg equilibrium approximation or two-timescale gradient descent-ascent for sequential-move games in preference optimization contexts.

---

9. Fast Nonlinear Two-Time-Scale Stochastic Approximation: Achieving Finite-Sample Complexity

URL: View paper

Brief Assessment

Fast Two-Timescale Approximation[29] focuses on general two-timescale stochastic approximation theory and convergence analysis, not on Stackelberg equilibrium approximation for preference optimization in language models.

---

10. Two-timescale Q-learning with function approximation in zero-sum stochastic games

URL: View paper

Brief Assessment

Two-Timescale Q-Learning[28] addresses two-player zero-sum stochastic games with Q-learning for Nash equilibrium computation, not Stackelberg equilibrium approximation in preference optimization. The candidate focuses on game-theoretic RL with payoff-based learning, while the original paper introduces a sequential-move framework for LLM alignment without explicit reward models.

Contribution 3: Inference-time refinement capability

Description: The authors demonstrate that SLHF's Leader-Follower structure naturally supports inference-time refinement, where the Follower policy can improve outputs from the Leader or other models through conditional generation. This capability enables performance gains through additional inference-time computation alone, without requiring further training or external feedback.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Adaptive inference-time compute: Llms can predict if they can do better, even mid-generation

URL: View paper

Brief Assessment

Adaptive Inference-Time Compute[44] focuses on adaptive sampling strategies using self-evaluation to decide when to generate more samples or prune unpromising ones, rather than the Leader-Follower conditional generation structure for refinement that SLHF proposes. The candidate's approach uses a generative reward model for mid-generation predictions, while SLHF's refinement comes from the Follower policy conditionally improving the Leader's committed outputs.

---

2. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review

URL: View paper

Brief Assessment

Reward-Guided Generation[42] focuses on diffusion models for protein design and biological applications, using inference-time guidance to optimize reward functions without fine-tuning. The original paper's contribution concerns language models with a Leader-Follower structure for conditional generation based on human preferences, which is a fundamentally different architecture and application domain.

---

3. Mask-predict: Parallel decoding of conditional masked language models

URL: View paper

Brief Assessment

Mask-Predict[37] focuses on iterative refinement in machine translation through masked language modeling, where the model repeatedly masks and regenerates low-confidence words. This is fundamentally different from SLHF's Leader-Follower structure for preference optimization, which enables refinement through conditional generation based on human preferences rather than confidence-based masking in translation tasks.

---

4. Self-improving language models for evolutionary program synthesis: A case study on ARC-AGI

URL: View paper

Brief Assessment

Self-Improving Program Synthesis[45] focuses on evolutionary program synthesis with iterative refinement during training phases, not inference-time refinement through conditional generation as in SLHF. The candidate's refinement occurs within evolutionary search loops rather than enabling performance gains through inference-time computation alone without additional training.

---

5. Inference-time scaling of diffusion language models with particle gibbs sampling

URL: View paper

Brief Assessment

Particle Gibbs Sampling[43] focuses on trajectory-level refinement for diffusion language models through resampling denoising trajectories, not on conditional generation with a Leader-Follower structure. The technical mechanisms differ fundamentally from SLHF's approach.

---

6. Accelerating blockwise parallel language models with draft refinement

URL: View paper

Brief Assessment

Draft Refinement[39] focuses on refining block drafts in blockwise parallel decoding through lattice rescoring with auxiliary language models, not on conditional generation where a follower policy improves leader outputs through iterative sampling as in SLHF.

---

7. Speculative decoding and beyond: An in-depth survey of techniques

URL: View paper

Brief Assessment

Speculative Decoding Survey[46] focuses on generation-refinement frameworks for accelerating autoregressive decoding through draft-verification mechanisms, not on preference optimization or conditional generation for improving outputs based on human feedback. The survey addresses computational efficiency rather than alignment or quality improvement through iterative refinement.

---

8. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models

URL: View paper

Brief Assessment

Meta-Reasoner[41] focuses on optimizing inference-time reasoning strategies through dynamic strategy adjustment using contextual multi-armed bandits, rather than the Leader-Follower conditional generation structure that enables SLHF's inference-time refinement capability.

---

9. Spar: Self-play with tree-search refinement to improve instruction-following in large language models

URL: View paper

Brief Assessment

Self-Play Refinement[40] focuses on tree-search based self-refinement for instruction-following tasks, where the model refines its own responses through search strategies. The original paper's inference-time refinement operates through a Leader-Follower structure with conditional generation on arbitrary preference signals, which is architecturally and methodologically distinct from tree-search self-play refinement.

---

10. OpenCodeInterpreter: Integrating Code Generation with Execution and Refinement

URL: View paper

Brief Assessment

OpenCodeInterpreter[38] focuses on code generation with execution feedback and iterative refinement in the code domain, while SLHF's inference-time refinement operates through a Leader-Follower structure for general preference optimization without requiring external feedback or domain-specific execution.

Appendix: Text Similarity Detection

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

References

- [0] Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game View paper
- [1] Nash learning from human feedback View paper
- [2] A Minimaximalist Approach to Reinforcement Learning from Human Feedback View paper
- [3] Sequential Preference Ranking for Efficient Reinforcement Learning from Human Feedback View paper
- [4] Sta-rlhf: Stackelberg aligned reinforcement learning with human feedback View paper
- [5] Online Iterative Reinforcement Learning from Human Feedback with General Preference Model View paper
- [6] Extragradient Preference Optimization (EGPO): Beyond Last-Iterate Convergence for Nash Learning from Human Feedback View paper
- [7] Dynamically optimized sequential experimentation (DOSE) for estimating economic preference parameters View paper
- [8] Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach View paper
- [9] Modelling the dynamic vaccination game with evolutionary feedback: exploring pairwise interactions and vaccine strategies View paper
- [10] Fundamental Limits of Game-Theoretic LLM Alignment: Smith Consistency and Preference Matching View paper
- [11] Preference learning along multiple criteria: A game-theoretic perspective View paper
- [12] 5. Sequential pairwise voting View paper
- [13] Cooperative intent: an exploration of computational learning in a discrete preference space View paper
- [14] A parametric approach to Bayesian optimization with pairwise comparisons View paper
- [15] Learning to maximize the social welfare from preference feedback View paper
- [16] Bandits with Preference Feedback: A Stackelberg Game Perspective View paper
- [17] Revealed preferences of individual players in sequential games View paper
- [18] Group sequential two-stage preference designs View paper
- [19] A Game Theoretic Based K-Nearest Neighbor Approach for Binary Classification View paper
- [20] Intransitivity cycles and their transformations: How dynamically adapting systems function View paper
- [21] Finite-time convergence to an -efficient Nash equilibrium in potential games View paper
- [22] Optimal designs for testing pairwise differences: A graphâ□□based game theoretic approach View paper
- [23] A game-theoretic approach to pairwise clustering and matching View paper
- [24] Strategic Manipulation in Group Decisions with Pairwise Comparisons: A Game Theoretical Perspective View paper
- [25] Deterrence and Risk Preferences in Sequential Attacker-Defender Games with Continuous Efforts. View paper
- [26] Multi-Step Preference Optimization via Two-Player Markov Games View paper
- [27] Independent policy gradient methods for competitive reinforcement learning View paper
- [28] Two-timescale Q-learning with function approximation in zero-sum stochastic games View paper
- [29] Fast Nonlinear Two-Time-Scale Stochastic Approximation: Achieving Finite-Sample Complexity View paper
- [30] Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games View paper
- [31] Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems View paper
- [32] Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning View paper
- [33] An online actorâ□□critic algorithm with function approximation for constrained markov decision processes View paper
- [34] Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation View paper
- [35] Convergence Guarantees for Gradient-Based Learning in Continuous Games. View paper
- [36] A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance View paper
- [37] Mask-predict: Parallel decoding of conditional masked language models View paper

- [38] OpenCodeInterpreter: Integrating Code Generation with Execution and Refinement View paper
- [39] Accelerating blockwise parallel language models with draft refinement View paper
- [40] Spar: Self-play with tree-search refinement to improve instruction-following in large language models View paper
- [41] Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models View paper
- [42] Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review View paper
- [43] Inference-time scaling of diffusion language models with particle gibbs sampling View paper
- [44] Adaptive inference-time compute: Llms can predict if they can do better, even mid-generation View paper
- [45] Self-improving language models for evolutionary program synthesis: A case study on ARC-AGI View paper
- [46] Speculative decoding and beyond: An in-depth survey of techniques View paper
- [47] Magnetic preference optimization: Achieving last-iterate convergence for language model alignment View paper
- [48] Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Models Alignment View paper
- [49] Adversarial preference optimization: Enhancing your alignment via rm-llm game View paper
- [50] LLM Driven Processes to Foster Explainable AI View paper
- [51] Chasing Moving Targets with Online Self-Play Reinforcement Learning for Safer Language Models View paper
- [52] Learning strategic language agents in the werewolf game with iterative latent space policy optimization View paper
- [53] Adversarial Preference Learning for Robust LLM Alignment View paper