Novelty Assessment Report

Paper: Token-Importance Guided Direct Preference Optimization
PDF URL: https://openreview.net/pdf?id=cMEnMVvMw9
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Aligning Large Language Models (LLMs) with human preferences is crucial for safe and effective AI interactions. While popular methods like Direct Preference Optimization (DPO) have simplified alignment, they remain sensitive to data noise and overlook the differential importance of individual tokens. Existing token-level approaches often rely on probability prediction or simplistic weighting schemes to obtain token importance, which still cannot fully address these issues. To solve this problem, we propose the Token-Importance Guided Direct Preference Optimization (TI-DPO), a framework that achieves fine-grained semantic control through two synergistic innovations. First, we propose a novel hybrid weighting mechanism that combines gradient attribution with a Gaussian prior, ensuring both the accuracy and robustness of token importance scores. Second, we employ a triplet loss to provide structured guidance for the optimization, explicitly guiding model outputs to approach preferred responses and diverge from non-preferred ones. Experimental results show that TI-DPO achieves higher accuracy and stronger generative diversity, providing more stable and computationally efficient solutions compared with DPO and other RLHF methods.

Core Task Landscape

This paper addresses: **Aligning Large Language Models with Human Preferences**
A total of **50 papers** were analyzed and organized into a taxonomy with **27 categories**.

Taxonomy Overview

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

- **Preference Learning Paradigms and Algorithms**
- **Preference Data and Feedback Sources**
- **Specialized Alignment Objectives and Domains**
- **Personalization and Diverse Preferences**
- **Evaluation and Analysis of Alignment**
- **Alignment Surveys and Overviews**
- **Advanced Alignment Techniques**

Complete Taxonomy Tree

- Aligning Large Language Models with Human Preferences Survey Taxonomy
- Preference Learning Paradigms and Algorithms
  - Reinforcement Learning from Human Feedback (RLHF)
  - RLHF Implementation and Practice (1 papers)
    - [13] ChatGLM-RLHF: Practices of Aligning Large Language Models with Human Feedback (Hou Zhenyu, 2024) View paper
  - AI Feedback as Supervision (2 papers)
    - [11] Rlaif: Scaling reinforcement learning from human feedback with ai feedback (H Lee, 2023) View paper
    - [29] Aligning Large Language Models from Self-Reference AI Feedback with one General Principle (Bao Rong, 2024) View paper
  - RL Algorithm Innovations for LLM Alignment (2 papers)
    - [37] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models (Khaki, 2024) View paper
    - [48] A Technical Survey of Reinforcement Learning Techniques for Large Language Models (Aggarwal, 2025) View paper
  - Direct Preference Optimization
  - Token-Level and Fine-Grained Optimization ★ (2 papers)
    - [0] Token-Importance Guided Direct Preference Optimization (Anon et al., 2026) View paper
    - [12] Aligning Large Language Models via Fine-grained Supervision (Dehong Xu, 2024) View paper
  - Game-Theoretic and Self-Play Methods (1 papers)
    - [41] Self-Play Preference Optimization for Language Model Alignment (Wu Yue, 2024) View paper
  - Ranking and Contrastive Optimization (2 papers)
    - [30] Noise contrastive alignment of language models with explicit rewards (Huayu Chen, 2024) View paper
    - [47] Preference ranking optimization for human alignment (Feifan Song, 2024) View paper
  - Online and Adaptive Preference Optimization (1 papers)
    - [44] Human alignment of large language models through online preference optimisation (Calandriello, 2024) View paper
  - Alternative Alignment Frameworks
  - Representation and Latent Space Methods (2 papers)
    - [5] Aligning Large Language Models with Human Preferences through Representation Engineering (Huang, 2023) View paper
    - [21] Latent Preference Coding: Aligning Large Language Models via Discrete Latent Codes (Gong, 2025) View paper

    - Sampling and Inference-Based Alignment (2 papers)
        - [3] Aligning language models with human preferences (Korbak, 2024) View paper
        - [20] BoNBoN Alignment for Large Language Models and the Sweetness of Best-of-n Sampling (Cristina GÃ¢rbacea, 2024) View paper
    - Flow Matching and Continuous Optimization (1 papers)
        - [22] Preference alignment with flow matching (Song Chong, 2024) View paper
    - Training-Free and Prompt-Based Alignment (2 papers)
    - [15] Black-Box Prompt Optimization: Aligning Large Language Models without Model Training (Jiale Cheng, 2023) View paper
    - [38] Prompt optimization with human feedback (Lin Xiao-Qiang, 2024) View paper
- Preference Data and Feedback Sources
    - Human Preference Data Collection (1 papers)
    - [8] Pku-saferlhf: Towards multi-level safety alignment for llms with human preference (Jiaming Ji, 2025) View paper
    - Implicit and Behavioral Preference Extraction (3 papers)
    - [4] Aligning Large Language Models with Implicit Preferences from User-Generated Content (Zhaoxuan Tan, 2025) View paper
    - [9] Aligning large language models with human preferences using historical text edits (Jan Majkutewicz, 2025) View paper
    - [16] The Real, the Better: Aligning Large Language Models with Online Human Behaviors (Jiang Guan-ying, 2024) View paper
    - Synthetic Preference Generation (2 papers)
    - [24] Learning to summarize from llm-generated feedback (Hwanjun Song, 2025) View paper
    - [43] Aligning large language models through synthetic feedback (Sungdong Kim, 2023) View paper
    - Preference Data Quality and Noise (1 papers)
    - [6] Impact of preference noise on the alignment performance of generative language models (Gao Yang, 2024) View paper
- Specialized Alignment Objectives and Domains
    - Multimodal Alignment (5 papers)
    - [7] Omnialign-v: Towards enhanced alignment of mllms with human preference (Zhao, 2025) View paper
    - [18] Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback (Tianyu Yu, 2024) View paper
    - [31] Vision-R1: Evolving Human-Free Alignment in Large Vision-Language Models via Vision-Guided Reinforcement Learning (Zhan Yufei, 2025) View paper
    - [34] Aligning Audio Captions with Human Preferences (Hegde, 2025) View paper
    - [49] Safe RLHF-V: Safe Reinforcement Learning from Human Feedback in Multimodal Large Language Models (Ji, 2025) View paper
    - Domain-Specific Alignment (3 papers)
    - [2] Pedagogical alignment of large language models (Baraniuk, 2024) View paper
    - [10] Evaluating and aligning codellms on human preference (Yang Jian, 2024) View paper
    - [42] Improving attributed text generation of large language models via preference learning (Dongfang Li, 2024) View paper
    - Safety and Multi-Objective Alignment (1 papers)
    - [46] Multi-Objective Alignment of Large Language Models Through Hypervolume Maximization (Mukherjee, 2024) View paper
- Personalization and Diverse Preferences
    - Modeling Diverse Preference Distributions (2 papers)
    - [14] MaxMin-RLHF: Alignment with diverse human preferences (Chakraborty, 2024) View paper
    - [32] Fine-tuning language models to find agreement among humans with diverse preferences (Bakker, 2022) View paper
    - Interactive and Adaptive Personalization (2 papers)
    - [40] Persona-judge: Personalized Alignment of Large Language Models via Token-level Self-judgment (Xiaotian Zhang, 2025) View paper
    - [50] Aligning llms with individual preferences via interaction (Wu Shujin, 2025) View paper
    - Personalization Frameworks and Policies (2 papers)
    - [33] Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback (Kirk, 2023) View paper
    - [36] A Survey on Personalized Alignment--The Missing Piece for Large Language Models in Real-World Applications (Jian Guan, 2025) View paper
    - Lifelong and Continual Preference Learning (1 papers)
    - [39] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization (Li Junsong, 2025) View paper
- Evaluation and Analysis of Alignment
    - Alignment Evaluation Benchmarks (2 papers)
    - [25] Aligning with human judgement: The role of pairwise preference in large language model evaluators (Liu, 2024) View paper
    - [35] HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition (Yuxuan Liu, 2024) View paper
    - Theoretical Analysis of Alignment Methods (2 papers)
    - [19] On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization (Jiancong Xiao, 2024) View paper
    - [28] Rethinking reward modeling in preference-based large language model alignment (H Sun, 2025) View paper
    - Preference Consistency and Reasoning (2 papers)
    - [17] Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences? (Gu ZhuoJun, 2025) View paper
    - [23] Can language models reason about individualistic human values and preferences? (Liwei Jiang, 2025) View paper
- Alignment Surveys and Overviews (3 papers)
    - [1] Aligning large language models with human: A survey (Wang Yu-Fei, 2023) View paper
    - [26] Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges (Lu Haoran, 2025) View paper
    - [27] A survey on human preference learning for large language models (Jiang, 2024) View paper
- Advanced Alignment Techniques (1 papers)
    - [45] Aligning Large Language Models with Counterfactual DPO (Butcher, 2024) View paper

Narrative

Core task: aligning large language models with human preferences. The field has matured into a rich taxonomy spanning several major branches. Preference Learning Paradigms and Algorithms explores foundational methods—ranging from reinforcement learning from human feedback (RLHF) to direct preference optimization (DPO) and its token-level refinements—that translate preference signals into model updates. Preference Data and Feedback Sources examines where preference information originates, including human annotations, AI-generated feedback, and implicit behavioral cues. Specialized Alignment Objectives and Domains addresses domain-specific challenges such as safety, code generation, and multimodal tasks, while Personalization and Diverse Preferences investigates how models can respect heterogeneous user values. Evaluation and Analysis of Alignment provides benchmarks and diagnostic tools to measure alignment quality, and Alignment Surveys and Overviews offer integrative perspectives on the rapidly evolving landscape. Advanced Alignment Techniques captures emerging innovations in optimization and representation engineering that push beyond standard paradigms.

Within Preference Learning Paradigms, a particularly active line of work focuses on token-level and fine-grained optimization, moving beyond coarse sequence-level rewards to credit assignment at finer granularities. Token-Importance DPO[0] exemplifies this direction by weighting tokens according to their contribution to preference outcomes, aiming to sharpen the learning signal where it matters most. This contrasts with approaches like Fine-grained Supervision[12], which also targets sub-sequence structure but may emphasize different decomposition strategies or supervision sources. Meanwhile, broader DPO variants explore noise robustness, online learning, and multi-objective trade-offs, reflecting ongoing debates about how to balance sample efficiency, stability, and alignment fidelity. Token-Importance DPO[0] sits naturally among these fine-grained methods, sharing the goal of more precise credit assignment while differing in its specific mechanism for identifying and emphasizing critical tokens during optimization.

Related Works in Same Category

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

1. Aligning Large Language Models via Fine-grained Supervision

Authors: Dehong Xu, Liang Qiu, Min-Seok Kim, Faisal Ladhak, Jaeyoung Do | **Year/Venue**: 2024 • Annual Meeting of the Association for Computational Linguistics | **URL**: View paper

Abstract

Pre-trained large-scale language models (LLMs) excel at producing coherent articles, yet their outputs may be untruthful, toxic, or fail to align with user expectations. Current approaches focus on using reinforcement learning with human feedback (RLHF) to improve model alignment, which works by transforming coarse human preferences of LLM outputs into a feedback signal that guides the model learning process. However, because this approach operates on sequence-level feedback, it lacks the precis...

Relationship Analysis

Both papers belong to the Token-Level and Fine-Grained Optimization category, addressing the limitation of sequence-level preference learning by providing granular, token-level supervision signals for alignment. The candidate paper proposes a fine-grained PPO approach using minimal edits to create token-level reward signals, while the original paper (TI-DPO) introduces a hybrid weighting mechanism combining gradient attribution with Gaussian priors and employs triplet loss for direct preference optimization. The key difference lies in their optimization frameworks: the candidate uses token-level rewards within RL-based PPO, whereas TI-DPO operates in the direct preference optimization paradigm with importance-weighted tokens and structured triplet objectives.

Contributions Analysis

Overall novelty summary. The paper proposes TI-DPO, a token-level direct preference optimization framework combining gradient-based importance weighting with Gaussian priors and triplet loss guidance. It resides in the Token-Level and Fine-Grained Optimization leaf, which contains only two papers including this one. This leaf sits within the broader Direct Preference Optimization branch, indicating a relatively sparse but emerging research direction focused on granular credit assignment beyond sequence-level optimization.

The taxonomy reveals that token-level methods occupy a small niche within DPO, which itself branches into game-theoretic approaches, ranking-based methods, and online optimization. Neighboring leaves address noise robustness and contrastive learning at the sequence level, while the broader Preference Learning Paradigms category includes RLHF variants and alternative frameworks like representation engineering. The scope notes clarify that token-level methods emphasize fine-grained supervision signals, distinguishing them from coarser sequence-level or game-theoretic formulations in sibling categories.

Among 29 candidates examined, the core TI-DPO framework contribution shows three refutable candidates from nine examined, suggesting moderate prior work overlap in token-level preference optimization. The hybrid weighting mechanism examined ten candidates with zero refutations, indicating this specific combination of gradient attribution and Gaussian priors may be less explored. The theoretical analysis contribution also found no refutations across ten candidates, though this reflects the limited search scope rather than exhaustive coverage of theoretical alignment literature.

Based on top-29 semantic matches, the work appears to occupy a relatively novel position within token-level DPO methods, particularly in its hybrid weighting design. However, the limited search scope and presence of some overlapping prior work in the core framework suggest careful positioning relative to existing fine-grained optimization approaches would strengthen claims of distinctiveness.

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

Contribution 1: Token-Importance Guided Direct Preference Optimization (TI-DPO) framework

Description: The authors introduce TI-DPO, a novel alignment framework that combines a hybrid weighting mechanism (gradient attribution with Gaussian prior) and triplet loss to achieve fine-grained control over token-level importance in preference optimization, addressing limitations of sequence-level methods like DPO.

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

---

1. Optimal Transport-Based Token Weighting scheme for Enhanced Preference Optimization

URL: View paper

Prior Art Analysis

Optimal Transport Weighting[65] demonstrates that prior work exists on token-level preference optimization with fine-grained semantic control. Both papers address the same fundamental limitation: that sequence-level methods like DPO treat all tokens equally, ignoring differential token importance. The candidate paper proposes a token-level weighting mechanism based on optimal transport to identify semantically meaningful tokens, while the original paper proposes gradient attribution combined with Gaussian prior. Both methods decompose DPO loss at the token level and apply dynamic token weighting schemes to achieve fine-grained control, directly challenging the novelty claim that TI-DPO was the first to combine token-level importance weighting with preference optimization.

Evidence

Evidence 1 - **Rationale**: Both papers identify the identical core problem: DPO's uniform token treatment leads to suboptimal alignment by allowing less important tokens to dominate the optimization signal. - **Original**: while popular methods like direct preference

optimization (dpo) have simplified alignment, they remain sensitive to data noise and overlook the differential importance of individual tokens. existing token-level approaches often rely on probability prediction or simplistic weighting schemes to obtain... - **Candidate**: The dpo loss treats each token equally, which can bias the model to overlook less important factors and learn by shortcuts, leading to suboptimal results (park et al., 2024). In Fig. 1, tokens less relevant to the question dominate the reward, while important parts like "cat likes to eat fish" shoul...

Evidence 2 - **Rationale**: Both papers propose token-level weighting mechanisms for DPO to achieve fine-grained semantic control. The candidate's optimal transport approach predates the original's gradient attribution method, demonstrating prior work on this contribution. - **Original**: we propose the token-importance guided direct preference optimization (ti-dpo), a framework that achieves fine-grained semantic control through two synergistic innovations. first, we propose a novel hybrid weighting mechanism that combines gradient attribution with a gaussian prior, ensuring both th... - **Candidate**: we propose an optimal transport-based weighting scheme for direct preference optimization (otpo), a novel unsupervised framework for calculating token weights in direct preference optimization. Our key innovation lies in emphasizing tokens where the responses agree (similar tokens) as indicators of ...

Evidence 3 - **Rationale**: Both methods dynamically assign token-level weights to identify important tokens. The candidate's optimal transport method provides an alternative approach to the same problem the original claims to solve first. - **Original**: we introduce a novel hybrid weighting mechanism to accurately and robustly identify key tokens. this mechanism combines gradient attribution with a gaussian prior, overcoming the problem of existing methods relying on biased proxies. - **Candidate**: we utilize an unbalanced optimal transport approach to dynamically assign a fixed total weight budget to token-level weights based on the similarity between tokens in chosen and rejected responses, allocating higher weights to more semantically relevant tokens. This allows for estimating the minimum...

---

2. Selective preference optimization via token-level reward function estimation

URL: View paper

Prior Art Analysis

Selective Preference Optimization[67] demonstrates that token-level preference optimization with fine-grained control was already explored prior to the original paper's submission. The candidate paper proves DPO can serve as a token-level reward function estimator (Theorem 1) and proposes selective optimization on key tokens identified through gradient-based scoring. While the original paper introduces a hybrid weighting mechanism combining gradient attribution with Gaussian prior, the candidate already established the core concept of using DPO-derived token-level rewards for selective training. Both papers address the same fundamental problem: moving beyond sequence-level optimization to token-level control, with the candidate providing theoretical foundations (Theorem 1) that enable token selection without requiring the original paper's specific Gaussian prior correction.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same core challenge: not all tokens contribute equally to preference alignment, and key token identification is crucial. The candidate provides empirical evidence that key tokens dominate rewards, motivating selective optimization. - **Original**: achieving fine-grained alignment requires addressing a core challenge: we not only need to accurately identify the key tokens that have a decisive impact on human preferences, but also need a subtle optimization objective to guide the model to adjust its preference - **Candidate**: though achieving outstanding performance, most of these methods are optimized on all available tokens from the training dataset. to validate the effectiveness of this setup, in figure 1, we present the token-level reward accumulations for 1,000 samples from an instruction following dataset (cui et a...

Evidence 2 - **Rationale**: The original paper uses gradient attribution to compute token importance. The candidate provides the theoretical foundation (Theorem 1) showing DPO inherently estimates token-level rewards through the log-ratio of policy and reference models, which is mathematically equivalent to gradient-based importance. - **Original**: this mechanism combines gradient attribution with a gaussian prior, overcoming the problem of existing methods relying on biased proxies. here, gradient attribution is a technique used to determine the contribution of each input feature (in our work, each token) to the model's output - **Candidate**: theorem 1. with a reference model πref, fitting any reward functions r that are consistent to the bradley-terry model with the dpo algorithm leads to an optimal estimation of another reward function ˆrthat decouples the response-level reward values into the token level, which satisfies: ˆr(st,at) ∝l...

Evidence 3 - **Rationale**: Both papers employ contrastive objectives for token-level optimization. While the original uses triplet loss with intermediate outputs, the candidate uses a contrastive objective on selected tokens. Both achieve fine-grained preference alignment through structured comparison. - **Original**: we adopt a structured triplet objective based on the identified key weights to achieve fine-grained optimization by incorporating the intermediate generated outputs (nguyen et al., 2018). this triplet structure explicitly guides the intermediate output to approach human preferences and distance from... - **Candidate**: we design a simple contrastive preference optimization objective (meng et al., 2024) on the target policy model πt with the selected tokens. specifically, the objective function lsepo is designed as follows: -e(q,yw,yl)∼dlog σ (ˆu(q,yw,iw kw) -ˆu(q,yl,il kl) -λ)

Evidence 4 - **Rationale**: Both papers critique existing token-level methods for inefficient or biased token selection. The candidate addresses this by proposing DPO-based selection, while the original proposes gradient attribution with Gaussian prior. Both aim to solve the same identified limitation in prior work. - **Original**: existing token-level methods fall short in dealing with this challenge for two reasons. first, their approaches to identifying key tokens often rely on biased probability proxies (liu et al., 2024a) or overly simplified weighting schemes (lin et al., 2024). - **Candidate**: some works explored only optimizing on selected response fragments, but their selection strategies are complex and expensive, utilizing monte-carlo tree search (xie et al., 2024; chen et al., 2024b) or annotations from human/capable llms (lai et al., 2024; yoon et al., 2024). the above limitations u...

---

3. Optimizing human-controlled preference alignment in large language models via dense token masking: A methodological approach

URL: View paper

Brief Assessment

Dense Token Masking[68] focuses on vector representation-level alignment and semantic consistency with preferences. The candidate does not provide sufficient technical detail about token-level importance weighting mechanisms or gradient attribution methods that would refute TI-DPO's novelty claim of combining hybrid weighting (gradient attribution with Gaussian prior) and triplet loss for fine-grained token-level control.

---

4. TGDPO: Harnessing Token-Level Reward Guidance for Enhancing Direct Preference Optimization

URL: View paper

Brief Assessment

TGDPO[66] focuses on decomposing sequence-level PPO into token-level problems and deriving closed-form optimal policies with token-level reward guidance for DPO. In contrast, the original paper's TI-DPO uses a hybrid weighting mechanism combining gradient attribution with Gaussian prior and triplet loss for fine-grained semantic control. These are distinct technical approaches to token-level preference optimization.

---

5. Token-level proximal policy optimization for query generation

URL: View paper

Brief Assessment

Token-level PPO[62] focuses on query generation for web search engines using reinforcement learning with token-level rewards, while TI-DPO addresses preference optimization for LLM alignment using gradient attribution and triplet loss. These are fundamentally different technical approaches for different tasks.

---

6. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment

URL: View paper

Brief Assessment

Fine-grained Verifiers[69] focuses on vision-language alignment using visual encoders as verifiers for token-level rewards in VLLMs, not general LLM preference optimization frameworks like TI-DPO.

---

7. Fine-grained Preference Optimization Improves Zero-shot Text-to-Speech

URL: View paper

Brief Assessment

Fine-grained TTS Optimization[63] focuses on speech synthesis with temporal modeling and semantic-phonetic alignment errors, not general LLM token-level preference optimization. The candidate addresses speech-specific issues (mispronunciation, prosody) rather than text generation alignment.

---

8. Fine-grained video dubbing duration alignment with segment supervised preference optimization

URL: View paper

Brief Assessment

Video Dubbing Alignment[61] addresses duration alignment in video dubbing translation using segment-level preference optimization (SSPO), not token-level importance weighting for general LLM alignment. The tasks and technical approaches differ fundamentally.

---

9. T-reg: Preference optimization with token-level reward regularization

URL: View paper

Prior Art Analysis

T-reg[64] demonstrates that prior work exists on token-level preference optimization with fine-grained control mechanisms. Both papers address the same core problem: moving beyond sequence-level rewards in preference optimization to achieve token-level credit assignment. T-reg[64] proposes using token-level rewards as regularization during preference optimization, while the original paper proposes TI-DPO with hybrid weighting and triplet loss. The candidate paper's approach of integrating token-level rewards into preference optimization frameworks predates or parallels the original's claims of novelty in this space.

Evidence

Evidence 1 - **Rationale**: Both papers claim novelty in token-level preference optimization frameworks. T-reg[64] explicitly proposes using token-level rewards as regularization to guide sequence-level optimization, which directly overlaps with the original paper's claim of achieving 'fine-grained semantic control' through token-level mechanisms. - **Original**: we propose the token-importance guided direct preference optimization (ti-dpo), a framework that achieves fine-grained semantic control through two synergistic innovations. first, we propose a novel hybrid weighting mechanism that combines gradient attribution with a gaussian prior, ensuring both th... - **Candidate**: we propose token-level reward regularization (t-reg), a novel approach that leverages both sequence-level and token-level rewards for preference optimization. harnessing the self-refinement capabilities of llms, our method uses contrastive prompting to enable llms to self-generate token-level reward...

Evidence 2 - **Rationale**: Both papers identify the same limitation of sequence-level optimization and propose token-level solutions. T-reg[64] explicitly references prior work on token-level rewards, suggesting this is an established research direction rather than a novel contribution. - **Original**: however, both dpo and rlhf have a fundamental flaw during optimization: they optimize at the sequence level, leading to the neglect of the influence of specific tokens, which in turn destabilizes the training process due to shifts in the sampling distribution - **Candidate**: current rlhf methods typically use sequencelevel rewards, where reward signals are provided at the end of an entire sequence. recent studies have shown that incorporating finer-grained reward signals (wu et al., 2024; lightman et al., 2023), ultimately at the token level (zhong et al., 2024), can si...

Evidence 3 - **Rationale**: T-reg[64] demonstrates that combining sequence-level and token-level rewards for preference optimization was already proposed. The candidate's approach of using token-level rewards as 'weak supervision' to guide sequence-level optimization parallels the original's claim of achieving fine-grained control. - **Original**: achieving fine-grained alignment requires the model not only to distinguish the quality of the entire sequence, but also to understand and precisely control the key morphemes that constitute the semantics of the sequence. existing sequence-level techniques often lead to a decrease in generation dive... - **Candidate**: we introduce t-reg, a novel tokenlevel preference optimization algorithm that leverages both sequence-level and token-level rewards. our approach is motivated by the observation that dpo inherently learns token-level rewards. instead of directly optimizing the policy model with auto-labeled token-le...

Evidence 4 - **Rationale**: Both papers claim novelty in methods for token-level credit assignment in preference optimization. T-reg[64]'s approach of deriving token-level rewards to guide optimization demonstrates that similar fine-grained alignment mechanisms were already established in the literature. - **Original**: we propose ti-dpo, a novel framework designed for achieving fine-grained alignment. this framework innovatively integrates a hybrid weighting mechanism, jointly formed by gradient attribution and a gaussian prior, with triplet loss, significantly enhancing the robustness and stability of weight allo... - **Candidate**: t-reg utilizes token-level rewards derived through contrastive prompting to guide the token-level rewards learned during preference optimization, enabling effective token-level credit assignment without the need for external token-level reward annotations.

Contribution 2: Hybrid weighting mechanism combining gradient attribution and Gaussian prior

Description: A new method for computing token importance that merges gradient-based attribution with a Gaussian prior distribution to counteract architectural biases (such as Lost-in-the-Middle) and provide stable, accurate token weights for preference alignment.

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. Investigating mysteries of cot-augmented distillation

URL: View paper

Brief Assessment

CoT-Augmented Distillation[53] uses gradient attribution (integrated gradients) to identify important tokens in chain-of-thought rationales for model distillation, not for token importance weighting in preference alignment. The candidate focuses on distillation from teacher to student models in reasoning tasks, while the original addresses preference optimization in LLM alignment with a hybrid weighting mechanism that combines gradient attribution with a Gaussian prior to counteract architectural biases.

---

2. CipherPrune: Efficient and Scalable Private Transformer Inference

URL: View paper

Brief Assessment

CipherPrune[52] focuses on encrypted token pruning for private transformer inference using cryptographic protocols, not on preference alignment or combining gradient attribution with Gaussian priors for token importance in RLHF contexts.

---

3. Learning explainable models using attribution priors

URL: View paper

Brief Assessment

Attribution Priors Learning[56] uses gradient-based attribution methods (expected gradients) for regularization during training, but does not combine gradient attribution with a Gaussian prior for token importance weighting in preference optimization contexts. The candidate focuses on regularizing feature attributions to encode domain knowledge, not on addressing architectural biases like Lost-in-the-Middle in LLM alignment.

---

4. Gradient based feature attribution in explainable ai: A technical review

URL: View paper

Brief Assessment

Gradient Feature Attribution[51] is a technical review of gradient-based feature attribution methods in explainable AI for neural networks, not a method for token importance weighting in preference alignment for LLMs. The candidate focuses on explaining model predictions through gradient-based saliency maps, while the original contribution addresses token-level optimization in direct preference optimization with architectural bias correction.

---

5. Incorporating priors with feature attribution on text classification

URL: View paper

Brief Assessment

Attribution Priors Classification[54] uses gradient attribution (integrated gradients) to compute token importance for text classification, but does not combine it with a Gaussian prior to counteract architectural biases like Lost-in-the-Middle. The candidate focuses on incorporating priors through L2 distance loss between attributions and target values for fairness and data scarcity, not on addressing positional biases in LLM preference alignment.

---

6. The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

URL: View paper

Brief Assessment

OOD Explainability Problem[57] focuses on feature importance explanations for model interpretability and addresses out-of-distribution issues in counterfactual inputs during explanation generation. It does not propose gradient attribution combined with priors for token importance weighting in preference alignment contexts.

---

7. Structural prior-driven feature extraction with gradient-momentum combined optimization for convolutional neural network image classification.

URL: View paper

Brief Assessment

Structural Prior Feature[60] focuses on convolutional neural network image classification with structural priors for feature extraction, not on token importance weighting for language model alignment using gradient attribution combined with Gaussian priors.

---

8. Better Explain Transformers by Illuminating Important Information

URL: View paper

Brief Assessment

Illuminating Important Information[55] focuses on refining attention-based explanations in transformers by masking irrelevant information flows, not on token importance weighting for preference alignment. The candidate does not address preference optimization, DPO, or combining gradient attribution with priors for alignment tasks.

---

9. On the Interaction of Belief Bias and Explanations

URL: View paper

Brief Assessment

Belief Bias Explanations[58] focuses on gradient-based attribution methods for explainability evaluation in reading comprehension, not on token importance weighting for preference alignment in LLMs. The paper does not address preference optimization, DPO, or architectural biases like Lost-in-the-Middle.

---

10. A Weibull gradient prior for image restoration

URL: View paper

Brief Assessment

Weibull Gradient Prior[59] focuses on image restoration using Weibull distributions for gradient priors in computer vision, not token importance weighting in language models or preference alignment.

Contribution 3: Theoretical analysis proving TI-DPO superiority over DPO

Description: The authors provide formal theoretical guarantees demonstrating that TI-DPO achieves a strictly lower loss bound compared to standard DPO and yields higher expected rewards under fixed KL divergence constraints, offering a rigorous foundation for the method's empirical advantages.

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. Statistical Rejection Sampling Improves Preference Optimization

URL: View paper

Brief Assessment

Statistical Rejection Sampling[71] focuses on improving preference optimization through rejection sampling and unified loss functions (logistic vs. hinge), not on proving theoretical loss bounds comparing token-level versus sequence-level DPO methods.

---

2. Human-in-the-loop: Real-time Preference Optimization

URL: View paper

Brief Assessment

Real-time Preference Optimization[74] focuses on online feedback optimization with pairwise comparisons in engineering control systems, not on theoretical loss bounds comparing preference optimization methods like DPO and TI-DPO for language model alignment.

---

3. Personalizing reinforcement learning from human feedback with variational preference learning

URL: View paper

Brief Assessment

Variational Preference Learning[76] focuses on multi-modal reward modeling for diverse user preferences using variational inference, not on proving loss bounds comparing token-level versus sequence-level preference optimization methods like TI-DPO does.

---

4. Uncertainty-penalized direct preference optimization

URL: View paper

Brief Assessment

Uncertainty-Penalized DPO[73] focuses on uncertainty penalization schemes to address overfitting in DPO, not on proving loss bounds or expected reward superiority through token-level importance weighting and triplet loss as in the original paper.

---

5. On the generalization of preference learning with dpo

URL: View paper

Brief Assessment

DPO Generalization[72] focuses on generalization under diverse human values with finite-step training dynamics, not on token-level importance weighting or triplet loss mechanisms that form the basis of TI-DPO's theoretical superiority claims.

---

6. Design considerations in offline preference-based rl

URL: View paper

Brief Assessment

Offline Preference Design[80] focuses on general design choices (loss functions, base policies, constraints) across multiple offline RLHF methods rather than comparing a specific token-level method against DPO. The theoretical framework analyzes curvature properties and coverage assumptions but does not address token-importance weighting or triplet loss mechanisms that are central to TI-DPO's claimed novelty.

---

7. Multiplayer Nash Preference Optimization

URL: View paper

Brief Assessment

Multiplayer Nash Optimization[75] focuses on multi-player game-theoretic frameworks for preference optimization, not on token-level importance weighting or loss bound comparisons between DPO variants.

---

8. Adversarial Policy Optimization for Offline Preference-based Reinforcement Learning

URL: View paper

Brief Assessment

Adversarial Policy Optimization[77] focuses on offline preference-based RL with trajectory-level preferences and sample complexity bounds, not on token-level direct preference optimization or loss bound comparisons between DPO variants.

---

9. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning

URL: View paper

Brief Assessment

Negative Preference Optimization[79] focuses on unlearning tasks and catastrophic collapse prevention, not on proving superiority of token-importance guided methods over standard DPO for alignment tasks.

---

10. Mitigating Hallucination Through Theory-Consistent Symmetric Multimodal Preference Optimization

URL: View paper

Brief Assessment

Symmetric Multimodal Optimization[78] focuses on multimodal hallucination mitigation through symmetric preference learning with vision-oriented objectives, not on general token-level theoretical loss bounds for preference optimization methods.

Appendix: Text Similarity Detection

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

References

- [0] Token-Importance Guided Direct Preference Optimization View paper
- [1] Aligning large language models with human: A survey View paper
- [2] Pedagogical alignment of large language models View paper
- [3] Aligning language models with human preferences View paper
- [4] Aligning Large Language Models with Implicit Preferences from User-Generated Content View paper
- [5] Aligning Large Language Models with Human Preferences through Representation Engineering View paper
- [6] Impact of preference noise on the alignment performance of generative language models View paper
- [7] Omnialign-v: Towards enhanced alignment of mllms with human preference View paper
- [8] Pku-saferlhf: Towards multi-level safety alignment for llms with human preference View paper
- [9] Aligning large language models with human preferences using historical text edits View paper
- [10] Evaluating and aligning codellms on human preference View paper
- [11] Rlaif: Scaling reinforcement learning from human feedback with ai feedback View paper
- [12] Aligning Large Language Models via Fine-grained Supervision View paper
- [13] ChatGLM-RLHF: Practices of Aligning Large Language Models with Human Feedback View paper
- [14] MaxMin-RLHF: Alignment with diverse human preferences View paper

- [15] Black-Box Prompt Optimization: Aligning Large Language Models without Model Training View paper
- [16] The Real, the Better: Aligning Large Language Models with Online Human Behaviors View paper
- [17] Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences? View paper
- [18] Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback View paper
- [19] On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization View paper
- [20] BoNBoN Alignment for Large Language Models and the Sweetness of Best-of-n Sampling View paper
- [21] Latent Preference Coding: Aligning Large Language Models via Discrete Latent Codes View paper
- [22] Preference alignment with flow matching View paper
- [23] Can language models reason about individualistic human values and preferences? View paper
- [24] Learning to summarize from llm-generated feedback View paper
- [25] Aligning with human judgement: The role of pairwise preference in large language model evaluators View paper
- [26] Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges View paper
- [27] A survey on human preference learning for large language models View paper
- [28] Rethinking reward modeling in preference-based large language model alignment View paper
- [29] Aligning Large Language Models from Self-Reference AI Feedback with one General Principle View paper
- [30] Noise contrastive alignment of language models with explicit rewards View paper
- [31] Vision-R1: Evolving Human-Free Alignment in Large Vision-Language Models via Vision-Guided Reinforcement Learning View paper
- [32] Fine-tuning language models to find agreement among humans with diverse preferences View paper
- [33] Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback View paper
- [34] Aligning Audio Captions with Human Preferences View paper
- [35] HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition View paper
- [36] A Survey on Personalized Alignment--The Missing Piece for Large Language Models in Real-World Applications View paper
- [37] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models View paper
- [38] Prompt optimization with human feedback View paper
- [39] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization View paper
- [40] Persona-judge: Personalized Alignment of Large Language Models via Token-level Self-judgment View paper
- [41] Self-Play Preference Optimization for Language Model Alignment View paper
- [42] Improving attributed text generation of large language models via preference learning View paper
- [43] Aligning large language models through synthetic feedback View paper
- [44] Human alignment of large language models through online preference optimisation View paper
- [45] Aligning Large Language Models with Counterfactual DPO View paper
- [46] Multi-Objective Alignment of Large Language Models Through Hypervolume Maximization View paper
- [47] Preference ranking optimization for human alignment View paper
- [48] A Technical Survey of Reinforcement Learning Techniques for Large Language Models View paper
- [49] Safe RLHF-V: Safe Reinforcement Learning from Human Feedback in Multimodal Large Language Models View paper
- [50] Aligning llms with individual preferences via interaction View paper
- [51] Gradient based feature attribution in explainable ai: A technical review View paper
- [52] CipherPrune: Efficient and Scalable Private Transformer Inference View paper
- [53] Investigating mysteries of cot-augmented distillation View paper
- [54] Incorporating priors with feature attribution on text classification View paper
- [55] Better Explain Transformers by Illuminating Important Information View paper
- [56] Learning explainable models using attribution priors View paper
- [57] The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations View paper
- [58] On the Interaction of Belief Bias and Explanations View paper
- [59] A Weibull gradient prior for image restoration View paper
- [60] Structural prior-driven feature extraction with gradient-momentum combined optimization for convolutional neural network image classification. View paper
- [61] Fine-grained video dubbing duration alignment with segment supervised preference optimization View paper
- [62] Token-level proximal policy optimization for query generation View paper
- [63] Fine-grained Preference Optimization Improves Zero-shot Text-to-Speech View paper
- [64] T-reg: Preference optimization with token-level reward regularization View paper
- [65] Optimal Transport-Based Token Weighting scheme for Enhanced Preference Optimization View paper
- [66] TGDPO: Harnessing Token-Level Reward Guidance for Enhancing Direct Preference Optimization View paper
- [67] Selective preference optimization via token-level reward function estimation View paper
- [68] Optimizing human-controlled preference alignment in large language models via dense token masking: A methodological approach View paper
- [69] Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment View paper
- [70] ASPO: Adaptive Sentence-Level Preference Optimization for Fine-Grained Multimodal Reasoning View paper
- [71] Statistical Rejection Sampling Improves Preference Optimization View paper
- [72] On the generalization of preference learning with dpo View paper
- [73] Uncertainty-penalized direct preference optimization View paper
- [74] Human-in-the-loop: Real-time Preference Optimization View paper
- [75] Multiplayer Nash Preference Optimization View paper
- [76] Personalizing reinforcement learning from human feedback with variational preference learning View paper
- [77] Adversarial Policy Optimization for Offline Preference-based Reinforcement Learning View paper
- [78] Mitigating Hallucination Through Theory-Consistent Symmetric Multimodal Preference Optimization View paper
- [79] Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning View paper
- [80] Design considerations in offline preference-based rl View paper