Multi-level Style Preference Optimization: An Adaptive Detection Framework for Human-Machine Hybrid Text

Zehao Wang[1,2], Lianwei Wu[1,2*], Wenbo An[1], Hang Zhang[1], Yaxiong Wang[3]
[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Research Development Institute of Northwestern Polytechnical University in Shenzhen, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, China
202526923@mail.nwpu.edu.cn, wlw@nwpu.edu.cn, 2074137878@mail.nwpu.edu.cn, zh2003@mail.nwpu.edu.cn, wangyx15@stu.xjtu.edu.cn

Abstract

Large language model (LLM) generated texts now rival human quality, creating four text categories: purely machine-generated, machine-rewritten, machine-polished, and human-written content. Traditional detection methods face significant challenges in human-machine hybrid scenarios where LLMs perform rewriting or polishing, as existing approaches focus on single-level features and fail to capture subtle, multi-layered machine traces. To address this, we propose the Multi-level Style Preference Optimization (MSPO) framework, capturing machine style features at multiple granularities: sequence-level (overall consistency), phrase-level (distinctive n-gram patterns), and lexical-level (word selection distributions). We further incorporate four text complexity indicators (Type-Token Ratio, Average Sentence Length, Average Word Length, and Punctuation Ratio) to dynamically adjust optimization parameters based on human-machine text complexity differences, enhancing adaptability across diverse text types. Additionally, we construct a comprehensive detection dataset spanning three representative domains (scientific writing, news articles, and creative writing) across four text types (human-written, purely machine-generated, machine-rewritten, and machine-polished), generated using state-of-the-art LLMs for robust evaluation. Experimental results demonstrate that MSPO significantly outperforms existing methods across all text types. On the challenging rewritten texts, MSPO achieves up to 82.14% AUROC, representing an improvement of 11.15 percentage points over the strongest baseline ImBD, while maintaining robust cross-domain generalizability across scientific, news, and creative writing domains.

Code — https://github.com/wxdtk2/MSPO

Introduction

Large language models (LLMs) have achieved breakthrough progress, generating text nearly indistinguishable from human writing in fluency and coherence (OpenAI 2023a,b; Chowdhery et al. 2023; Touvron et al. 2023). However, their widespread application in news reporting, academic research, and content creation has raised serious concerns regarding academic integrity and information authenticity (M Alshater 2022; Wu et al. 2023b,a, 2025), making effective detection methods increasingly urgent (Bao et al. 2024; Mai et al. 2024; He et al. 2025; Tian et al. 2025).

In practical applications, the interaction between humans and LLMs has resulted in four distinct text categories, each presenting varying detection challenges. **Human-written text** serves as the baseline, consisting entirely of human-created content without any LLM processing. **Pure machine-generated text** is entirely produced by language models based on brief prompts, containing no human-authored content. More challenging are the hybrid scenarios: **Machine-rewritten text** emerges when LLMs participate in local content rewriting, where portions of human text are selectively rewritten while maintaining the original meaning but using different phrasing and vocabulary. **Machine-polished text** represents the most subtle form, where LLMs perform sentence-level enhancement to improve clarity and flow while preserving the core meaning and most of the original structure.

These human-machine hybrid text types, particularly machine-rewritten and machine-polished content, pose significant detection challenges because the boundaries between human and machine contributions become increasingly blurred. The subtle modifications often alter stylistic characteristics while preserving the majority of human content, making traditional detection approaches insufficient for these sophisticated scenarios (Zhang et al. 2024; Chawla 2024; Sadasivan et al. 2023).

Current detection methods can be broadly categorized into two types based on their target text scenarios. **For purely machine-generated text**, existing approaches include training-based methods (such as GPT-2 detector (Solaiman et al. 2019), RoBERTa-based detectors (Liu et al. 2019), and GPTZero (Tian and Cui 2023)) and zero-shot methods based on probability distributions (Gehrmann, Strobelt, and Rush 2019), text reconstruction (Yang et al. 2023), or probability curvature (Mitchell et al. 2023; Bao et al. 2024; Su et al. 2023). These methods are effective on purely machine-generated content by leveraging statistical and linguistic patterns. **For human-machine hybrid texts**, detection methods remain limited. Recent work like ImBD (Chen et al. 2025) addresses hybrid text detection through distributional feature imitation, while CheckGPT (Liu et al. 2024) develops a deep neural framework to capture semantic

*Corresponding author.

and linguistic patterns in ChatGPT-generated academic writing. Despite these efforts, existing methods typically operate at a single granularity level and lack adaptive mechanisms for different text complexities, limiting their effectiveness in sophisticated human-machine hybrid scenarios where machine modifications may be minimal and localized.

To address these challenges, we propose the Multi-level Style Preference Optimization (MSPO) framework. MSPO captures machine style features at three distinct granularities: **sequence-level** (evaluating overall text coherence and style consistency), **phrase-level** (identifying distinctive n-gram patterns), and **lexical-level** (capturing subtle word choice distributions). We jointly optimize these levels during training through a mixed loss mechanism. Furthermore, to enhance adaptability, we introduce a complexity-based dynamic parameter adjustment mechanism that modulates optimization strength based on the detected complexity differences between human and machine text.

In summary, our main contributions are as follows:

- We propose the Multi-level Style Preference Optimization framework (MSPO), which models style differences from sequence, lexical, and phrase levels, significantly improving detection capability across texts with varying degrees of machine involvement.
- We introduce a complexity-based dynamic parameter adjustment mechanism that adaptively adjusts detection parameters based on text complexity differences, demonstrating enhanced robustness across diverse text types.
- We demonstrate superior detection performance across multiple domains and LLMs. On the most challenging rewritten texts, MSPO achieves up to 82.14% AUROC, representing an improvement of 11.15 percentage points over the strongest baseline ImBD. On polished texts in news domain, MSPO achieves 88.09% average AUROC with 6.01 percentage points improvement, showing particular strength in human-machine hybrid scenarios where existing methods struggle.

Related Work

Purely Machine-Generated Text Detection

Detection methods for purely machine-generated text can be divided into two categories: training-based detection methods and zero-shot detection methods. **Training-based detection methods** rely on large-scale annotated datasets to train specialized models for identifying machine-generated text features. These methods typically employ binary classification architectures that learn to distinguish between human and machine text by capturing statistical patterns, linguistic structures, and stylistic differences in the training data. The GPT-2 detector (Solaiman et al. 2019) was among the first to demonstrate the feasibility of large-scale training-based detection approaches. Fine-tuned RoBERTa-based detectors (Liu et al. 2019) further improved performance by leveraging pre-trained language representations. Commercial solutions like GPTZero (Tian and Cui 2023) have made these approaches accessible to broader audiences, though they face challenges with model generalization across different generation architectures.

Zero-shot detection methods directly utilize the inherent properties of pre-trained language models without requiring task-specific training or domain-specific labeled data. Building on two key statistical properties of machine-generated text (which we formally define in Section Preliminaries), these methods can be divided into two categories:

Probability distribution-based methods: These methods utilize the characteristic that machine text exhibits higher log-likelihood values and lower entropy than human text. Methods like GLTR (Gehrmann, Strobelt, and Rush 2019) identify text by analyzing entropy distribution characteristics, while DNA-GPT (Yang et al. 2023) adopts a "truncate and regenerate" approach, regenerating parts of text and analyzing n-gram distribution differences. These methods distinguish text sources by modeling the more predictable output characteristics in language model generation processes.

Probability curvature-based methods: These methods utilize the characteristic that machine text typically lies in negative curvature regions in log-probability space. Representative methods include DetectGPT (Mitchell et al. 2023), Fast-DetectGPT (Bao et al. 2024), and NPR (Su et al. 2023), which identify machine text by perturbing text and analyzing probability changes. Since machine text often occupies local maxima of probability surfaces, perturbed samples generally exhibit lower probabilities than the original, enabling unsupervised detection through this characteristic probability gap.

Human-Machine Hybrid Text Detection

Widespread LLM application has shifted detection research toward human-machine hybrid text. However, research in this area is still in its early stages with relatively limited work.

Imitate Before Detect (ImBD) (Chen et al. 2025) establishes a scoring model by imitating the distributional features of human-machine hybrid texts and measuring their deviation from human text distributions for detection. This probability deviation-based approach avoids dependence on specific generation models or domains and demonstrates good generalizability across different scenarios. CheckGPT (Liu et al. 2024) proposes a deep learning framework to detect ChatGPT-generated content in academic writing. The framework captures subtle semantic and linguistic patterns in ChatGPT-generated academic abstracts, demonstrating improved performance over existing detection tools and hand-crafted features while exhibiting better cross-discipline transferability and robustness against prompt engineering.

Preliminaries

We first introduce two key statistical properties that characterize machine-generated text, which also serve as the theoretical foundation for current mainstream detection methods. Subsequently, we introduce an efficient preference alignment technique—Direct Preference Optimization (DPO) (Rafailov et al. 2023), which provides theoretical support for our proposed style preference optimization method.

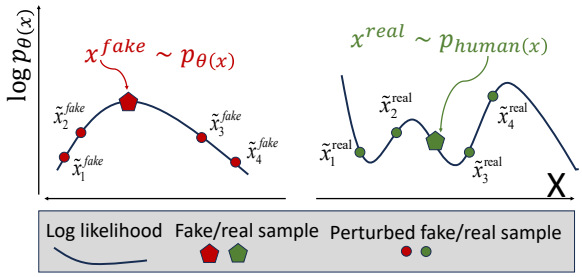

Figure 1: DetectGPT's illustration of probability curvature differences: machine-generated text (left) lies in negative curvature regions while human-written text (right) does not exhibit clear negative curvature (Mitchell et al. 2023).

Key Statistical Properties of Machine-Generated Text

Machine-generated text exhibits two key statistical differences from human-written text. First, LLM-generated text typically exhibits higher log-likelihood values and lower entropy due to LLMs' training objective to minimize negative log-likelihood (Lavergne, Urvoy, and Yvon 2008; Gehrmann, Strobelt, and Rush 2019; Su et al. 2023; Hashimoto, Zhang, and Liang 2019). Second, as shown in Figure 1, machine text often lies in negative curvature regions of log-probability space, where perturbations consistently yield lower probabilities (Mitchell et al. 2023).

Direct Preference Optimization

Direct Preference Optimization (DPO) is an efficient technique aimed at aligning language models with human preferences (Rafailov et al. 2023). Unlike traditional reinforcement learning-based methods (such as RLHF (Christiano et al. 2017; Bai et al. 2022)), DPO does not require explicit training of a reward model, but directly utilizes preference pair data to optimize language models. This approach significantly reduces computational complexity while maintaining or improving text generation quality.

The core idea of DPO is to directly transform human preference data into optimization objectives by maximizing the probability difference between preferred and rejected responses. Specifically, for each text pair (y_w, y_l), where y_w is the human-preferred response and y_l is the relatively less preferred response, DPO optimizes model parameters through the following formula:

$$\mathcal{L}_{\text{DPO}}(\theta;\theta_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Big(\beta\log\frac{p_\theta(y_w|x)}{p_{\theta_{\text{ref}}}(y_w|x)} - \beta\log\frac{p_\theta(y_l|x)}{p_{\theta_{\text{ref}}}(y_l|x)}\Big)\Bigg] \tag{1}$$

where p_θ is the model to be optimized, $p_{\theta_{\text{ref}}}$ is the reference model (usually the initial model), and β is a parameter controlling reward strength. This formula is based on the Bradley-Terry model, linking relative preference probabilities with differences in model scores.

In our work, DPO provides the theoretical foundation for our method. Although DPO's original intention is to align models with human preferences, we adapt this framework to make our scoring model "prefer" machine-generated text features. This optimization objective enables our model to learn unique patterns of machine text and assign higher scores to machine-generated content, thus establishing a solid foundation for effective detection of human-machine hybrid texts.

Methodology

We propose the Multi-level Style Preference Optimization (MSPO) framework. Unlike traditional binary classification methods, we optimize scoring differences between human and machine text through preference learning that captures multi-level style features.

As shown in Figure 2, our framework consists of two core components: multi-level style feature extraction and dynamic weight optimization. Through preference optimization training, these components work together to produce a machine-style scoring model for detection.

Problem Formulation

Given a text $x = (x_1, x_2, \ldots, x_n)$, our task is to determine whether it is human-written x_h or machine-generated x_m. We formulate this as a preference learning problem, training a scoring model $r_\theta(x)$ that quantifies machine-generation likelihood. The model is optimized to satisfy:

$$r_\theta(x_m) > r_\theta(x_h) \tag{2}$$

where higher scores indicate stronger machine-style characteristics. This preference-based formulation directly optimizes the scoring differences between human-written and machine-generated text. Our optimization objective is:

$$\max_\theta \mathbb{E}_{(x_m,x_h)\sim D}\left[r_\theta(x_m) - r_\theta(x_h)\right] \tag{3}$$

which maximizes the expected scoring gap, ensuring machine-generated text receives consistently higher scores than human-written text.

Multi-level Style Preference Optimization

Building upon the Direct Preference Optimization (DPO) framework (Rafailov et al. 2023), we propose Multi-level Style Preference Optimization (MSPO). While sequence-level optimization captures overall text probability, it may overlook fine-grained stylistic patterns at local textual units. To address this limitation, we introduce phrase-level and lexical-level constraints that capture stylistic differences across three complementary granularities: sequence, phrase, and lexical. The overall objective function is formulated as:

$$L_{\text{MSPO}} = L_{\text{seq}} + \gamma L_{\text{phrase}} + \alpha L_{\text{lexical}} \tag{4}$$

where L_{seq}, L_{phrase}, and L_{lexical} capture sequence-level, phrase-level, and lexical-level stylistic features respectively, with γ and α controlling their relative contributions.

Sequence-level Style Optimization Following DPO, we construct the scoring function $r_\theta(x)$ based on the language model $p_\theta(x)$:

$$r_\theta(x) = \beta\log\frac{p_\theta(x)}{p_{\text{ref}}(x)} \tag{5}$$

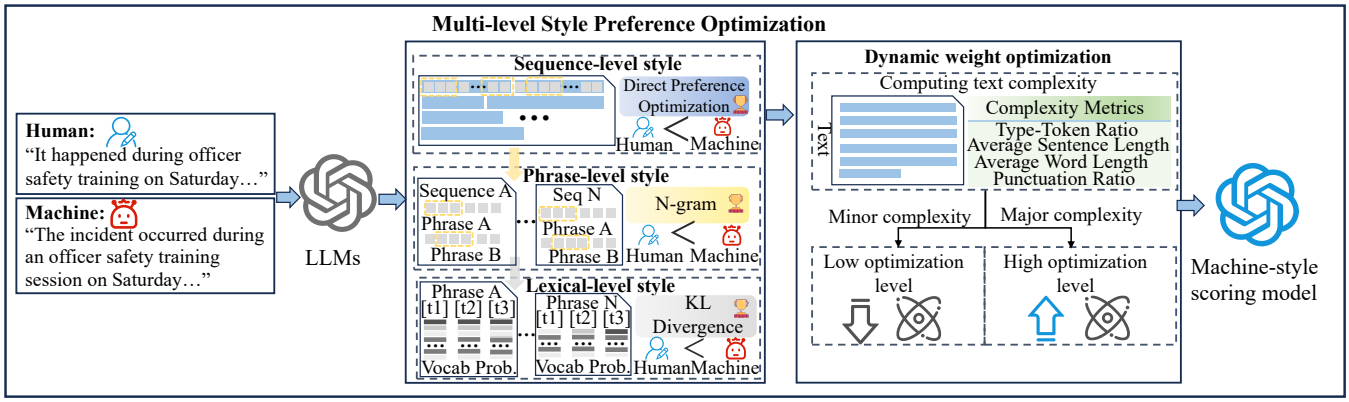

Figure 2: Multi-level Style Preference Optimization (MSPO) framework.

where $p_\theta(x)$ is the model to be optimized, $p_{\text{ref}}(x)$ is the reference model, and β controls the reward amplitude.

Using the Bradley-Terry model (Huang et al. 2006), we convert scoring differences into preference probabilities, yielding the sequence-level loss:

$$L_{\text{seq}} = -\log \sigma \left(\beta \log \frac{p_\theta(x_m)}{p_{\text{ref}}(x_m)} - \beta \log \frac{p_\theta(x_h)}{p_{\text{ref}}(x_h)} \right) \quad (6)$$

where (x_m, x_h) denotes a machine-human text pair, σ is the sigmoid function, and β controls the strength of the preference constraint.

Phrase-level Style Optimization The phrase-level constraint focuses on n-gram combination patterns, capturing consistent differences in local phrase structures:

$$L_{\text{phrase}} = \frac{1}{K} \sum_{i=1}^{K} \max(0, \log p_{\text{ref}}(g_i) - \log p_\theta(g_i) + \lambda) \quad (7)$$

where g_i represents the i-th n-gram sequence extracted from the text, $p_\theta(g_i)$ and $p_{\text{ref}}(g_i)$ denote the joint probability of this n-gram under the optimized model and reference model respectively, λ is a margin parameter, and K is the total number of n-grams. This hinge loss encourages the optimized model to assign higher probabilities to n-gram patterns when processing machine-generated text compared to the reference model, thereby amplifying stylistic differences at the phrase level.

Lexical-level Style Constraint The lexical-level constraint measures word selection distributional differences through KL divergence at each token position:

$$L_{\text{lexical}} = \frac{1}{n} \sum_{i=1}^{n} D_{KL}(p_\theta(\cdot|x_{h,<i}) \| p_\theta(\cdot|x_{m,<i})) \quad (8)$$

where $p_\theta(\cdot|x_{h,<i})$ and $p_\theta(\cdot|x_{m,<i})$ represent the vocabulary probability distributions at position i under the optimized model p_θ, conditioned on the preceding context of human text (x_h) and machine text (x_m) respectively. This term prevents the model from overfitting to context-specific styles, ensuring that p_θ maintains a generalized output distribution. This complements L_{seq} and L_{phrase}, which focus on learning discriminative features.

Dynamic Weight Optimization

In the DPO framework, the parameter β controls the extent of deviation of the optimized model from the reference model. Traditional methods use a static β value, which fails to account for varying complexity differences between human-written and machine-generated texts across different samples. To address this limitation, we introduce a dynamic β adjustment mechanism that adapts to text complexity variations.

We measure text complexity using four metrics: Type-Token Ratio (TTR), Average Sentence Length (ASL), Average Word Length (AWL), and Punctuation Ratio (PR). The complexity score is computed as:

$$C(x) = w_1 \text{TTR}(x) + w_2 \text{ASL}(x) + w_3 \text{AWL}(x) + w_4 \text{PR}(x) \quad (9)$$

where w_1, w_2, w_3, w_4 are the weights for each indicator. Based on the complexity difference between human and machine text pairs, we dynamically adjust β using:

$$\beta_{\text{dynamic}} = \beta_{\text{base}} \left(1 + \delta \tanh \left(\kappa \left(\frac{C(x_h) - C(x_m)}{C(x_m)} \right) \right) \right) \quad (10)$$

where β_{base} is the baseline value, δ controls the maximum adjustment range, and κ controls the adjustment sensitivity. This dynamic mechanism enables the model to adapt the optimization pressure based on the relative complexity difference in each pair, thereby improving detection performance across diverse text samples.

Style Probability Curvature Detection

The overall detection pipeline is illustrated in Figure 3. After multi-level style preference optimization training, we obtain a machine-style scoring model p_θ for detection. We employ the D-score metric based on probability curvature (Bao et al. 2024).

For text $x = (x_1, \ldots, x_n)$, the D-score measures the deviation of observed log-likelihood from expected log-likelihood:

$$d(x, p_\theta) = \frac{\log p_\theta(x) - \mu}{\sigma} \quad (11)$$

where $\log p_\theta(x) = \sum_{i=1}^{n} \log p_\theta(x_i|x_{<i})$. The expected

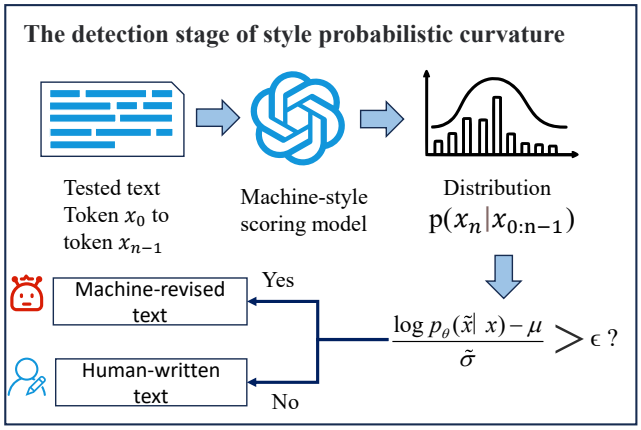

Figure 3: Detection pipeline with style probability curvature and threshold-based classification.

mean and standard deviation are:

$$\mu = \sum_{i=1}^{n} \mathbb{E}_{v\sim p_\theta(\cdot|x_{<i})}[\log p_\theta(v|x_{<i})]$$
$$\sigma^2 = \sum_{i=1}^{n} \mathrm{Var}_{v\sim p_\theta(\cdot|x_{<i})}[\log p_\theta(v|x_{<i})] \quad (12)$$

Since p_θ is optimized to assign higher probabilities to machine-generated patterns, machine text tends to have log-likelihoods above the model's expectation (positive D-scores), while human text exhibits lower log-likelihoods (negative or small D-scores). Classification is performed as:

$$f(x) = \begin{cases} 1, & \text{if } d(x, p_\theta) > \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where $f(x) = 1$ indicates machine-generated text.

Experiments

In this work, we conducted comprehensive experiments to validate the effectiveness of our proposed MSPO method. We designed our experiments to address the following research questions:

- **RQ1: Can MSPO outperform state-of-the-art baseline methods?**
- **RQ2: Can MSPO generalize across different LLMs and machine text types?**
- **RQ3: Which components are crucial for MSPO's performance?**

Dataset Construction

We construct a comprehensive dataset spanning three domains: scientific research, news, and creative writing. Each domain contains both human-written texts and three types of machine-generated variants.

Data Sources The human-written texts are sourced from established benchmarks published before 2019 (predating GPT-3 (Brown et al. 2020)) to ensure authenticity. For the scientific domain, we use the GPABench2 Dataset (Liu et al. 2024), which includes computer science (CS), health/social science (HSS), and physics (PHX) abstracts. The news domain is based on the XSum dataset (Narayan, Cohen, and Lapata 2018), while the creative writing domain uses the WritingPrompts dataset (Fan, Lewis, and Dauphin 2018). From each subdomain, we extract 1,000 human-written samples for our experiments.

Machine-Generated Text Variants We generate three types of machine-altered texts using five state-of-the-art language models: GPT-3.5(OpenAI 2022), GPT-4o(OpenAI 2024), LLaMA3.1-8B(Dubey et al. 2024), Qwen3(Yang et al. 2025), and DeepSeek-R1-Distill-Llama-8B(Guo et al. 2025).

Pure Machine-generated Text. We use domain-specific prompts based on original titles, such as "As an expert researcher, write... abstract... [title]" (scientific), "Write a concise... news summary... [title]" (news), and "As an expert English writer, write... paragraph... [title]" (creative). We match generation length to the corresponding human-authored samples in each domain.

Machine-rewritten Text. We first split the original text into sentences, then randomly select 50% of sentences for rewriting using language models. The prompt template is: "Rewrite the following sentence in your own words, maintaining the same meaning and accuracy, but using different phrasing and vocabulary". The unselected sentences remain unchanged, and all sentences are reassembled.

Machine-polished Text. We split the original text into individual sentences, then systematically polish each sentence using language models with the prompt: "Polish and enhance the following sentence to improve clarity, flow, and overall quality while preserving its core meaning." After polishing all sentences, we reassemble them to obtain the complete machine-polished version.

Statistics of the Dataset We collected 72,500 samples, consisting of 5,000 human texts from existing datasets and 67,500 machine variants across 5 models and 3 types (ranging from 500 to 1,000 samples each), as detailed in Table 1.

Category	Details	Count
Human	(From existing datasets)	5,000
Machine	(Total)	67,500
	5 models, 5 source datasets, 3 types	
	500–1,000 samples per combination	
Total		**72,500**

Table 1: Statistics of the Dataset

Experimental Setup

Model Configuration We fine-tune GPT-NEO-2.7B via LoRA ($r = 16, \alpha = 32$) on 1,000 GPT-3.5 polished pairs (200/subdomain). Training was conducted for 5 epochs

on a single A600 GPU using AdamW (LR 1×10^{-4}, batch 8, gradient accumulation). All results are reported from a single run. MSPO hyperparameters are $\beta_{\text{base}} = 0.08, \delta = 0.25, \kappa = 1.5$; complexity weights $\mathbf{w} = \{0.35, 0.30, 0.20, 0.15\}$ for TTR, ASL, AWL, and PR, respectively; and other parameters are $n = 3, \lambda = 0.12, \gamma = 0.25, \alpha = 0.35$.

Baselines and Evaluation Metrics

We compare MSPO against 8 baseline methods: Likelihood (Solaiman et al. 2019) computing average log-probability of tokens; Rank and LogRank (Gehrmann, Strobelt, and Rush 2019) analyzing token rank distributions in the vocabulary; Entropy (Gehrmann, Strobelt, and Rush 2019) measuring token sequence predictability; LRR (Su et al. 2023) using likelihood ratio ranking between models; Fast-DetectGPT (Bao et al. 2024) measuring probability curvature through text perturbation; DNA-GPT (Yang et al. 2023) comparing n-gram distributions between original and regenerated texts; and ImBD (Chen et al. 2025) learning to imitate machine text distributions and measuring deviation from human text.

We evaluate all methods using the Area Under the Receiver Operating Characteristic curve (AUROC) as the primary metric.

Main Results

Method	Qwen3	Llama3.1	DeepSeek	GPT-4o	Avg.
Likelihood	0.2033	0.2155	0.1799	0.2021	0.2002
Rank	0.3810	0.4121	0.3134	0.3564	0.3657
LogRank	0.8050	0.7836	0.8189	0.8102	0.8044
Entropy	0.6769	0.6525	0.7082	0.6802	0.6795
LRR	0.2241	0.2703	0.2368	0.2090	0.2351
Fast-DetectGPT	0.2625	0.2913	0.2553	0.2599	0.2673
ImBD	0.8247	0.8797	0.7441	0.8348	0.8208
MSPO(Ours)	**0.8849**	**0.8999**	**0.8428**	**0.8960**	**0.8809**

Table 2: Detection performance (AUROC) on machine-polished texts from the news domain (XSum) across four language models (Qwen3, LLaMA3.1, DeepSeek, GPT-4o).

Comparison with State-of-the-Art Methods (RQ1) Polished texts represent a commonly encountered scenario in real-world applications, such as academic writing assistance and content optimization tools. We focus on the polish task for core method validation, as it presents notable detection challenges where machine modifications are subtle and localized.

We evaluate on news domain texts (XSum dataset). Table 2 demonstrates the detection performance comparison between MSPO and all baseline methods. MSPO achieves an average AUROC of 88.09% across four language models, outperforming the strongest baseline ImBD (82.08%) with an improvement of 6.01 percentage points. Other baselines show poor performance with AUROC scores ranging from 20% to 68%, indicating their ineffectiveness on this challenging task.

Cross-Model and Cross-Domain Evaluation (RQ2) We evaluate MSPO's performance across three task types with increasing complexity: pure generation, polishing, and rewriting. We use GPT-3.5 as the representative model and conduct tests across three academic domains (CS, HSS, PHX) to ensure a robust evaluation while controlling for model-specific variations.

Table 3 reveals MSPO's superior adaptability across complexity levels. On generated texts, MSPO achieves performance comparable to ImBD, with AUROC scores around 99%, demonstrating strong baseline performance. For polished texts, MSPO consistently outperforms all baselines across domains, maintaining high performance above 96% while other methods show significant degradation. On the most challenging rewritten texts, MSPO demonstrates remarkable robustness, achieving up to 82.14% AUROC in the PHX domain, compared to ImBD's 70.99%. This represents an absolute improvement of over 11 percentage points.

To validate MSPO's generalization beyond academic domains, Table 4 presents detection results on machine-rewritten texts in news summarization (XSum) and creative writing domains across five language models. The results demonstrate MSPO's strong generalization to these informal domains. Traditional statistical methods and other neural approaches show substantially lower performance across all models and domains. On news summarization (XSum), MSPO consistently outperforms all baselines across models. The most dramatic improvement is seen on LLaMA3.1, where MSPO achieves 77.15% AUROC, compared to the strongest baseline (ImBD) at only 59.32%, representing an improvement of 17.83 percentage points. For creative writing, MSPO maintains superior performance, reaching up to 73.29% AUROC on LLaMA3.1, while ImBD achieves 70.82%.

These results demonstrate MSPO's consistent superiority across complexity levels and domains, with performance gaps widening in more challenging scenarios.

Discussion

Ablation Study (RQ3) To understand each component's contribution in our MSPO framework, we conduct comprehensive ablation studies on CS domain polished text detection. We systematically remove key components to isolate their individual contributions:

We evaluate five variations. The **Full MSPO** includes all components: sequence, lexical, and phrase-level losses, plus the dynamic β. We then create three ablations: **W/O Lexical**, which removes the token-level KL divergence constraints but keeps sequence and phrase levels with dynamic β; **W/O Phrase**, which removes the n-gram-level structural constraints, keeping sequence and lexical levels with dynamic β; and **Fixed Beta**, which disables the complexity-based dynamic β adjustment, using a fixed β across all text types. Finally, we compare against a **Sequence Only** minimal baseline, which uses only the sequence-level MSPO loss and a fixed β.

Figure 4 presents comprehensive ablation results across four evaluation metrics. The results reveal clear component importance and validate our design choices. The ab-

Method	Generated			Rewritten			Polished		
	CS	HSS	PHX	CS	HSS	PHX	CS	HSS	PHX
Likelihood	0.9154	0.7503	0.7712	0.3380	0.4016	0.3278	0.3359	0.4297	0.3750
Rank	0.6586	0.7691	0.6080	0.4201	0.4404	0.3608	0.4214	0.5119	0.3714
LogRank	0.7320	0.2308	0.2291	0.6726	0.6027	0.6844	0.6850	0.5856	0.6521
Entropy	0.4777	0.5448	0.5512	0.6042	0.5605	0.5920	0.6151	0.5581	0.5550
LRR	0.9140	0.8184	0.7252	0.3245	0.3865	0.3071	0.2801	0.3682	0.2808
Fast-DetectGPT	0.9936	0.9800	0.9678	0.3535	0.3478	0.3131	0.3648	0.4500	0.3550
DNA-GPT	0.4328	0.6147	0.6892	0.3649	0.4410	0.4021	0.3818	0.4786	0.4480
ImBD	**0.9997**	**0.9999**	**0.9983**	0.7159	0.6590	0.7099	0.9548	0.9371	0.9339
MSPO(Ours)	0.9907	0.9812	0.9934	**0.8047**	**0.7703**	**0.8214**	**0.9774**	**0.9656**	**0.9727**

Table 3: Detection performance (AUROC) of GPT-3.5 across three task types (generated, rewritten, polished) in three scientific domains: computer science (CS), health and social science (HSS), and physics (PHX).

Method	Qwen3		LLaMA3.1		DeepSeek		GPT-3.5		GPT-4o	
	XSum	Writing	XSum	Writing	XSum	Writing	XSum	Writing	XSum	Writing
Likelihood	0.1746	0.3133	0.2644	0.3245	0.3011	0.3289	0.3380	0.2988	0.2904	0.3293
Rank	0.3527	0.5056	0.4310	0.5261	0.4289	0.5023	0.4201	0.4811	0.3956	0.5042
LogRank	0.1693	0.3291	0.2578	0.3303	0.2964	0.3427	0.3274	0.3112	0.2804	0.3434
Entropy	0.6990	0.6521	0.6557	0.6657	0.6296	0.6482	0.6042	0.6981	0.6355	0.6568
LRR	0.1890	0.3956	0.2758	0.3716	0.3185	0.4009	0.3245	0.3745	0.2820	0.4076
Fast-DetectGPT	0.1439	0.3461	0.2734	0.3928	0.3036	0.3787	0.3397	0.4122	0.2825	0.3996
ImBD	0.5683	0.6012	0.5932	0.7082	0.5913	0.6391	0.5723	0.6206	0.5571	0.6216
MSPO	**0.7556**	**0.6905**	**0.7715**	**0.7329**	**0.7431**	**0.7186**	**0.7665**	**0.7625**	**0.7641**	**0.7094**

Table 4: Detection performance (AUROC) on machine-rewritten texts in news (XSum) and creative writing (WritingPrompts) domains across five language models (Qwen3, LLaMA3.1, DeepSeek, GPT-3.5, GPT-4o).

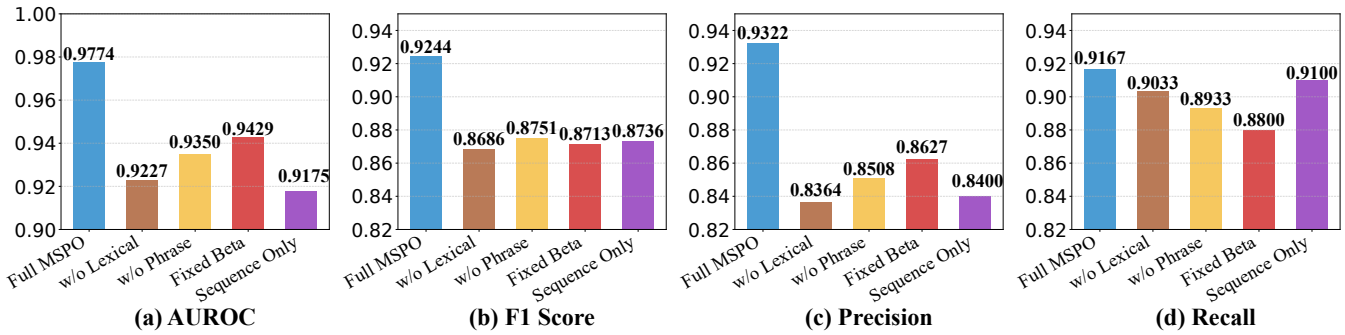

Figure 4: Comprehensive Ablation Study Results. Performance comparison across AUROC, F1 Score, Precision, and Recall metrics on CS domain polished text detection task. Each variant systematically removes specific components to isolate their contributions.

lation study demonstrates that the multi-level architecture provides the most significant contribution to MSPO's performance. Removing all multi-level enhancements ("Sequence Only") causes the largest performance drop of 5.99 percentage points in AUROC, highlighting the substantial value of our multi-level design. Among individual components, lexical-level modeling contributes more significantly (drop of 5.47 percentage points when removed) than phrase-level modeling (drop of 4.24 percentage points), while dynamic parameter adjustment provides additional optimization benefits (improvement of 3.45 percentage points over fixed parameters).

Conclusion

This work presents the Multi-level Style Preference Optimization (MSPO) framework for detecting human-machine hybrid texts using multi-level style modeling, dynamic parameter adjustment, and preference optimization. Comprehensive experiments across three domains and five LLMs demonstrate MSPO's superior performance, achieving 97.02% average AUROC on challenging polished text detection tasks. Ablation studies confirm each component's meaningful contribution, establishing MSPO as an effective solution for real-world machine text detection scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (U22B2036, 62572403, and 62202381), in part by Guangdong Basic and Applied Basic Research Foundation (2024A1515010087), and General Program of Chongqing Natural Science Foundation (No. CSTB2022NSCQ-MSX1284). We would like to thank the anonymous reviewers for their constructive comments.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bao, G.; Zhao, Y.; Teng, Z.; Yang, L.; and Zhang, Y. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chawla, D. S. 2024. Is ChatGPT corrupting peer review? Telltale words hint at AI use. *Nature*, 628: 483–484.

Chen, J.; Zhu, X.; Liu, T.; Chen, Y.; Xinhui, C.; Yuan, Y.; Leong, C. T.; Li, Z.; Tang, L.; Zhang, L.; et al. 2025. Imitate Before Detect: Aligning Machine Stylistic Preference for Machine-Revised Text Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23559–23567.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.*, 24(1): 11324–11436.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898.

Gehrmann, S.; Strobelt, H.; and Rush, A. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116. Florence, Italy: Association for Computational Linguistics.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of NAACL-HLT*, 1689–1701.

He, Z.; Huang, J.; Lu, M.; Huang, Z.; Liu, S.; Tian, Z.; and Li, D. 2025. GCML: Gradient Coherence Guided Meta-Learning for Cross-Domain Emerging Topic Rumor Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 19159–19173.

Huang, T.-K.; Weng, R. C.; Lin, C.-J.; and Ridgeway, G. 2006. Generalized Bradley-Terry Models and Multi-Class Probability Estimates. *Journal of Machine Learning Research*, 7(1).

Lavergne, T.; Urvoy, T.; and Yvon, F. 2008. Detecting Fake Content with Relative Entropy Scoring. *Pan*, 8(27-31): 4.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Liu, Z.; Yao, Z.; Li, F.; and Luo, B. 2024. On the detectability of ChatGPT content: Benchmarking, methodology, and evaluation through the lens of academic writing. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2236–2250.

M Alshater, M. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. *Available at SSRN 4312358*.

Mai, G.; Huang, W.; Sun, J.; Song, S.; Mishra, D.; Liu, N.; Gao, S.; Liu, T.; Cong, G.; Hu, Y.; et al. 2024. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 10(2): 1–46.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*, 24950–24962. PMLR.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/. Accessed: 2022-11-30.

OpenAI. 2023a. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt. Accessed: 2023-12-01.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2024. GPT-4o: Omni-modal AI for Everyone. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-05-13.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2023. Can AI-Generated Text be Reliably Detected? arXiv:2303.11156.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; McCain, M.; Newhouse, A.; Blazakis, J.; McGuffie, K.; and Wang, J. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203.

Su, J.; Zhuo, T. Y.; Wang, D.; and Nakov, P. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. arXiv:2306.05540.

Tian, E.; and Cui, A. 2023. GPTZero: Towards Detection of AI-Generated Text Using Zero-Shot and Supervised Methods. Technical report, GPTZero. https://gptzero.me.

Tian, Z.; Huang, J.; He, Z.; Huang, Z.; Lu, M.; Qiao, L.; Mei, S.; Wang, Y.; and Li, D. 2025. LLM-based rumor detection via influence guided sample selection and game-based perspective analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 28402–28414.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wu, L.; Liu, P.; Zhao, Y.; Wang, P.; and Zhang, Y. 2023a. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 211–225.

Wu, L.; Long, Y.; Gao, C.; Wang, Z.; and Zhang, Y. 2023b. MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection. *Information Fusion*, 100: 101944.

Wu, L.; Wang, K.; Nie, K.; Guo, S.; Gao, C.; Wang, Z.; and Li, S. 2025. TFGIN: Tight-Fitting Graph Inference Network for Table-based Fact Verification. *ACM Transactions on Information Systems*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, X.; Cheng, W.; Wu, Y.; Petzold, L.; Wang, W. Y.; and Chen, H. 2023. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. arXiv:2305.17359.

Zhang, Q.; Gao, C.; Chen, D.; Huang, Y.; Huang, Y.; Sun, Z.; Zhang, S.; Li, W.; Fu, Z.; Wan, Y.; and Sun, L. 2024. LLM-as-a-Coauthor: Can Mixed Human-Written and Machine-Generated Text Be Detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, 409–436. Mexico City, Mexico: Association for Computational Linguistics.