Beyond Stylistic Suppression: Adaptive Reliance Modeling for Unseen-Generator AI-Text Detection

Anonymous ACL submission

Abstract

Detecting AI-generated text from unseen large language models (LLMs) remains highly challenging due to substantial stylistic variation across generators. Existing approaches typically improve robustness by suppressing stylistic information from either lexical or global perspectives, but they often treat stylistic behaviors in a homogeneous manner. However, different stylistic behaviors exhibit varying transferability under generator shift, making overly aggressive stylistic suppression suboptimal for robust generalization. To address this issue, we propose an adaptive stylistic modeling framework for unseen-generator AI-text detection. Our framework progressively reduces reliance on unstable lexical stylistic features while dynamically regulating global stylistic behavior to balance transferable semantic and stylistic evidence under diverse unseen-generator conditions. Experiments on the MAGE benchmark, covering 27 representative LLMs across 7 categories, demonstrate consistent improvements over state-of-the-art methods under varying unseen-generator conditions and generator shifts. Our source code will be publicly available at https://github.com/xxx.

1 Introduction

With the rapid evolution of LLMs, detecting AI-generated text from unseen generators has become increasingly challenging. Unlike closed-world settings that rely on familiar generator patterns, real-world deployments must generalize to continuously evolving LLMs with unknown architectures, training data, and decoding behaviors. These models can produce fluent and stylistically diverse text that closely resembles human writing, exposing the limitations of existing detection approaches.

Humans often rely on both lexical choices and broader stylistic behaviors when judging whether a text is machine-generated, as different generators may exhibit distinct word preferences and document-level writing tendencies. To examine this phenomenon, we collect a set of human-written stories from Aesop's Fables and prompt multiple LLMs to rewrite them while preserving the original semantics. We then analyze the resulting texts from both lexical and global stylistic perspectives. As illustrated in Figure 1, different generators exhibit distinguishable lexical preferences under identical semantics, while also showing clear variation in global stylistic statistics. These observations suggest that stylistic variation across generators emerges at multiple linguistic levels rather than from isolated lexical artifacts alone.

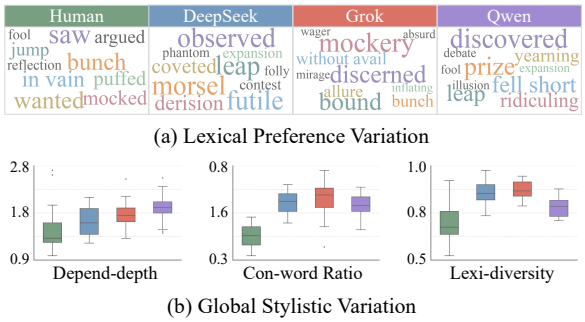

Figure 1: Visualization of lexical and global stylistic variation across humans and different LLMs using rewritten stories from Aesop's Fables. (a) Word clouds illustrate distinct lexical preferences under similar semantics. (b) Global stylistic statistics reveal differences in document-level writing patterns across generators.

Since stylistic patterns often vary significantly across generators, existing approaches improve unseen-generator detection by reducing reliance on generator-dependent stylistic cues, either from lexical-level behaviors such as token-level distributional patterns (Bao et al., 2024; Sun et al., 2025) and perturbation sensitivity (Mitchell et al., 2023; Liu et al., 2024; Zhu et al., 2025b; Li et al., 2025b), or from broader document-level statistical patterns (Su et al., 2023; Venkatraman et al., 2024; Hans et al., 2024) and linguistic structural repre-

sentations (Zanotto and Aroyehun, 2025; Li et al., 2025a). However, stylistic variations across generators often emerge jointly at lexical and global levels, where different stylistic behaviors exhibit varying degrees of transferability under generator shift. Uniform stylistic suppression may discard transferable authenticity cues, leading to unstable generalization. Thus, the key challenge is to jointly model multi-level stylistic behaviors while adaptively regulating the detector's reliance on them.

To address this issue, we propose an adaptive stylistic modeling framework that jointly models lexical stylistic preference and global stylistic behavior. At the lexical level, the framework identifies generator-sensitive lexical preferences and suppresses unstable stylistic reliance through neutral-aware semantic aggregation. At the global level, it introduces a dual-branch adaptive regulation mechanism that dynamically balances transferable semantic information and residual stylistic evidence under different unseen-generator conditions.

In summary, our contributions are as follows:

- We revisit unseen-generator AI-text detection from the perspective of stylistic reliance, showing that stylistic variation emerges at both lexical and global levels with varying transferability under generator shift.
- We propose an adaptive stylistic modeling framework that suppresses unstable generator-sensitive lexical reliance through neutral-aware semantic aggregation, while dynamically regulating global stylistic behavior via a dual-branch adaptive regulation mechanism.
- Extensive experiments across diverse LLM families demonstrate consistent improvements over state-of-the-art methods under both moderate and large generator shifts.

2 Related Work

Recent studies on AI-generated text detection suggest that discriminative signals can emerge from both lexical stylistic preferences and broader global writing behaviors (Liu et al., 2023; Mitchell et al., 2023; Wang et al., 2023). Accordingly, existing methods mainly improve robustness from either lexical or higher-level stylistic perspectives.

2.1 Lexical Stylistic Modeling

A large body of recent work exploits lexical-level stylistic cues for AI-generated text detection, including token distributions (Sun et al., 2025; Shi et al., 2024), token-level probability signals (Zhou et al., 2025a; Chen et al., 2025; Basani and Chen, 2025), and perturbation sensitivity (Liu et al., 2024; Li et al., 2025b; Zhu et al., 2025b). These studies suggest that different LLMs exhibit distinguishable lexical preferences that can serve as effective detection signals. Other methods further improve robustness by modeling lexical consistency patterns or perturbation-invariant behaviors under distribution shifts (Li et al., 2025b; Liu et al., 2024; Zhu et al., 2025b; Zhou et al., 2025a). While these approaches effectively exploit lexical stylistic signals, they often rely heavily on generator-sensitive word usage patterns, which may generalize poorly under unseen-generator shifts.

2.2 Global and Structural Stylistic Modeling

Another line of work improves robustness by modeling broader global or structural stylistic behaviors beyond token-level lexical patterns. Existing methods exploit contextual representations and perturbation-invariant embeddings to improve robustness under distribution shifts (Pu et al., 2026; Kuznetsov et al., 2024), while others further incorporate topic disentanglement, hierarchical representation learning, or contrastive objectives to capture transferable generation characteristics (Gui et al., 2025; He et al., 2025). Some studies additionally model higher-level structural signals such as entity graphs, syntactic regularities, or language-specific linguistic patterns to characterize generation behaviors beyond lexical statistics (Wu et al., 2025; Park et al., 2025; Su et al., 2025). However, these approaches often model higher-level stylistic information in a homogeneous manner, without explicitly regulating the detector's reliance on transferable versus generator-sensitive stylistic behaviors.

Overall, existing methods primarily model stylistic information from either lexical or higher-level global perspectives in isolation. In contrast, our work jointly models lexical stylistic preferences and global stylistic behaviors, while dynamically regulating their influence under generator shift.

3 Methodology

We propose an adaptive stylistic modeling framework for robust AI-generated text detection under unseen generators. Our framework is motivated by the observation that stylistic discrepancies across generators emerge from multiple comple-

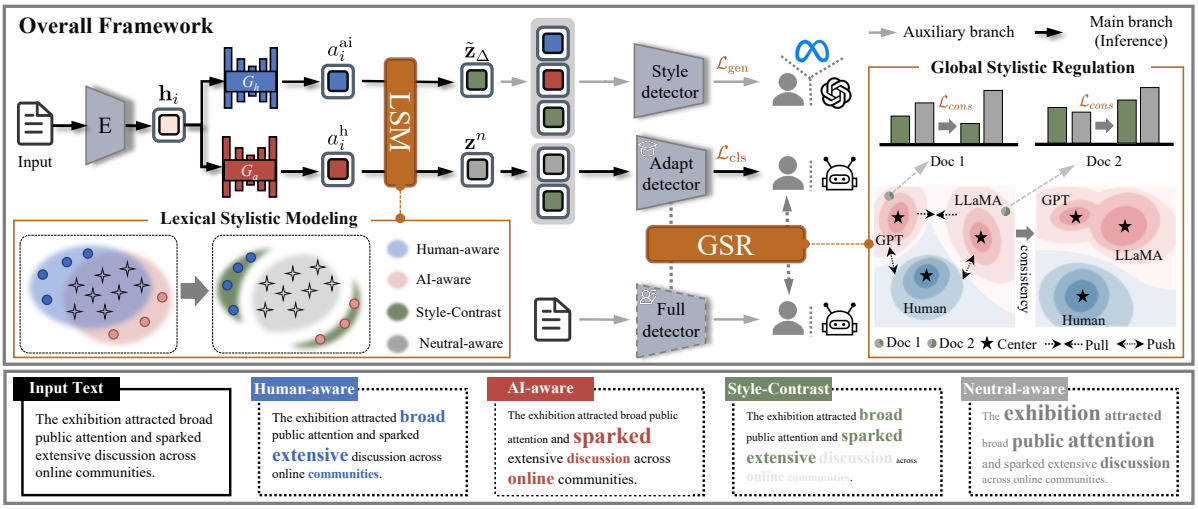

Figure 2: Overview of the proposed framework. The framework consists of two stages: lexical stylistic modeling and adaptive stylistic regulation. The first stage models token-level stylistic preferences to suppress unstable lexical stylistic reliance through neutral-aware semantic aggregation. The second stage introduces a dual-branch regulation mechanism to dynamically regulate global stylistic reliance under unseen-generator shifts.

mentary levels, including lexical preferences and global stylistic behaviors. Rather than uniformly suppressing stylistic information, we distinguish unstable, generator-sensitive patterns from transferable stylistic cues and adaptively regulate their influence across different generator shifts. This design reduces overreliance on generator-specific styles while preserving transferable stylistic evidence that is beneficial for generalization.

3.1 Problem Setup

We study AI-generated text detection under an unseen-generator setting. Each training instance is represented as (x_i, y_i, s_i), where x_i denotes a text sample, $y_i \in \{0, 1\}$ indicates whether the text is human-written or AI-generated, and s_i denotes the corresponding source generator (e.g., GPT or LLaMA). During training, samples are drawn from a set of seen generators, whereas at inference time the detector may encounter text generated by unseen generators with different stylistic characteristics. The goal is to learn a detector that remains robust under generator shift.

3.2 Lexical Stylistic Modeling

We begin by explicitly modeling token-level stylistic preference behaviors. Our motivation is that different generators often exhibit distinct lexical tendencies, with certain tokens more likely to align with human-written patterns while others tend to reflect generator-preferred stylistic behaviors. Instead of treating stylistic information as a single homogeneous signal, we characterize stylistic preference through competing AI-aware and human-aware directions.

3.2.1 Lexical Preference Estimation

Given an input text $x = \{t_1, \ldots, t_n\}$, a pre-trained BERT encoder first produces contextualized token representations $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\} \in \mathbb{R}^{n \times d}$. To characterize token-level stylistic tendencies, we compute two competing preference scores for each token:

$$a_i^{\text{ai}} = G_{\text{ai}}(\mathbf{h}_i), \qquad a_i^{\text{h}} = G_{\text{h}}(\mathbf{h}_i), \tag{1}$$

where $G_{\text{ai}}(\cdot)$ and $G_{\text{h}}(\cdot)$ denote learnable projections corresponding to AI-aware and human-aware stylistic directions, respectively. The two preference scores are jointly normalized via a softmax operation, encouraging each token to exhibit a relative preference for one stylistic direction.

Based on the normalized preference scores, token representations are aggregated into AI-aware and human-aware semantic representations through weighted aggregation:

$$\mathbf{z}^{\text{ai}} = \sum_{i=1}^{n} a_i^{\text{ai}} \mathbf{h}_i, \qquad \mathbf{z}^{\text{h}} = \sum_{i=1}^{n} a_i^{\text{h}} \mathbf{h}_i. \tag{2}$$

To further characterize the relative stylistic discrepancy between competing preference directions, we introduce a style-contrast representation:

$$\mathbf{z}_\Delta = \mathbf{z}^{\text{ai}} - \mathbf{z}^{\text{h}}. \tag{3}$$

3.2.2 Lexical Preference Suppression

Having identified token-level lexical preferences, we next suppress unstable preference dominance through a neutral-aware aggregation mechanism. Our intuition is that tokens with balanced AI-aware and human-aware activations are more likely to capture semantically stable content shared across generators, whereas strongly asymmetric activations tend to reflect generator-sensitive stylistic patterns.

Tokens with more balanced preference activations are assigned larger neutrality weights:

$$a_i^n = \exp\left(-\frac{(a_i^{\text{ai}} - a_i^{\text{h}})^2}{\tau}\right), \tag{4}$$

where τ controls the sensitivity to stylistic preference contrast.

Using these neutrality weights, the neutral-aware semantic representation is obtained through weighted aggregation of contextualized token representations:

$$\mathbf{z}^n = \sum_{i=1}^{n} a_i^n \mathbf{h}_i. \tag{5}$$

Using these neutrality weights, tokens with balanced stylistic preferences contribute more substantially to the final semantic representation, encouraging the model to focus on semantically stable content shared across generators. Importantly, the proposed mechanism does not discard stylistic information entirely, but instead suppresses excessive reliance on highly preference-sensitive lexical patterns.

3.2.3 Residual Stylistic Purification

While lexical preference suppression reduces unstable stylistic dominance, residual stylistic information may still provide complementary evidence beneficial for unseen-generator detection. To further reduce semantic interference in stylistic modeling, we purify the style-contrast representation through a lightweight orthogonalization operation:

$$\tilde{\mathbf{z}}_\triangle = \mathbf{z}_\triangle - \frac{\mathbf{z}_\triangle^\top \mathbf{z}^n}{\|\mathbf{z}^n\|_2^2}\mathbf{z}^n, \tag{6}$$

where $\tilde{\mathbf{z}}_\triangle$ denotes the purified style-contrast representation.

This operation removes the component aligned with the neutral semantic direction from the stylistic representation, encouraging stylistic cues to act as complementary evidence rather than reusing transferable semantic dimensions for discrimination.

To further organize stylistic structures across different generators, we introduce an auxiliary generator attribution objective. Specifically, the AI-aware, human-aware, and purified style-contrast representations jointly characterize stylistic behavior from complementary preference perspectives:

$$\mathcal{L}_{\text{gen}} = -\mathbb{E}\Big[\log C_{\text{gen}}^{(z_i)}([\mathbf{z}^{\text{ai}}; \mathbf{z}^{\text{h}}; \tilde{\mathbf{z}}_\triangle])\Big], \tag{7}$$

where $C_{\text{gen}}(\cdot)$ denotes the generator attribution classifier.

This objective encourages the learned preference representations to better capture generator-sensitive lexical stylistic patterns across different source domains.

3.3 Adaptive Stylistic Regulation

Although lexical stylistic modeling suppresses unstable token-level preference dominance, document-level representations may still retain strong generator-specific global stylistic behaviors. However, fully removing stylistic information is also undesirable, since some stylistic characteristics may remain transferable when unseen generators remain related to seen ones. Therefore, rather than enforcing complete stylistic invariance, the key objective is to adaptively regulate the detector's reliance on global stylistic information under different unseen-generator conditions.

To this end, we construct a dual-branch regulation framework consisting of an adaptive branch and a full-information branch. The adaptive branch performs AI/Human classification using the global neutral-aware semantic representation $\mathbf{z}^n$ together with the global style-contrast representation $\tilde{\mathbf{z}}_\triangle$, encouraging prediction behavior anchored in transferable semantic information with controlled stylistic evidence. In contrast, the full-information branch directly operates on the original document-level contextual representation produced by BERT, preserving richer stylistic information from the raw document distribution.

Different generators often exhibit substantially different global stylistic tendencies. Directly optimizing a single detector on such heterogeneous stylistic distributions can easily cause the prediction behavior to drift toward dominant generator-specific styles, resulting in unstable generalization under unseen generators. To alleviate this issue, we stabilize global stylistic behavior through the

interaction between the adaptive branch and the full-information branch.

Specifically, the full-information branch is not directly optimized by back-propagation. Instead, its parameters are updated through a temporally smoothed trajectory from the adaptive branch:

$$\theta_{\text{full}} \leftarrow m \times \theta_{\text{full}} + (1 - m) \times \theta_{\text{adapt}}, \quad (8)$$

where the momentum parameter m follows a cosine schedule:

$$m \triangleq 1 - \frac{(1 - m_{\text{base}}) \cdot (\cos(\pi k/K) + 1)}{2}, \quad (9)$$

where m_{base} denotes the initial momentum coefficient, while k and K represent the current and total training steps, respectively.

By continuously accumulating stylistic tendencies across different generators, the full-information branch captures a temporally smoothed global stylistic tendency rather than overfitting to dominant styles of individual generators. Meanwhile, because the full-information branch continues to observe unsuppressed document-level representations, transferable stylistic evidence can still be preserved when stylistic similarity remains informative for detecting unseen generators.

To transfer this stabilized stylistic behavior to the adaptive branch, we further align the prediction distributions of the two branches through a consistency objective:

$$\mathcal{L}_{\text{cons}} = T^2 \, \text{KL}\Big(\text{softmax}(p_{\text{full}}/T) \parallel \text{softmax}(p_{\text{adapt}}/T)\Big). \quad (10)$$

where p_{full} and p_{adapt} denote the prediction logits produced by the full-information branch and the adaptive branch, respectively, and T denotes the temperature parameter.

Unlike directly enforcing generator-invariant representations, the proposed mechanism dynamically regulates stylistic reliance according to different unseen-generator conditions. Consequently, the detector learns to suppress unstable generator-dependent global styles under large generator shifts while still preserving transferable stylistic evidence when stylistic similarity remains informative.

The overall training objective is formulated as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{gen}} + \beta \mathcal{L}_{\text{cons}}, \quad (11)$$

where α and β control the contributions of generator attribution supervision and consistency-based stylistic regulation, respectively.

	MAGE Corpus (Li et al., 2024)						
Cate.	FLAN.	GPT	LLa.	OPT	GLM	BSci.	Eleu.
AI	15K	10K	9K	12K	7K	8K	6.8K
Human	15K	10K	9K	12K	7K	8K	6.8K

Table 1: Number of AI-generated and human-written samples across datasets. Balanced sampling is applied to categories with multiple LLMs, ensuring each category contains a comparable number of samples, resulting in 27 representative models overall.

4 Experiments

4.1 Experimental Details

Datasets. As summarized in Table 1, we conduct experiments on the MAGE benchmark (Li et al., 2024), which contains texts generated by 27 LLMs across seven generator categories. Following the original MAGE setting, we use all provided LLMs and category splits. For each category, AI-generated and human-written samples are balanced. For In-Category evaluation, samples are randomly split into training, validation, and test sets in a 7:1:2 ratio. For Cross-Category evaluation, we follow the leave-one-generator-out setting by treating one category as unseen during training.

Competitive Methods. We compare our method with representative unseen-generator detectors from the perspectives of lexical perturbation and representation regularization. PECOLA (Liu et al., 2024) and DP-Net (Zhou et al., 2025b) improve robustness through perturbation-based strategies. SCRN (Huang et al., 2024), D-SVDD (Zeng et al., 2025), and DRGD (Pu et al., 2026) reduce generator-specific bias through semantic regularization, OOD modeling, and representation disentanglement. We further include MCP (Zhu et al., 2025a) and DNA (Zhu et al., 2025b) as representative zero-shot detectors, together with BERT (Devlin et al., 2019) as the encoder-only baseline.

Implementation Details. We use the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 16. The adaptive branch is updated through momentum-based parameter smoothing with $m_{\text{base}} = 0.996$, and the consistency temperature is set to $T = 3.0$. The loss weights are set to $\alpha = 0.9$ for $\mathcal{L}_{\text{gen}}$ and $\beta = 0.8$ for $\mathcal{L}_{\text{cons}}$. All experiments are conducted on two NVIDIA GeForce RTX 3090 GPUs.

	Accuracy (%) / F_1-Measure						
Cross-Category	FLAN-T5	GPT	LLaMA	OPT	GLM	BigScience	EleutherAI
Zero-shot methods							
$\text{MCP}_{\text{ACL'25}}$	60.2 / 77.5	73.2 / 84.1	76.6 / 81.2	78.9 / 78.6	75.7 / 75.4	85.3 / 85.2	83.1 / 83.5
$\text{DNA}_{\text{NIPS'25}}$	63.9 / 75.7	74.4 / 77.6	75.8 / 78.9	77.6 / 77.8	74.6 / 74.9	82.1 / 83.7	86.7 / 86.3
Training-based methods							
$\text{BERT}_{\text{NAACL'19}}$	75.9 / 76.4	78.8 / 78.1	84.1 / 81.4	87.4 / 86.8	85.9 / 84.5	88.6 / 87.9	88.5 / 88.0
$\text{PECOLA}_{\text{ACL'24}}$	76.1 / 75.9	81.4 / 80.7	84.7 / 84.5	87.6 /87.5	78.3 / 78.9	82.9 / 81.7	78.2 / 79.5
$\text{SCRN}_{\text{ACL'24}}$	75.8 / 75.7	80.9 / 80.8	81.7 / 81.3	87.4 / 87.3	72.8 / 71.6	81.8 / 80.6	89.3 / 89.2
$\text{DP-Net}_{\text{NAACL'25}}$	69.3 / 72.5	73.6 / 75.2	85.8 / 85.7	80.5 / 80.4	88.8 / 88.7	80.3 / 81.3	85.9 / 85.0
$\text{D-SVDD}_{\text{NIPS'25}}$	80.3 / 81.4	82.7 / 82.9	88.4 / 89.0	89.4 / 89.6	84.2 / 84.6	90.1 / 89.8	91.6 / 91.8
$\text{DRGD}_{\text{ACL'26}}$	77.0 / 76.8	81.5 / 81.6	86.8 / 87.8	90.1 / 89.4	88.4 / 89.2	90.6 / 90.7	91.0 / 91.6
Ours	**81.8 / 81.7**	**83.5 / 83.1**	**89.3 / 89.1**	**90.4 / 90.3**	**89.9 / 89.6**	**91.4 / 91.2**	**92.9 / 92.7**
In-Category	FLAN-T5-base	Davinci-002	LLaMA-30B	OPT-6.7B	N/A*	T0-11B	GPT-NeoX-20B
Zero-shot methods							
$\text{MCP}_{\text{ACL'25}}$	60.3 / 77.6	73.5 / 84.7	76.7 / 81.4	78.8 / 78.2	- / -	85.6 / 85.5	82.8 / 83.2
$\text{DNA}_{\text{NIPS'25}}$	63.4 / 75.2	75.2 / 76.3	75.7 / 78.5	75.6 / 76.3	- / -	82.2 / 83.8	87.3 / 86.9
Training-based methods							
$\text{BERT}_{\text{NAACL'19}}$	83.3 / 83.9	85.3 / 85.6	86.1 / 85.6	93.3 / 93.3	- / -	90.1 / 90.6	90.6 / 92.2
$\text{PECOLA}_{\text{ACL'24}}$	85.7 / 85.5	82.3 / 82.2	88.3 / 88.2	94.8 / 94.4	- / -	84.4 / 84.2	92.6 / 92.7
$\text{SCRN}_{\text{ACL'24}}$	81.1 / 80.5	84.4 / 84.2	86.1 / 85.9	93.1 / 93.0	- / -	83.9 / 83.5	90.2 / 90.3
$\text{DP-Net}_{\text{NAACL'25}}$	80.3 / 79.7	76.4 / 78.0	82.2 / 82.5	82.4 / 83.1	- / -	79.4 / 78.1	88.3 / 88.2
$\text{D-SVDD}_{\text{NIPS'25}}$	84.2 / 85.1	81.9 / 83.4	90.6 / 90.3	95.0 / 94.9	- / -	91.7 / 91.8	93.3 / 93.2
$\text{DRGD}_{\text{ACL'26}}$	86.2 / 86.0	86.8 / 87.6	90.3 / 89.9	95.0 / 95.1	- / -	91.0 / 91.1	93.6 / 93.5
Ours	**87.2 / 87.0**	**89.4 / 89.3**	**91.3 / 91.1**	**97.3 / 97.2**	- / -	**92.6 / 92.2**	**94.2 / 93.9**

Table 2: Cross-generator generalization results on the MAGE benchmark under leave-one-generator-out protocol. Cross-Category evaluates larger stylistic shifts across generator families, while In-Category evaluates stylistically related unseen generators. Our method consistently achieves the best overall performance across both settings. N/A*: Since GLM contains only a single subclass, in-category generalization is not applicable.

4.2 Overall Performance

4.2.1 Quantitative Generalization Results

We evaluate all methods under the leave-one-generator-out setting. As shown in Table 2, our method consistently achieves the best overall performance across both Cross-Category and In-Category settings. The gains are particularly notable under larger stylistic shifts. For example, on the challenging unseen FLAN-T5 category in Cross-Category evaluation, our method improves accuracy by 5.9% over BERT and 1.5% over D-SVDD. These results suggest that the proposed adaptive stylistic regulation effectively suppresses unstable generator-specific styles while preserving transferable stylistic evidence under varying unseen-generator conditions.

4.2.2 Representation Generalization Analysis

To further understand the generalization improvements observed in Table 2, we visualize representation distributions under different unseen-generator shifts using t-SNE visualizations in Figure 3. Vanilla BERT exhibits noticeable train/test distribution mismatch under unseen generators. Introducing lexical stylistic modeling ($+L$) improves AI/Human separability, while our full framework ($+LG$) further produces better-aligned train/test distributions across both In-Category and Cross-Category settings. These results demonstrate the complementary effects of lexical stylistic suppression and adaptive stylistic regulation under varying unseen-generator shifts.

4.3 Analysis of Stylistic Modeling

To better understand how lexical and global stylistic modeling affect robustness under unseen-generator shifts, we conduct a series of analyses from both representation and behavioral perspectives. Unless otherwise specified, we consistently compare three variants throughout this section: BERT as the encoder-only baseline, $+L$ with lexical stylistic modeling only, and $+LG$ (Ours) with additional adaptive stylistic regulation. We further include representative perturbation-based (DP-Net), distribution-based (D-SVDD), and zero-shot (DNA) detectors for comparison.

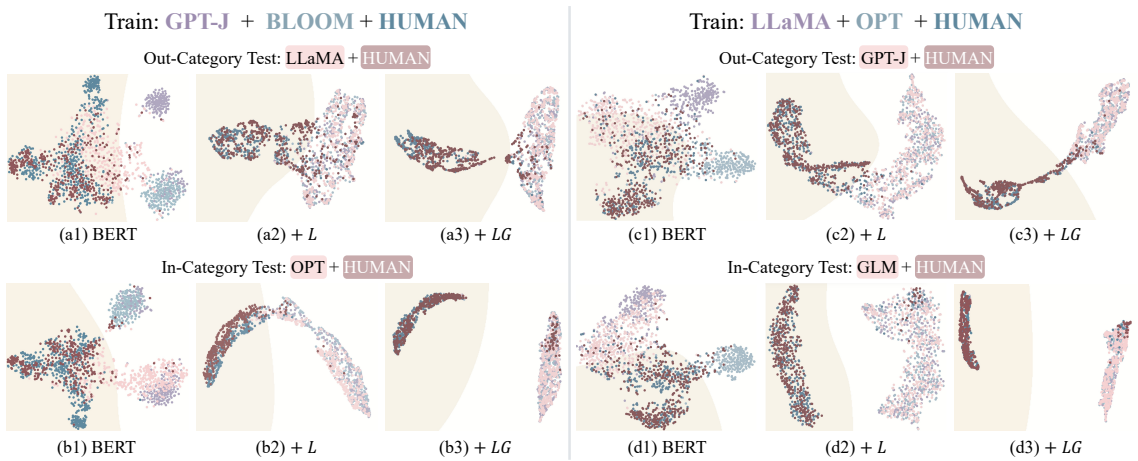

Figure 3: T-SNE visualizations under unseen generators with different stylistic relatedness. The top row corresponds to Cross-Category evaluation with larger generator shifts, while the bottom row shows In-Category evaluation with stylistically closer unseen generators. From left to right, BERT, lexical stylistic suppression ($+L$), and our full framework ($+LG$) progressively improve AI/Human separability and reduce train/test mismatch.

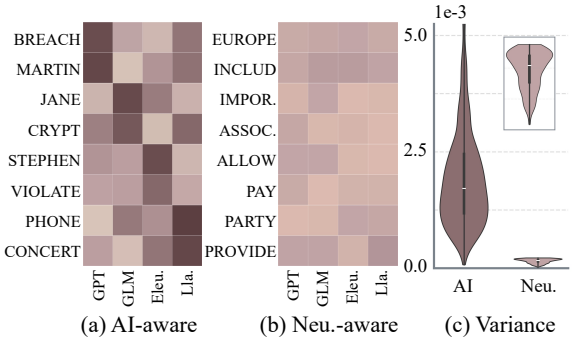

Figure 4: Qualitative and quantitative analysis of lexical preference across generators. (a) Stylistic words show cross-generator variation (darker blocks indicate higher attention), whereas (b) semantically neutral words remain relatively consistent. (c) Stylistic words exhibit larger preference variance than neutral words.

	Type	DNA*	DP-Net	D-SVDD*	BERT	+L	+LG
Perturbation	*Accuracy (Acc ↑)*						
	Original	63.9	69.3	80.3	75.9	78.6	**81.2**
	Delete	59.6	62.3	70.1	66.9	71.8	**75.1**
	Swap	59.3	64.2	73.4	69.9	71.2	**72.4**
	Insert	59.8	62.8	71.8	67.1	72.5	**75.5**
	Replace	59.8	63.0	70.8	67.5	71.3	**74.6**
	Avg.	59.6	63.1	71.5	67.9	71.7	**74.4**
Adversarial	*Attack Success Rate (ASR ↓)*						
	TextFooler	-	24.7	34.3	35.9	33.8	**23.1**
	BertAttack	-	**19.8**	28.6	31.8	25.7	21.5
	CLARE	-	31.6	40.5	41.3	38.3	**30.2**

Table 3: Robustness evaluation on unseen FLAN-T5 under lexical perturbations and adversarial attacks. DNA* is omitted from adversarial evaluation since it is a zero-shot detector. D-SVDD*: In adversarial evaluation, we report results using HRN from the same work since D-SVDD is incompatible with TextAttack-based attacks.

4.3.1 Lexical Stylistic Analysis

Cross-Generator Preference Analysis. We first analyze whether the proposed lexical stylistic modeling captures generator-sensitive lexical behaviors across different LLM families. As shown in Figure 4(a)/(b), stylistically sensitive words exhibit clear preference variation across generators, while semantically neutral words remain relatively stable with consistent activations. Figure 4(c) further shows substantially larger cross-generator variance for stylistically sensitive words than neutral words. These results suggest that the proposed modeling effectively distinguishes generator-sensitive lexical behaviors from transferable semantic content.

Robustness under Lexical Perturbation. To evaluate the proposed lexical stylistic modeling for capturing generator-sensitive lexical patterns, we evaluate robustness under lexical perturbations and adversarial attacks. As shown in Table 3, introducing lexical stylistic modeling consistently improves robustness over the encoder-only baseline. On the challenging unseen FLAN-T5 setting, $+L$ improves the average perturbed accuracy from 67.9% to 71.7%, while the full framework further improves it to 74.3%. Under adversarial attacks, $+LG$ also achieves consistently lower attack success rates than both BERT and lexical-only variants. These results suggest that suppressing generator-sensitive lexical reliance improves robustness under generator shift.

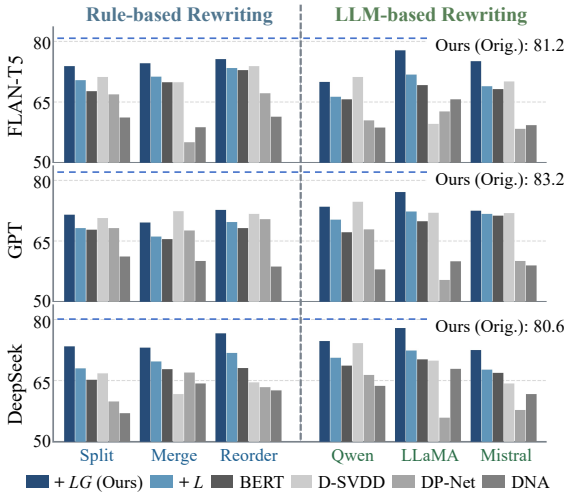

Figure 5: Robustness evaluation under document-level stylistic rewriting. Both rule-based (left) and LLM-based (right) rewriting are applied while preserving the original semantics. Experiments are conducted on the public FLAN-T5 and GPT datasets, together with a newly collected DeepSeek dataset, to evaluate robustness under global stylistic shift.

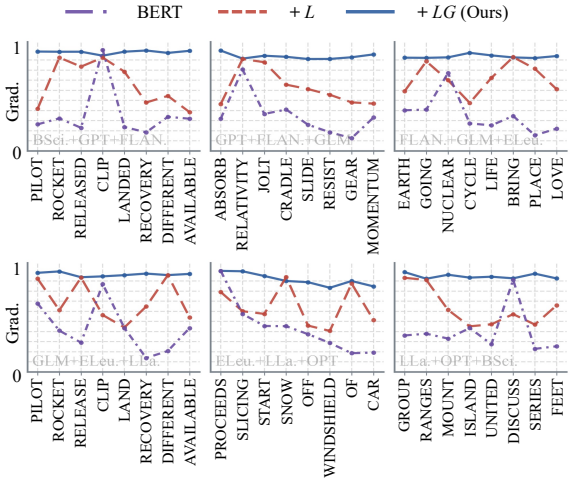

Figure 6: Gradient-based token attribution distributions for BERT, +*L*, and +*LG*. Each subplot corresponds to a different combination of three generator categories selected from six MAGE categories (excluding EleutherAI). Six representative settings are shown due to highly consistent attribution patterns across combinations. Adaptive stylistic regulation (+*LG*) produces substantially smoother and less concentrated token reliance distributions across generators.

4.3.2 Global Stylistic Analysis

Robustness to Global Stylistic Shift. Beyond lexical stylistic behaviors, we further analyze robustness under global stylistic variation using document-level rewriting while preserving the original semantics. As shown in Figure 5, all methods exhibit performance degradation after rewriting, indicating substantial sensitivity to global stylistic variation. In contrast, the proposed framework consistently maintains more stable performance across diverse rewriting conditions. For instance, on the challenging FLAN-T5 setting, +*LG* improves the average rewritten accuracy from 70.8% (+*L*) and 70.2% (BERT) to 74.5%. These results suggest that adaptive stylistic regulation improves robustness under global stylistic shift by reducing excessive reliance on unstable stylistic patterns.

Token Reliance Distribution Analysis. To further understand how adaptive stylistic regulation improves robustness under global stylistic shift, we visualize token-level attribution distributions before and after regulation. As shown in Figure 6, vanilla BERT exhibits highly concentrated attribution patterns, where a few dominant stylistic tokens repeatedly receive disproportionately large attribution scores across different generator settings. While $+L$ partially alleviates this behavior, $+LG$ (Ours) produces substantially smoother and more balanced attribution distributions. These results suggest that adaptive stylistic regulation effectively reduces excessive reliance on dominant global stylistic cues, encouraging more transferable prediction behavior across generators.

5 Conclusion

We present an adaptive stylistic modeling framework for unseen-generator AI-text detection. The proposed framework suppresses unstable generator-sensitive reliance while preserving transferable stylistic evidence across lexical and global levels. Experiments demonstrate improved robustness under diverse unseen-generator shifts.

Limitation

Although our framework models lexical and global stylistic behaviors separately, the specific linguistic patterns driving robustness under different generator shifts are still not fully understood. In addition, robustness against more adaptive adversarial attacks and deliberately style-obfuscated generations remains underexplored.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via

conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pages 24814–24836.

Advik Raj Basani and Pin-Yu Chen. 2025. Diversity boosts AI-generated text detection. In *Data in Generative Models Workshop: The Bad, the Ugly, and the Greats at ICML 2025*.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Xinhui Chen, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Tang Long, Lei Zhang, Chenyu Yan, Guanghao Mei, Jie Zhang, and Lefei Zhang. 2025. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23559–23567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jiayi Gui, Baitong Cui, Xiaolian Guo, Ke Yu, and Xiaofei Wu. 2025. AIDER: A robust and topic-independent framework for detecting AI-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9299–9310.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. 2025. DETree: Detecting human-AI collaborative texts via tree-structured hierarchical representation learning. In *The 39th Conference on Neural Information Processing Systems*.

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are AI-generated text detectors robust to adversarial perturbations? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Robust AI-generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17036–17055.

Xiang Li, Zhiyi Yin, Hexiang Tan, Shaoling Jing, Du Su, Yi Cheng, Huawei Shen, and Fei Sun. 2025a. PrDetect: Perturbation-robust LLM-generated text detection based on syntax tree. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8290–8301.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53.

Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. 2025b. Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3091–3113.

Shengchao Liu, Xiaoming Liu, Yichen Wang, Zehua Cheng, Chengzhengxu Li, Zhaohan Zhang, Yu Lan, and Chao Shen. 2024. Does DetectGPT fully utilize perturbation? bridging selective perturbation to fine-tuned contrastive learning detector would be better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023. CoCo: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24950–24962.

Shinwoo Park, Shubin Kim, Do-Kyung Kim, and Yo-Sub Han. 2025. KatFishNet: Detecting LLM-generated korean text through linguistic feature analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21189–21222.

Xiao Pu, Zepeng Cheng, Lin Yuan, Yu Wu, and Xiuli Bi. 2026. Breaking the generator barrier: Disentangled representation for generalizable AI-text detection. *arXiv preprint arXiv:2604.13692*.

Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. 2024. Ten words only still help: Improving black-box AI-generated text detection via proxy-guided efficient re-sampling. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22015–22036.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-Who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yang Wu, Ruijia Wang, and Jie Wu. 2025. Non-existent relationship: Fact-aware multi-level machine-generated text detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3757–3768.

Sergio E. Zanotto and Segun Aroyehun. 2025. Linguistic and embedding-based profiling of texts generated by humans and large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22852–22869.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. 2025. Human texts are outliers: Detecting LLM-generated texts via out-of-distribution detection. In *The 39th Conference on Neural Information Processing Systems*.

Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, Ying Yang, Shakeel A. O. B. Gavioli-Akilagun, and Chengchun Shi. 2025a. AdaDetectGPT: Adaptive detection of LLM-generated text with statistical guarantees. In *The 39th Conference on Neural Information Processing Systems*.

Yinghan Zhou, Wen-Juan Hou, Wanli Peng, Yiming Xue, Zhengxian Wu, and Ziwei Zhang. 2025b. Kill two birds with one stone: Generalized and robust AI-generated text detection via dynamic perturbations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8864–8875.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. 2025a. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12298–12319.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. 2025b. DNA-DetectLLM: Unveiling AI-generated text via a DNA-inspired mutation-repair paradigm. In *The 39th Conference on Neural Information Processing Systems*.