MGT-Prism: Enhancing Domain Generalization for Machine-Generated Text Detection via Spectral Alignment

Shengchao Liu[1], Xiaoming Liu[1,*], Chengzhengxu Li[1], Zhaohan Zhang[2], Guoxin Ma[1], Yu Lan[1], Shuai Xiao[3]

[1]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University
[2]Queen Mary University of London,
[3]Alibaba
{liusc, czx.li, guoxin.ma}@stu.xjtu.edu.cn, {xm.liu, ylan2020}@xjtu.edu.cn,
zhaohan.zhang@qmul.ac.uk, shuai.xsh@alibaba-inc.com

Abstract

Large Language Models have shown growing ability to generate fluent and coherent texts that are highly similar to the writing style of humans. Current detectors for Machine-Generated Text (MGT) perform well when they are trained and tested in the same domain but generalize poorly to unseen domains, due to domain shift between data from different sources. In this work, we propose MGT-Prism, an MGT detection method from the perspective of the frequency domain for better domain generalization. Our key insight stems from analyzing text representations in the frequency domain, where we observe consistent spectral patterns across diverse domains, while significant discrepancies in magnitude emerge between MGT and human-written texts (HWTs). The observation initiates the design of a low frequency domain filtering module for filtering out the document-level features that are sensitive to domain shift, and a dynamic spectrum alignment strategy to extract the task-specific and domain-invariant features for improving the detector's performance in domain generalization. Extensive experiments demonstrate that MGT-Prism outperforms state-of-the-art baselines by an average of 0.90% in accuracy and 0.92% in F1 score on 11 test datasets across three domain-generalization scenarios.

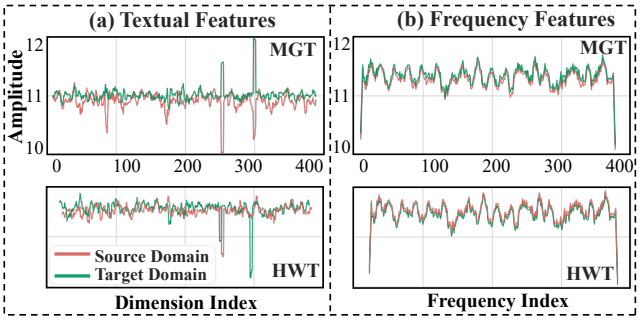

Figure 1: Comparison of the textual features and their corresponding frequency features. The horizontal axis represents the feature dimension index for both textual and frequency-domain features, while the vertical axis shows the magnitude of raw feature values in the text domain and transformed frequency components, respectively. Frequency features exhibit more consistent patterns between intra-class samples from different domains, providing the foundation for frequency-based detection under domain shift[1].

Introduction

Large Language Models (LLMs) are becoming popular as writing assistants in daily work for their incredible ability to generate fluent and coherent texts following users' instructions. However, the widespread applications of LLMs have raised substantial concerns regarding their misuse in fake news generation (Liu et al. 2024a), ghostwriting (Kumar et al. 2025), spamming, etc. (Wang et al. 2024; Li et al. 2025b), which calls for an urgent need to detect MGTs precisely and reliably.

Due to the diversity in application scenarios, model architectures, and scales of LLMs, an ideal MGT detector should perform consistently well in *domain generalization* (DG) settings where the detector is trained on datasets from different source domains and tested on unseen target domains. Two mainstream approaches dominate current MGT detection methods, *i.e.*, fine-tuning and metrics-based methods, yet both largely neglect the detector's ability to generalize. Existing fine-tuning methods either focus on capturing distinguishing features between MGT and HWT, such as text coherence (Liu et al. 2023), token probabilities (Chen et al. 2025), and attention patterns (Kushnareva et al. 2021), or explore novel training strategies, including contrastive learning (Liu et al. 2024b), and adversarial training (Hu, Chen, and Ho 2023; Li et al. 2025b). However, these methods do not explicitly disentangle task-specific and domain-specific features during training, resulting in limited DG. Metric-based methods work in an unsupervised manner. They calculate a single score for distinguishing MGTs from HWTs using perplexity (Bao et al. 2024), entropy (Shum, Diao, and Zhang 2023), or re-sampling (Shi et al. 2024). Metric-based methods also suffer from performance degradation in DG settings, as they rely on a specific training set for deciding

*Corresponding author

[1]The data used in our experiments consist of 3,000 randomly sampled HWTs and 3,000 MGTs from both the source and target domains, covering all three datasets used in domain generalization.

the optimized classification threshold.

Inspired by the potential of the frequency domain to disentangle the feature representations into orthogonal components (Tamkin, Jurafsky, and Goodman 2020a; Sun et al. 2024; Guo et al. 2023), we examine the inter- and intra-class features from the perspective of spectra using Discrete Fourier Transform (DFT) (Bracewell 1986). The preliminary analysis in Sec. reveals a critical property of text representations in the frequency domain. As shown in Figure 1, the spectrum of intra-class text representation from different domains exhibits differences in the magnitude but is more consistent in the distribution pattern, indicating that *i)* distinguishable features between MGT and HWT exist in the frequency domain; *ii)* transfer from features space to frequency domain mitigates the domain shift.

Based on the natures of spectra, which are decomposability and insensitivity to domain shift, we propose MGT-Prism, an MGT detection framework from the perspective of the frequency domain, to align intra-class features from different domains for emphasizing the task-specific and domain-invariant inter-class features, therefore enabling strong DG ability of MGT detectors. Specifically, we design a frequency spectrum filter block to suppress domain-related features (Tamkin, Jurafsky, and Goodman 2020a). To mitigate the domain shift in finer-grain, we further introduce a frequency spectrum alignment strategy to reduce the distribution discrepancy among intra-class instances in the frequency domain. By aligning intra-class features in the frequency domain, MGT-Prism successfully extracts and utilizes the task-specific and domain-invariant features from the training set and generalizes well to unseen domains. Extensive experiments demonstrate the strong generalization ability of MGT-Prism among multiple settings. Compared to state-of-the-art approaches in accuracy and F1 score, MGT-Prism achieves average improvements of 0.92% and 1.56%, 0.90% and 0.69%, and 0.88% and 1.24% in cross-generator, cross-domain, and cross-scale, respectively, highlighting its efficacy. Our contributions are summarized as follows:

- We propose to analyze MGT detection in the frequency domain and observe a similar distribution pattern with varied magnitudes between MGT and HWT spectra, offering a new insight and broad applicability in MGT detection.
- We design a novel model, MGT-Prism, which aligns the intra-class features in text spectra from different domains to extract task-specific and domain-invariant features for enhancing domain generalizability in MGT detection.
- We conduct extensive experiments on 11 test sets across three generalization scenarios. The results demonstrate that our model consistently outperforms state-of-the-art methods in both effectiveness and generalization capability for MGT detection.

Related Work

Domain Generalization. In the context of MGT detection, DG means a detector should remain reliable when faced with text from new generators or domains not encountered during training. Many metric-based detectors rely on head-token analysis (Gehrmann, Strobelt, and Rush 2019), logistic regression on perplexity features (Bao et al. 2024, 2023), or token-wise log probabilities (Wang et al. 2023a; Hans et al. 2024) from white-box LLMs to address this challenge. While metric-based methods are training-free, their prediction accuracy is often lower than supervised approaches. In contrast, fine-tuned detectors benefit from perturbation-based methods (Li et al. 2025a; Shum, Diao, and Zhang 2023), contrastive learning (Tack et al. 2020; Gunel et al. 2021), or a combination of both (Liu et al. 2024a), which have proven effective in improving model generalization. In addition, adversarial training frameworks (Li et al. 2025b; Hu, Chen, and Ho 2023) have been employed to address robustness issues. Unlike prior methods based on overall text features, our approach introduces the Fourier Transform to extract multi-scale frequency information, emphasizing both global and local patterns, and mitigates distributional bias to enable effective detection in DG.

Fourier Transforms. Recently, frequency-based techniques have been increasingly integrated into the computer vision field (Yi et al. 2023; Guo et al. 2023). By analyzing how low frequencies in an image typically capture global structures (Fan et al. 2022) and color information (Cao et al. 2020), while high frequencies contain fine details of objects, a series of methods to enhance predictive capabilities have emerged (Fan et al. 2022). Some works have applied these techniques to the natural language processing (NLP) field (Wu et al. 2021; Lee-Thorp et al. 2021), combining them with attention mechanisms to accelerate computations (Tamkin, Jurafsky, and Goodman 2020a; Lee-Thorp et al. 2021; Choromanski et al. 2020), and embedding spectral filters in different tasks (Fang and Xu 2024; Tamkin, Jurafsky, and Goodman 2020a; Khan, Hayat, and Porikli 2019), where the low-frequency components are considered to represent slower-changing features. Inspired by this, we observe that transforming representations from the feature space to the frequency domain reduces dimensional complexity. It also enhances the structural regularity of feature distributions in frequency domain. This transformation amplifies distributional irregularities (*e.g.*, lexical repetition, unnatural conjunctions, and templated phrasing), making MGT and HWT more distinguishable for downstream detection.

Preliminary

This section briefly introduces the problem formulation and the DFT in traditional signal processing and its application in our work.

MGT Detection Under DG

The DG[2] task is defined as follows: given a training set consisting of multiple observed source domains $\mathcal{D}_S = \{\mathcal{D}_n\}_{n=1}^{N}$, where each domain $\mathcal{D}_n = \{(x_i^{(n)}, y_i^{(n)})\}_{i=1}^{T_n}$ contains T_n labeled samples, the goal is to learn a model from $\mathcal{D}_S$ that generalizes to arbitrary unseen target domains $\mathcal{D}_T$ whose distributions differ from those of the sources.

[2]We use *domain generalization* to refer to generalization across domains, generators, and generator scales in MGT detection.

Discrete Fourier Transform

The Continuous Fourier Transform (CFT) is one of the core mathematical tools in the field of signal processing (Bracewell 1986). The idea is that any complex signal can be decomposed into a superposition of sine waves (or complex exponentials) of different frequencies. Specifically, for any time-domain signal $x(t)$, its frequency component $X(f)$ at frequency f in the frequency domain is expressed as:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi j f t} dt, \tag{1}$$

where $e^{-2\pi j f t}$ is the complex exponential basis and j denotes the imaginary unit. The DFT extends the CFT to discrete signals, where $x[n]$ denotes uniformly sampled values of $x(t)$. The k-th component in the frequency domain is:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-2\pi j \frac{kn}{N}}, \quad 0 \le k \le (N-1), \tag{2}$$

where N represents the length of the signal, which is the number of sampling points of the input sequence $x[n]$.

Applying DFT to Contextual Word Representations

Given an input sequence $x[n]$, an encoder-only Transformer (*e.g.*, RoBERTa) produces final-layer token vectors $V = (v_0, \ldots, v_{N-1})$, where $v_t \in \mathbb{R}^d$. Fix a neuron $i \in \{0, \ldots, d-1\}$ and consider its trajectory across tokens $s^{(i)} = (v_0[i], \ldots, v_{N-1}[i]) \in \mathbb{R}^N$. Building on Eq. (2), which treats the final-layer `[CLS]` vector as a length-d signal along the hidden dimension, we generalize to the token axis: fixing neuron i, we apply the same DFT to its activations across tokens $s^{(i)}$. The spectral coefficients $h_k^{(i)}$ are:

$$h_k^{(i)} = \sum_{t=0}^{N-1} s_t^{(i)} \, e^{-2\pi \mathrm{i} k t / N}, \qquad k = 0, \ldots, N-1. \tag{3}$$

Here $k = 0$ is the lowest component (Direct Current component). Low-, high-, or band-pass effects are obtained by applying a spectral mask m_k (*e.g.*, zeroing selected $h_k^{(i)}$).

Why Frequency Helps MGT Detection Generalize Across Domains

A common property of modern NLP models is their ability to produce *contextualized* token representations by modeling a sequence of tokens (e.g., characters or subword units). We collect the frequency-domain features as $\mathbf{H} = \{\mathbf{h}_k\}_{k=0}^{N-1}$, where $\mathbf{h}_k = [h_k^{(0)}, \ldots, h_k^{(d-1)}]^\top$. To capture global structural patterns, we keep components that complete at most one full oscillation over the input. Consequently, we define the low-frequency band as:

$$\mathbf{H}_{\text{low}} = \{\mathbf{h}_k\}_{k=0}^{d_{\text{low}}}, \quad d_{\text{low}} = \left\lceil \frac{N}{t_{\text{num}}} \right\rceil. \tag{4}$$

where t_{num} is the number of tokens in the input text. Similarly, to capture sentence-level features, we retain components with at most one oscillation per sentence, based on the sentence count s_{num}:

$$\mathbf{H}_{\text{mid}} = \{\mathbf{h}_k\}_{k=d_{\text{low}}+1}^{d_{\text{mid}}}, \quad d_{\text{mid}} = \left\lceil \frac{N}{s_{\text{num}}} + (N - d_{\text{low}}) \right\rceil. \tag{5}$$

After obtaining the mid frequency band $\mathbf{H}_{\text{mid}}$, we divide the remaining frequency features into high frequency band:

$$\mathbf{H}_{\text{high}} = \{\mathbf{h}_k\}_{k=d_{\text{mid}}+1}^{N-1}. \tag{6}$$

The inverse DFT (IDFT; Lee-Thorp et al. (2021)), which transforms the input from the frequency domain $\mathbf{H}_{\text{f}}$ back to the text feature space H_f, can be expressed as:

$$H_f = \text{IDFT}(\mathbf{H}_{\text{f}}) \tag{7}$$

Previous studies (Tamkin, Jurafsky, and Goodman 2020b) have shown that the high-frequency components reflect word-level features (*e.g.*, perplexity, log-probability), while low-frequency components are associated with document-level characteristics (*e.g.*, topic, style). Correspondingly, recent studies DetectGPT(Mitchell et al. 2023) and Binoculars (Hans et al. 2024) demonstrate that the distinct writing preference at the word-level plays a key role in MGT detection. Therefore, we conduct two complementary validations [3]: ***i)*** **in the frequency domain**, we examine whether high-frequency components are particularly sensitive to token-level changes; and ***ii)*** **in the feature space** , we assess whether low-frequency components effectively capture global structural features. These findings support the subsequent use of low-frequency filtering and frequency-domain alignment to improve DG in MGT detection.

Analyzing the Correlation Between Frequency Components and Feature Space. To validate whether different frequency components effectively capture distinct linguistic features, we apply three types of content-preserving perturbations in the feature space, namely token perturbations (word-level), sentence reordering (sentence-level), and theme transformations (document-level). As shown in Table 1, token-level perturbations tend to cause larger MAE[4] values in the high-frequency components, while theme transformations primarily increase the MAE in the low-frequency components compared to other perturbations. This indicates that high-frequency components are more sensitive to token-level changes and predominantly encode token-level features. Conversely, low-frequency components are more sensitive to document-level perturbations, capturing domain-specific information (*e.g.*, topic, style).

Perturbation	low-frequency	mid-	high-
Theme-Transformation	**0.0072**	0.0166	0.0349
Sentence-Reordering	0.0002	**0.0744**	0.0250
Token-Replacement	0.0041	0.0867	**0.3225**
Token-Delete	0.0034	0.2453	**0.4122**
Token-Repetition	0.0048	0.2703	**0.5592**

Table 1: MAE Shift in Frequency Bands.

[3] Our experimental data are identical to those used in Figure 1.

[4] Mean Absolute Error is a standard metric for measuring signal-level differences, and a higher MAE reflects greater spectral deviation between the perturbed and original inputs in the frequency domain (Yi et al. 2023). The perturbation rate is 15%.

Topic Coherence in Feature Space. To evaluate the coherence between frequency components and global semantic features, we project each frequency band back into the feature space via IDFT and compute its similarity to the original document-level embedding using BERTScore[5] . As shown in Table 2, low-frequency components exhibit the highest similarity, whereas high-frequency components score the lowest, indicating that low-frequency features better preserve global topical information.

Perturbation	low-frequency	mid-	high-
BERTSore	**0.8671**	0.4062	0.1945

Table 2: Topic Coherence Compared to the Original Text.

Methodology

The workflow of the proposed MGT-Prism is shown in Figure 2, comprising two main components: a Low Frequency Filtering Module (LFF), and a Spectrum Alignment Module (SAM), to reduce the domain gap for intra-class samples. LLF cuts out the low-frequency band to reduce domain-sensitive but task-irrelevant features. SAM sets an optimization objective to mitigate domain gap for data within the same class.

Low Frequency Filtering Module

Previous studies (Tamkin, Jurafsky, and Goodman 2020b; Sun et al. 2024) on frequency domain analysis in the field of natural language processing and analysis about the discrepancy between MGT and HWT (Chen et al. 2025; Liu et al. 2024b) observe that the low-frequency band corresponds with document-level features. We transform the text input into a frequency domain representation (Lee-Thorp et al. 2021; Yi et al. 2023) and decompose it into three frequency bands (*i.e.*, low, mid, and high), and propose a low-frequency filtering module to suppress the common features between MGT and HWT.

With the improvement of text generation capabilities of LLMs, MGT hardly differs from HWT at the discourse level (Chen et al. 2025; Liu et al. 2024a). In frequency domain analysis, this phenomenon is reflected in the fact that low frequency components are often smoother (Tamkin, Jurafsky, and Goodman 2020a). Furthermore, in the DG settings, changes in style, generator, or theme are more likely to cause the migration of text document-level features. Therefore, we filter out low-frequency components $\mathbf{H}_{\text{low}}$ while retaining mid and high frequency components to enhance the classifier's ability to distinguish between MGT and HWT under DG. The final frequency domain features after low-frequency filtering can be expressed as:

$$d_{\text{mid}} = \left\lceil \frac{N}{s_{\text{num}}} \cdot \tau + (N - d_{\text{low}}) \cdot (1 - \tau) \right\rceil, \tag{8}$$

[5]BERTScore (Angelov and Inkpen 2024) is a widely used metric for assessing the topic coherence of learned representations. Higher scores indicate stronger global semantic coherence between the reconstructed embeddings and the original embeddings.

$$\mathbf{H}_{\text{mid}} = \{\mathbf{h}_k\}_{k=d_{\text{low}}+1}^{d_{\text{mid}}}, \quad \mathbf{H}_{\text{f}} = \mathbf{H}_{\text{mid}} \oplus \mathbf{H}_{\text{high}}, \tag{9}$$

where $\tau \in [0, 1]$ is the scaling factor. By adjusting τ, the mid-frequency band will be affected by the number of sentences and the remaining frequency bands, avoiding the overflow problem caused by too few or too many sentences. $\oplus$ denote the concatenation operation. From the perspective of natural text, $\mathbf{H}_{\text{f}}$ retains more fine-grained information of words and sentences in the text input x.

Spectral Alignment Module

Frequency Spectrum Reconstruction. After removing the low-frequency features that are easily affected by domain-shift, we restore the remaining frequency-domain features to the original features to reduce the detector's excessive attention to specific domain information during the training process. Specifically, given a batch of training data $D = \{x_b, y_b\}_{b=0}^{B-1}$ with B text-label pairs. Through the previous section, we can obtain the frequency domain features representation $\{\mathbf{H}_{\text{f}}^b\}_{b=0}^{B-1} = \{\mathbf{H}_{\text{mid}}^b \oplus \mathbf{H}_{\text{high}}^b\}_{b=0}^{B-1}$ of each input text T_b. Subsequently, we compute the average modulus values of the frequency-domain features within the mid- and high-frequency intervals for the current batch D. These values are then compared with the corresponding average modulus values of the entire training dataset to determine the weighting factors α_{mid} and α_{high} for reconstructing the features in the mid and high frequency bands.

$$\mu_{\text{mid}} = \frac{1}{B}\sum_{b=0}^{B-1}\left|\mathbf{H}_{\text{mid}}^b\right|, \ \ \mu_{\text{high}} = \frac{1}{B}\sum_{n=0}^{B-1}\left|\mathbf{H}_{\text{high}}^b\right|, \tag{10}$$

$$\alpha_{\text{mid}} = \frac{\bar{\mu}_{\text{mid}}}{\mu_{\text{mid}}}, \ \ \alpha_{\text{high}} = \frac{\bar{\mu}_{\text{high}}}{\mu_{\text{high}}}, \tag{11}$$

where $|.|$ represents the modulus operation, $\bar{\mu}_{\text{mid}}$ and $\bar{\mu}_{\text{high}}$ represent the average modulus values of the mid and high frequency of the entire training set, respectively. Therefore, the final frequency domain features can be expressed as:

$$\mathbf{H}_{\text{f}}^b = \alpha_{\text{mid}}\mathbf{H}_{\text{mid}}^b \oplus \alpha_{\text{high}}\mathbf{H}_{\text{high}}^b, \ \ 0 \le b \le (B-1). \tag{12}$$

The natural text features after filtering and reconstruction can be expressed as:

$$H_f^b = \text{IDFT}(\mathbf{H}_{\text{f}}^b), \ \ 0 \le b \le (B-1), \tag{13}$$

By low frequency filtering and reconstructing frequency domain features, we strip away the global information that is less discriminative between MGT and HWT, thereby promoting more effective learning of local features.

Frequency Spectrum Alignment. To further promote feature alignment among samples from the same class and enhance generalization to unseen domains, we introduce a frequency domain feature alignment loss $\mathcal{L}_{\text{MAE}}$. Building on the standard signal level dissimilarity metric, mean absolute error (MAE) (Yi et al. 2023), $\mathcal{L}_{\text{MAE}}$ reduces distributional differences by minimizing the average L_1 distance between samples from the same class in the frequency domain. Specifically, given a batch of training data D and corresponding frequency features $\{\mathbf{H}_{\text{f}}^b\}_{b=0}^{B-1}$, for any frequency features $\mathbf{H}_{\text{f}}^b$, let the set of remaining frequency features with

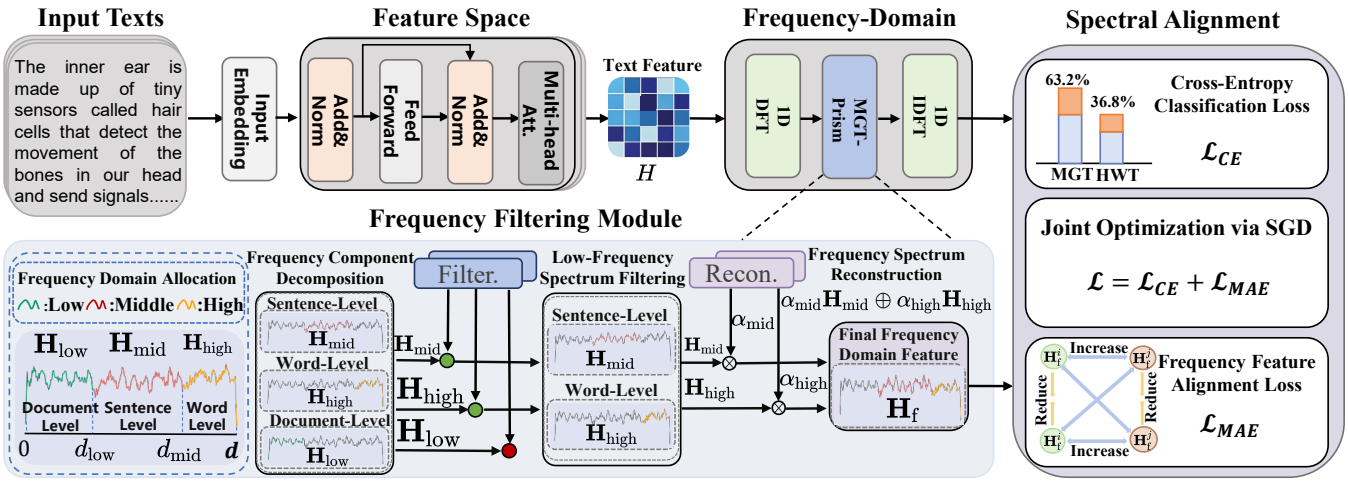

Figure 2: Overview of MGT-Prism. In the Low Frequency Filtering Module (Sec.), we transform features from feature space into the frequency domain, where we analyze information distribution across multiple frequency bands (*i.e.*, low-, mid-, and high-frequency). Then, we propose a low-frequency filtering module to suppress redundant features shared between MGT and HWT. In the Spectral Alignment stage (Sec.), we compute the global frequency spectrum distribution and reconstruct the frequency components accordingly. Furthermore, we design a frequency alignment loss to enhance DG.

the same label as D_{pos}, and the set of remaining frequency features with different labels be denoted as D_{neg}. Then the loss function $\mathcal{L}_{\text{MAE}}$ can be defined as:

$$\mathcal{L}_{\text{pos}} = \mathbb{E}_{\mathbf{H}_{\text{f}}^{b} \in D} \mathbb{E}_{\mathbf{H}_{\text{f}}^{i} \in D_{\text{pos}}} (\|(|\mathbf{H}_{\text{f}}^{b}| - |\mathbf{H}_{\text{f}}^{i}|)\|_1), \tag{14}$$

$$\mathcal{L}_{\text{neg}} = \mathbb{E}_{\mathbf{H}_{\text{f}}^{b} \in D} \mathbb{E}_{\mathbf{H}_{\text{f}}^{i} \in D_{\text{neg}}} (\max(0, (\xi - \|(|\mathbf{H}_{\text{f}}^{b}| - |\mathbf{H}_{\text{f}}^{i}|)\|_1))), \tag{15}$$

$$\mathcal{L}_{\text{MAE}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}, \tag{16}$$

where $\| \cdot \|_1$ denotes the MAE computed over the modulus values, ξ represents the maximum same class distance, which is used to control the distance between samples with different labels. Then, the total loss is computed by:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MAE}}, \tag{17}$$

where the cross-entropy classification loss $\mathcal{L}_{\text{CE}}$ is computed on natural text features H_f^b, describing in Eq. (13), and corresponding labels Y_b.

Experiments

Experiment Settings

To evaluate DG in MGT detection, we conduct extensive experiments on three open-source datasets, including M4 (Wang et al. 2023b), DetectRL (Wu et al. 2024), and MAGE (Li et al. 2024) dataset, under three experimental settings: ***i)*** **Cross-Domain**, composed of the DetectRL and MAGE datasets, and including MGTs from Opinion Statement (Opinion.S), Question Answering (Question.A), Story generation (Story.G), and Scientific Writing (Scientific.W). ***ii)*** **Cross-Generator**, composed of the M4 and MAGE datasets, in which we obtain MGTs generated by Flan-T5 (Chung et al. 2024), ChatGPT (OpenAI 2025), GLM (Zeng et al. 2022), and LLaMA (Touvron et al. 2023). ***iii)*** **Cross-Scale**, based on the MAGE dataset, including MGTs generated by LLaMA2-13B, LLaMA2-30B, and LLaMA2-65B.

We randomly sample 1,000 training instances balanced across datasets, domains, and categories. Training is performed for 30 epochs using AdamW ($\epsilon = 2 \times 10^{-5}$), with a learning rate of 0.01 and a scaling factor τ of 0.6.

Competitors

We evaluate MGT-Prism against nine methods for MGT detection, including metric-based and fine-tuned detectors.

Metric-based detectors. Log-probabilities from a generative LM are often used for classification with a predefined threshold [6], including *GLTR* (Gehrmann, Strobelt, and Rush 2019), *DetectGPT* (Mitchell et al. 2023), *Fast-DetectGPT* (Bao et al. 2024), *Binoculars* (Hans et al. 2024) [7].

Fine-tuned detectors. Supervised detectors trained on a PLM, typically optimized with a classification loss, including *RoBERTa* (Liu et al. 2019), *Ghostbuster* (Verma et al. 2023), *RADAR* (Hu, Chen, and Ho 2023), *PECOLA* (Liu et al. 2024a), *ImBD* (Chen et al. 2025).

Performance Comparison

We report the experimental results in Table 3. We first compare MGT-Prism with state-of-the-art metric-based and model-based detectors. In the DG setting, compared to the strongest baselines RoBERTa and Binoculars, MGT-Prism achieves average improvements of 3.62% in accuracy and 4.75% in F1 score, demonstrating strong DG. Broadly speaking, the mostly fine-tuned detectors outperform metric-based methods across all datasets because the performance significantly drops when the scoring model differs from the target model (Mitchell et al. 2023). Moreover,

[6] We utilizing the GPT-Neo-2.7B (Black et al. 2021) to align with Fast-DetectGPT (Bao et al. 2024) experiments.

[7] Under the original Falcon-7B and Falcon-7B-Instruct setting.

Method			Metric-based				Model-based					
Dataset	**Test data**	Metric	*GLTR*	*DetectGPT*	*Fast-Dete.*	*Binoculars*	*RoBERTa*†	*RADAR*	*Ghostbuster*	*PECOLA*†	*ImBD*	*MGT-Prism*†
Cross–Generator	FLAN-T5	Acc	67.10	60.30	74.70	63.46	$\underline{86.80}_{1.90}$	$65.82_{4.06}$	$79.55_{4.16}$	$86.30_{1.80}$	$71.55_{2.06}$	$\mathbf{89.62_{1.67}}$
		F1	52.11	49.33	60.70	52.70	$\underline{85.95}_{1.76}$	$60.75_{5.06}$	$75.45_{4.26}$	$\underline{86.20}_{1.62}$	$65.54_{3.46}$	$\mathbf{89.02_{1.97}}$
	ChatGPT	Acc	78.12	60.71	76.60	$\underline{87.60}$	$85.01_{3.74}$	$61.50_{3.01}$	$77.06_{2.04}$	$\underline{84.21}_{3.50}$	$85.68_{3.76}$	$\mathbf{88.96_{2.03}}$
		F1	62.01	68.70	69.70	$\underline{87.63}$	$84.65_{1.76}$	$69.95_{5.15}$	$72.65_{5.07}$	$84.25_{2.01}$	$79.55_{3.26}$	$\mathbf{89.22_{1.67}}$
	GLM	Acc	75.50	73.20	76.02	**95.28**	$89.94_{1.68}$	$70.04_{4.09}$	$80.94_{2.58}$	$89.60_{2.70}$	$87.48_{2.06}$	$\underline{92.42}_{\mathbf{1.56}}$
		F1	60.10	62.43	69.62	**95.36**	$86.95_{2.06}$	$62.90_{3.09}$	$74.65_{4.13}$	$84.22_{2.01}$	$89.04_{3.76}$	$\underline{92.82}_{\mathbf{1.67}}$
	LLaMA	Acc	72.40	70.80	79.65	80.95	$\underline{85.31}_{2.67}$	$70.12_{4.67}$	$80.45_{2.76}$	$82.99_{2.61}$	$84.22_{3.07}$	$\mathbf{87.66_{1.74}}$
		F1	60.17	62.33	61.60	83.15	$\underline{80.85}_{3.66}$	$67.95_{3.05}$	$79.65_{4.01}$	$84.20_{1.92}$	$\underline{85.34}_{2.06}$	$\mathbf{87.20_{1.77}}$
Cross–Domain	Opinion.S	Acc	71.15	70.40	76.30	$\underline{95.04}$	$93.61_{2.65}$	$74.39_{3.05}$	$90.97_{2.91}$	$93.49_{3.06}$	$94.85_{0.33}$	$\mathbf{96.06_{1.08}}$
		F1	65.60	60.70	72.70	$\underline{95.20}$	$92.24_{1.93}$	$60.04_{2.56}$	$91.27_{1.83}$	$91.14_{2.20}$	$92.58_{2.06}$	$\mathbf{96.90_{1.53}}$
	Question.A	Acc	72.90	72.20	76.90	93.60	$\underline{93.92}_{1.15}$	$72.52_{1.95}$	$92.82_{4.06}$	$94.12_{2.95}$	$93.23_{0.92}$	$\mathbf{96.53_{1.45}}$
		F1	70.60	76.10	70.80	$\underline{93.62}$	$93.04_{1.93}$	$78.64_{3.64}$	$91.45_{2.74}$	$92.64_{2.10}$	$92.28_{1.06}$	$\mathbf{96.20_{1.13}}$
	Scientific.W	Acc	41.24	50.90	66.90	$\underline{79.92}$	$70.43_{8.29}$	$70.18_{10.79}$	$62.73_{9.19}$	$68.90_{9.17}$	$\mathbf{80.26_{5.10}}$	$79.88_{7.22}$
		F1	32.60	35.22	48.70	**79.95**	$61.94_{8.23}$	$50.64_{8.94}$	$55.95_{6.93}$	$60.25_{7.80}$	$74.28_{4.56}$	$\underline{77.80}_{\mathbf{7.43}}$
	Story.G	Acc	70.60	60.22	76.70	90.62	$\underline{98.62}_{0.42}$	$75.76_{3.23}$	$95.60_{1.32}$	$98.20_{0.40}$	$92.75_{0.54}$	$\mathbf{98.96_{0.25}}$
		F1	65.25	67.06	70.90	90.90	$\underline{98.04}_{1.63}$	$70.84_{3.06}$	$94.65_{2.63}$	$97.09_{2.00}$	$93.08_{2.76}$	$\mathbf{98.90_{1.03}}$
Cross–Scale	LLaMa-13b	Acc	73.92	71.20	82.90	94.65	$93.85_{1.72}$	$76.28_{2.60}$	$90.95_{1.92}$	$93.48_{2.12}$	$\underline{94.91}_{0.95}$	$\mathbf{95.94_{1.62}}$
		F1	70.60	75.02	78.98	$\underline{93.70}$	$92.74_{1.50}$	$70.60_{3.14}$	$92.50_{1.93}$	$92.14_{2.01}$	$\underline{92.78}_{2.06}$	$\mathbf{95.91_{1.70}}$
	LLaMa-30b	Acc	77.90	73.20	80.62	93.65	$93.28_{1.20}$	$77.49_{1.25}$	$94.44_{2.10}$	$\underline{94.44}_{1.19}$	$92.26_{1.71}$	$\mathbf{94.67_{1.20}}$
		F1	73.90	77.25	84.21	$\underline{94.02}$	$92.60_{1.93}$	$71.64_{3.14}$	$92.15_{2.76}$	$93.04_{1.40}$	$90.28_{3.66}$	$\mathbf{93.70_{1.73}}$
	LLaMa-65b	Acc	75.60	74.90	77.92	92.05	$\underline{92.24}_{1.63}$	$74.70_{2.60}$	$91.95_{2.03}$	$92.06_{1.10}$	$86.38_{2.90}$	$\mathbf{93.60_{2.13}}$
		F1	70.70	72.12	78.09	91.40	$\underline{92.10}_{1.43}$	$70.14_{3.20}$	$92.05_{2.13}$	$91.14_{1.70}$	$83.48_{3.96}$	$\mathbf{93.92_{1.73}}$
Average		Acc	70.58	67.09	76.83	87.82	$\underline{89.36}$	71.70	85.04	88.89	87.59	**92.21**
		F1	62.14	64.20	69.63	87.06	$\underline{87.37}$	66.73	82.94	86.93	85.29	**91.96**

Table 3: Accruacy and F1 score (%) of MGT-Prism and baseline methods for MGT detection under the DG setting. The results are average values of 10 runs with different random seeds. The subscript means the standard deviation (*e.g.*, $93.92_{1.73}$ means 93.92 ± 1.73). Metric-based methods' results are deterministic, so we do not report standard deviation. Also, these metric-based methods must have the white-box generator as the base model, which is different from the model-based methods. † denotes using the RoBERTa-base (125M) as the backbone model. The best and second-best are **bolded** and underlined respectively.

the detection results in the cross-generator setting are generally lower than those in the cross-domain and cross-scale settings, likely due to more severe domain shifts introduced by different generators. It demonstrates that the different generation paradigms across models lead to a larger domain shift, degrading the model performance. In the cross-domain setting, testing the model on the Scientific.W dataset leads to consistent low accuracy below 81% because of the large difference between general writing and the scitific writing which uses more technical words (Liu et al. 2024a).

Ablation Study

We conduct ablation experiments to understand the contribution of each component of MGT-Prism in the cross-generator setting. The core components we examine are: Low Frequency Filtering Module (LFF), Frequency Spectrum Reconstruction (FSR), and Frequency Spectrum Alignment (FSA). As shown in Table 4, we find that every block in MGT-Prism contributes to the DG of MGT detector, shown by the improvement over the RoBERTa baseline. Moreover, the FSA module contributes to the detection performance most, indicating the alignment of frequency-domain features plays an important role in mitigating domain shift. In average, the combination of two components is always superior to only incorporating one module into the training process,

Modules			Cross-Generator				
LFF	**FSR**	**FSA**	**Flan-T5**	**ChatGPT**	**GLM**	**LLaMa**	**Avg.**
–	–	–	85.95	84.65	86.95	80.85	84.60
✓	–	–	87.84	85.61	88.61	84.92	86.75
–	✓	–	86.40	84.82	89.15	83.91	86.07
–	–	✓	88.05	85.27	89.01	85.28	86.90
✓	✓	–	87.42	86.49	90.02	85.81	87.44
✓	–	✓	88.04	88.05	89.99	85.71	87.94
–	✓	✓	87.21	86.85	89.25	86.93	87.56
✓	✓	✓	**89.02**	**89.23**	**92.82**	**87.20**	**89.57**

Table 4: F1 score (%) for Ablation study on different module combinations. ✓ means that we keep the corresponding block and - means the block is removed.

indicating that different modules are complementary to each other and collectively boost the DG of MGT detector. We also discuss the effects of different scale factors and test lengths in Appendix.

Discussion

Effect of Individual Frequency Components. Compared to style and theme, which are coarse-grained and unstable

across domains or generators, sentence structure and token-level statistics provide more reliable signals for detection. While many existing methods exploit such fine-grained cues (e.g., LM probabilities and perturbations), our approach further aligns multi-granular features with frequency bands. As shown in Table 5, results reveal that using only the low-frequency band leads to an average F1 score of 83.22%, representing a 1.38% drop compared to the all-features model. In contrast, isolating the mid- and high-frequency bands individually leads to performance gains of 1.01% and 1.47%, respectively, compared to the all-features model.

Modules			Cross-Generator				
Low-	Mid-	High-	Flan-T5	ChatGPT	GLM	LLaMa	Avg.
✓	–	–	85.01	83.64	85.29	79.96	83.22
–	✓	–	86.01	**85.31**	88.64	**83.49**	85.61
–	–	✓	**87.65**	85.07	**89.49**	83.08	**86.07**
✓	✓	✓	85.95	84.65	86.95	80.85	84.60

Table 5: F1 score from Ablation on Individual Components.

Robustness under Perturbation. We test the robustness of detectors under four attacks (*i.e.*, Delete, Insert, Repeat, and Generate) with a perturbation rate of 15%, following Wang et al. (2024). As shown in Table 6, our method exhibits a consistently smaller drop in F1 score than the baselines RoBERTa and Binoculars. MGT-Prism achieves an average improvement of 5.29% in F1 score. Specifically, under the deletion and repetition perturbations, the averaged F1 score of MGT-Prism decreases by only 5.10% and 5.23% in the two scenarios (*i.e.*, cross-generator and cross-domain), underscoring its remarkable robustness. The complete evaluation results for all attacks are provided in the Appendix.

Perturbation	Delete	Insert	Repeat	Generate	Avg.
Binoculars					
Cross-Generator	65.41	65.12	65.66	59.59	63.95
Cross-Domain	83.97	81.10	82.45	67.45	<u>78.74</u>
Cross-Scale	84.78	76.19	68.53	61.55	<u>72.76</u>
RoBERTa-base					
Cross-Generator	75.065	59.91	81.285	60.03	<u>69.07</u>
Cross-Domain	85.35	74.81	81.40	72.90	78.61
Cross-Scale	83.15	73.22	68.16	65.06	72.39
MGT-Prism					
Cross-Generator	79.61	67.99	84.58	67.40	**74.90**
Cross-Domain	89.32	77.12	85.12	77.67	**82.30**
Cross-Scale	85.74	76.53	73.95	69.51	**76.43**

Table 6: Performance on diverse perturbation attacks. Results are reported as the average F1 score (%) across three domain generalization settings and four perturbation scenarios. GPT-2 XL (1.5B) (Solaiman et al. 2019) is employed to construct perturbations, including generating and inserting.

Applicability to Different Backbones. We evaluate the effectiveness of MGT-Prism with RoBERTa-large and BERT-large and Qwen3-0.6B (Zhang et al. 2025) as the backbones in both in-domain (IND) and DG settings. As shown in Table 7, MGT-Prism always outperforms vanilla fine-tuning with different backbones, in all settings. Specifically, MGT-Prism improves over the vanilla fine-tuning by 1.16% and 2.11% in RoBERTa-large, and by an average of 1.53% on Qwen3-0.6B. The results demonstrate that MGT-Prism is applicable to different backbones with different scales. More test results and feature distribution visualizations are provided in the Appendix.

Model	Method	Generator	Domain	Scale	Avg.
In-domain (IND)					
RoBERTa-base	base	93.45	96.01	94.90	94.78
	MGT-Prism	94.89	97.43	95.32	95.88
RoBERTa-large	base	94.75	97.87	96.24	96.28
	MGT-Prism	**96.05**	**98.98**	**97.29**	**97.44**
BERT-large	base	88.21	94.49	92.23	91.64
	MGT-Prism	90.01	95.24	92.94	92.73
QWen3-0.6B	base	93.81	96.82	95.82	95.48
	MGT-Prism	<u>95.28</u>	<u>98.24</u>	<u>96.49</u>	<u>96.67</u>
Domain Generalization (DG)					
RoBERTa-large	base	<u>88.13</u>	<u>93.10</u>	93.81	<u>91.49</u>
	MGT-Prism	**90.84**	**95.02**	**95.37**	**93.60**
BERT-large	base	81.4	82.68	89.88	84.17
	MGT-Prism	84.18	88.71	92.97	88.22
QWen3-0.6B	base	86.17	89.08	93.86	89.32
	MGT-Prism	87.47	92.04	<u>95.03</u>	91.19

Table 7: F1 score (%) Comparison with Different Backbones. For the DG setting, the reported results are averaged across all test subsets. For the IND setting, all subcategories are merged for training and testing. The best and second-best are **bolded** and <u>underlined</u> respectively.

Conclusion

In this work, we focus on enhancing the DG ability of MGT detectors with concentration on mitigating domain shift brought by data sources. We transfer the text representation to frequency domain using DFT and find that the spectra of text features mitigate the domain shift among intra-class samples. Based on the decomposability of frequency domain, we design a low-frequency spectrum filtering module to remove the low-frequency feature which is more affected by the domain shift. Moreover, we align the spectra of intra-class samples from different domains with frequency spectrum alignment module to mitigate the domain shift, therefore enabling the model to learn task-specific and domain-invarient features from frequency domain. Experimental results show that our method outperforms SOTA metric-based and model-based baselines. We hope that our work can inspire future research in AI-generated content detection in other modalities, and serve as a foundation for developing a unified detecting approach. The code and datasets are released at https://github.com/lsc-1/MGT-Prism.

Acknowledgements

We thank all the anonymous reviewers and the area chair for their helpful feedback, which aided us in greatly improving the paper. This work is supported by National Key R&D Program (2023YFB3107400), National Natural Science Foundation of China (62272371, 62103323, U21B2018, 62161160337, U20B2049), Initiative Postdocs Supporting Program (BX20190275, BX20200270), China Postdoctoral Science Foundation (2019M663723, 2021M692565), Fundamental Research Funds for the Central Universities under grant (xzy012024144), and Shaanxi Province Key Industry Innovation Program (2021ZDLGY01-02).

References

Angelov, D.; and Inkpen, D. 2024. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13528–13539.

Bao, G.; Zhao, Y.; Teng, Z.; Yang, L.; and Zhang, Y. 2023. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *The Twelfth International Conference on Learning Representations*.

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. In *The Twelfth International Conference on Learning Representations*.

Black, S.; Leo, G.; Wang, P.; Leahy, C.; and Biderman, S. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Bracewell, R. N. 1986. *The Fourier Transform and Its Applications*. McGraw-Hill, 2nd edition.

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33: 17766–17778.

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.

Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Belanger, D.; Colwell, L.; et al. 2020. Masked language modeling for proteins via linearly scalable long-context transformers. *arXiv preprint arXiv:2006.03555*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Fan, W.; Zheng, S.; Yi, X.; Cao, W.; Fu, Y.; Bian, J.; and Liu, T.-Y. 2022. DEPTS: Deep expansion learning for periodic time series forecasting. *arXiv preprint arXiv:2203.07681*.

Fang, R.; and Xu, Y. 2024. Addressing Spectral Bias of Deep Neural Networks by Multi-Grade Deep Learning. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 114122–114146. Curran Associates, Inc.

Gehrmann, S.; Strobelt, H.; and Rush, A. M. 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.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*.

Guo, J.; Wang, N.; Qi, L.; and Shi, Y. 2023. Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24132–24141.

Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Hu, X.; Chen, P.-Y.; and Ho, T.-Y. 2023. Radar: Robust ai-text detection via adversarial learning. *arXiv preprint arXiv:2307.03838*.

Khan, S. H.; Hayat, M.; and Porikli, F. 2019. Regularization of deep neural networks with spectral dropout. *Neural Networks*, 110: 82–90.

Kumar, S.; Garg, S.; Sengupta, S.; Ghosal, T.; and Ekbal, A. 2025. MixRevDetect: Towards Detecting AI-Generated Content in Hybrid Peer Reviews. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 944–953. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-190-2.

Kushnareva, L.; Cherniavskii, D.; Mikhailov, V.; Artemova, E.; Barannikov, S.; Bernstein, A.; Piontkovskaya, I.; Piontkovski, D.; and Burnaev, E. 2021. Artificial Text Detection via Examining the Topology of Attention Maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 635–649.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontanon, S. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Li, R.; Hao, W.; Zhao, W.; Yang, J.; and Mao, C. 2025a. Learning to Rewrite: Generalized LLM-Generated Text Detection. arXiv:2408.04237.

Li, Y.; Li, Q.; Cui, L.; Bi, W.; Wang, Z.; Wang, L.; Yang, L.; Shi, S.; and Zhang, Y. 2024. MAGE: Machine-generated Text Detection in the Wild. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), 36–53. Bangkok, Thailand: Association for Computational Linguistics.

Li, Y.; Zhang, Z.; Li, C.; Shen, C.; and Liu, X. 2025b. Iron Sharpens Iron: Defending Against Attacks in Machine-Generated Text Detection with Adversarial Training. arXiv:2502.12734.

Liu, S.; Liu, X.; Wang, Y.; Cheng, Z.; Li, C.; Zhang, Z.; Lan, Y.; and Shen, C. 2024a. Does detectgpt fully utilize perturbation? bridging selective perturbation to fine-tuned contrastive learning detector would be better. *arXiv preprint arXiv:2402.00263*.

Liu, S.; Liu, X.; Wang, Y.; Cheng, Z.; Li, C.; Zhang, Z.; Lan, Y.; and Shen, C. 2024b. Does DetectGPT Fully Utilize Perturbation? Bridging Selective Perturbation to Fine-tuned Contrastive Learning Detector would be Better. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1874–1889. Bangkok, Thailand: Association for Computational Linguistics.

Liu, X.; Zhang, Z.; Wang, Y.; Pu, H.; Lan, Y.; and Shen, C. 2023. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16167–16188.

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 preprint arXiv:1907.11692*.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *ICML 2023*.

OpenAI. 2025. ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/. Accessed: 2025-11-13.

Shi, Y.; Sheng, Q.; Cao, J.; Mi, H.; Hu, B.; and Wang, D. 2024. Ten Words Only Still Help: Improving Black-Box AI-Generated Text Detection via Proxy-Guided Efficient Re-Sampling. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, 494–502. International Joint Conferences on Artificial Intelligence Organization.

Shum, K.; Diao, S.; and Zhang, T. 2023. Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data. *arXiv preprint arXiv:2302.12822*.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Sun, P.; Zhu, Y.; Zhang, Y.; Yan, X.; Wang, Z.; and Ji, X. 2024. Unleashing the Potential of Large Language Models through Spectral Modulation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3892–3911. Miami, Florida, USA: Association for Computational Linguistics.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33: 11839–11852.

Tamkin, A.; Jurafsky, D.; and Goodman, N. 2020a. Language through a prism: A spectral approach for multiscale language representations. *Advances in Neural Information Processing Systems*, 33: 5492–5504.

Tamkin, A.; Jurafsky, D.; and Goodman, N. 2020b. Language Through a Prism: A Spectral Approach for Multiscale Language Representations. arXiv:2011.04823.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Verma, V.; Fleisig, E.; Tomlin, N.; and Klein, D. 2023. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. *arXiv preprint arXiv:2305.15047*.

Wang, P.; Li, L.; Ren, K.; Jiang, B.; Zhang, D.; and Qiu, X. 2023a. SeqXGPT: Sentence-Level AI-Generated Text Detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wang, Y.; Feng, S.; Hou, A.; Pu, X.; Shen, C.; Liu, X.; Tsvetkov, Y.; and He, T. 2024. Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2894–2925. Bangkok, Thailand: Association for Computational Linguistics.

Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Whitehouse, C.; Afzal, O. M.; Mahmoud, T.; Sasaki, T.; et al. 2023b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.

Wu, J.; Zhan, R.; Wong, D. F.; Yang, S.; Yang, X.; Yuan, Y.; and Chao, L. S. 2024. DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 100369–100401. Curran Associates, Inc.

Yi, K.; Zhang, Q.; Fan, W.; Wang, S.; Wang, P.; He, H.; An, N.; Lian, D.; Cao, L.; and Niu, Z. 2023. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36: 76656–76679.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.