VERITAS:Veracity-Enhanced Robust Identification of LLM-generated Text Against Style-shifts

Xiaoquan Yi[1], **Haozhao Wang**[1], **Yichen Li**[1] **Wenchao Xu**[2],
Jinghua Zhang[1]*, **Yuhua Li**[1], **Imran Razzak**[3], **Rui Zhang**[1]†, **Ruixuan Li**[1]
[1]School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2]School of Division of Integrative Systems and Design
Hong Kong University of Science and Technology, Hong Kong, China.
[3]Mohamed bin Zayed University of Artificial Intelligence
{yixiaoquan,hz_wang}@hust.edu.cn, u201211756@alumni.hust.edu.cn, rayteam@yeah.net

Abstract

With the rising prominence of large language models (LLMs), developing technologies to identify LLM-generated text has become increasingly critical. However, existing technologies depend on static linguistic features, which can be evaded as advanced models increasingly mimic a wide range of writing styles. This study reveals two crucial limitations in existing detection systems. First, state-of-the-art detectors experience a substantial accuracy decline of up to 16.45% when exposed to style-based adversarial attacks generated by LLMs. Furthermore, despite exhibiting remarkable zero-shot capabilities, general-purpose LLMs underperform specialized, robustness-tuned detectors in identifying adversarially manipulated text. To address these limitations, we propose a novel style-agnostic detection framework named VERITAS that enhances detection accuracy and robustness by prioritizing content-driven features over stylistic attributes. Our approach integrates a style-invariant training paradigm to disentangle content semantics from stylistic variations. We leverage adversarially enriched datasets constructed using LLMs fine-tuned for diverse style-based attacks. Furthermore, we utilize advanced representation learning techniques to extract content-centric features, emphasizing semantic coherence, logical consistency, and factual alignment. Experimental results across multiple datasets and detection models validate the effectiveness of our framework, showing improvements in detection accuracy and robustness against diverse adversarial manipulations.

1 Introduction

Large language models (LLMs) have recently demonstrated remarkable capabilities in generating textual data [1, 2], and have garnered widespread attention. However, malicious actors abuse LLMs to generate various-style text, creating misleading public-opinion content such as fake news or academic papers with false data to distort the truth [3]. It is an extremely pressing matter to develop detection technologies that remain unaffected by text styles and can precisely identify the text's origin.

Many state-of-the-art detection methods heavily rely on stylistic attributes [3, 4, 5], i.e., lexical choice, sentence structure, and syntactic patterns. Although effective under normal conditions, these features are highly susceptible to adversarial style-based attacks, where generated text mimics human writing styles. This reliance undermines the robustness of these detectors, as adversaries can easily manipulate

*Jinghua Zhang is corresponding authors.
†Homepage: https://www.ruizhang.info/

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

style while preserving semantic coherence, even for multi-modal contents [6]. These limitations highlight the pressing need for detection frameworks that are less reliant on stylistic attributes, incorporate content-focused strategies, and are rigorously evaluated against a comprehensive set of adversarial attacks.

In this work, we aim to develop a technique independent of stylistic attributes and unaffected by text styles, enabling precise identification of the text's origin. Specifically, we identify two critical limitations: *state-of-the-art detectors for generated text are significantly compromised by LLM-induced stylistic variations*, and *general-purpose LLMs exhibit markedly inferior performance in detecting adversarially manipulated text*. Motivated by two critical limitations, we present a novel style-agnostic detection framework named `VERITAS`. We incorporate a style-invariant training paradigm that involves disentangling content features from stylistic attributes during training using adversarial learning techniques. The model is explicitly trained to disregard stylistic variations by introducing adversarially perturbed examples that mimic diverse human writing styles. We generate a comprehensive adversarial dataset leveraging LLMs fine-tuned for style attacks. This dataset includes diverse stylistic manipulations across genres and domains, exposing the model to various adversarial scenarios. Experimental results across various datasets and models demonstrate that our method achieves substantial improvements in detection accuracy. Our contributions are as follows:

- We observe that current detectors for LLM-generated text exhibit considerable limitations when encountering texts with styles purposefully altered by other LLMs. To our knowledge, *this work is the first to systematically investigate the impact of these LLM-driven stylistic alterations on the performance of text detection systems*.
- We introduce a novel training framework that enhances the resilience of text generation detectors by learning style-invariant features. To the best of our knowledge, *this is the first approach to forgo stylistic artifacts in favor of content-driven analysis*.
- We construct an adversarially enriched dataset by leveraging LLMs fine-tuned for style-based attacks. *Extensive experiments on advanced and commercial LLMs (LLAMA-13B, GPT-4, Qwen-2, etc.) show that `VERITAS` outperforms state-of-the-art detection methods by up to* 6.11% *in accuracy.*

2 Related work

In this section, we discuss two critical dimensions of LLMs-generated text analysis, i.e., representative detection approaches and adversarial attacks.

Universal Detection Frameworks for LLM-Generated Content. Numerous methodologies leverage the sophisticated mechanisms of LLMs, encompassing intermediate layer outputs as well as weight parameters, to differentiate between texts authored by humans and those generated by LLMs [7, 8, 9]. These methodologies introduce subtle text modifications to monitor changes in log probabilities, where a notable decrease typically indicates LLM-generated text, while an increase or minor fluctuations suggest human authorship. To detect text generated by smaller models, statistical analyses are employed to examine features,i.e., word choice, sentence structure, and stylistic elements [10, 11, 7]. These features are compared against known human and LLMs-generated patterns to identify discrepancies indicative of non-human authorship. Despite their innovations, these methods largely depend on stylistic cues, which are easily manipulated by adversaries using style-based transformations [12, 13, 14]. This reliance exposes a critical vulnerability, as demonstrated by accuracy degradation under adversarial conditions.

Adversarial Attacks in Text Generation. Adversarial attacks have now emerged as a formidable threat to the detection of generated text [15, 16]. The existing adversarial sample attack techniques are characterized by their high level of stealth and aggressiveness, enabling them to effectively evade the current detection methods [17, 18]. Many approaches subtly perturb parts of the text or embed feature-specific text snippets [19, 20, 21, 22], causing NLP models to produce targeted incorrect outputs while preserving semantic and syntactic integrity. Although adversarial examples in other domains, i.e., image recognition [23, 24, 25] and fake news detection [26, 27], have been extensively studied, similar efforts in the context of LLMs-generated text are limited. Existing detection frameworks are not specifically designed to address these sophisticated style transformations, leading to accuracy degradation in adversarial scenarios.

3 Problem Definition

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent a dataset where x_i is the input text sample, and $y_i \in \{0, 1\}$ is the corresponding label, with $y_i = 1$ indicating LLMs-generated text and $y_i = 0$ for human-written text. The goal is to train a detector $f_\theta(x) : \mathcal{X} \to \{0, 1\}$ parameterized by θ, which can accurately classify x_i while maintaining robustness against adversarial manipulations.

An adversarial example x_i' is defined as a perturbed version of x_i generated by an adversary $\mathcal{A}$, and ϕ represents adversarial parameters.:

$$x_i' = \mathcal{A}(x_i; \phi) \tag{1}$$

The $\mathcal{X}'$ denotes the set of adversarially perturbed samples. . The detector $f_\theta(x)$ should satisfy the following robustness criterion:

$$f_\theta(x_i) = f_\theta(x_i'), \quad \forall x_i \in \mathcal{D},\, x_i' \in \mathcal{X}', \tag{2}$$

4 Motivation

In this section, we utilize LLMs to synthesize text across various writing styles to simulate realistic adversarial scenarios. We conduct a preliminary analysis to systematically evaluate the robustness and efficacy of state-of-the-art detection systems when confronted with this stylistically altered content.

4.1 Simulating Style-Based Adversarial Attacks

The advanced natural language capabilities of LLMs enable users to seamlessly transform text styles via tailored prompts, posing a significant challenge to the robustness of current detection systems. In this study, we investigate direct style-based attacks by leveraging distinctive writing paradigms. We employ highly recognizable styles, ranging from the whimsical narratives of Andersen's fairy tales to the rigorous scientific prose of prestigious academic journals like *Nature* and *Science*. Because these writing styles are marked by pronounced lexical choices, narrative elements, and tonal qualities, they serve as highly effective vehicles for adversarial manipulation. For instance, a standard LLM-generated narrative can be adversarially rewritten to incorporate the moral undertones and vivid imagery typical of classical fairy tales. Our prompt template for these stylistic transformations is structured as follows:

> **Prompt:** Transform the following text to match the stylistic conventions of [publisher/book]: *[input text]*

Specifically, we map target domains to their respective stylistic proxies: Andersen's fairy tales for narrative texts, CNN's journalistic tone for political discourse, and the academic rigor of *Nature* and *Science* for scientific literature. These stylistically perturbed test samples are subsequently utilized to systematically benchmark the robustness of advanced detectors against style-driven evasion tactics.

4.2 Performance Degradation Under Style-Shifts

As shown in Table 1, the original accuracy (O) reflects the detection methods' performance on unaltered texts, while the adversarial accuracy (A) measures their robustness against style-based adversarial attacks, with performance drops indicated by (↓). Among the evaluated methods, DetectGPT [28] demonstrates the highest overall performance, particularly excelling on the Story (O: 78.24%, A: 70.13%) and PolitiFact (O: 75.91%, A: 71.96%) datasets. Nevertheless, its accuracy declines significantly on the Science dataset (O: 69.72%, A: 60.43%), exposing the heightened difficulty of defending against adversarial manipulations in structured, technical domains. Conversely, GLTR [29] and

Method	Story		PolitiFact		Science	
	O	**A (↓)**	**O**	**A (↓)**	**O**	**A (↓)**
DetectGPT	78.24	70.13	75.91	71.96	69.72	60.43
GLTR	63.24	54.88	59.24	52.32	60.66	53.73
LLMDet	72.35	60.69	70.08	62.57	60.52	57.63
LLaMA-7B	69.34	61.45	73.45	60.59	68.59	52.14
Neo-2.7B	62.25	50.82	65.39	52.75	60.97	48.33

Table 1: The performance comparison of different methods across Story, PolitiFact, and Scientists datasets. **O** represents the original accuracy, and **A** (↓) represents the adversarial accuracy.

LLMDet [3] exhibit weaker baseline and adversarial performance. GLTR is especially vulnerable, as evidenced by its accuracy drop on PolitiFact (O: 59.24%, A: 52.32%), whereas LLMDet maintains a moderate advantage (O: 70.08%, A: 62.57%). Furthermore, general-purpose LLMs such as LLaMA-7B and Neo-2.7B struggle across the board. Neo-2.7B experiences the most drastic degradation on the Science dataset (O: 62.25%, A: 48.33%), underscoring the limitations of smaller foundational models in zero-shot detection tasks. Crucially, these empirical results reveal two fundamental limitations in current detection paradigms. First, state-of-the-art detectors are highly susceptible to LLM-driven stylistic variations, suffering substantial performance drops of up to 16.45% when evaluated on stylistically altered test sets. Second, despite their remarkable zero-shot capabilities in general tasks, general-purpose LLMs perform markedly worse in detecting adversarially manipulated text compared to specialized detection systems or pre-trained language models fine-tuned for robustness.

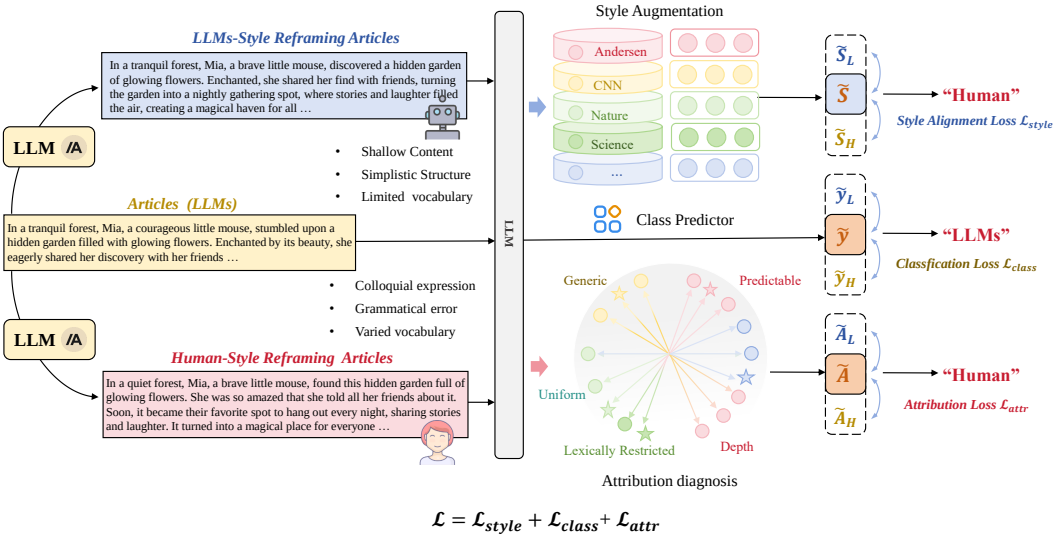


Figure 1: Overview of the VERITAS framework. It consists of three main components, e.g., style alignment loss, class detection Loss, veracity attribution loss. VERITAS is designed to enhance the robustness of text generation detectors against style-based adversarial attacks by focusing on content-driven features and minimizing reliance on stylistic cues.

5 The VERITAS Framework

In this section, we detail the proposed methodology for detecting LLM-generated content. VERITAS is inspired by SheepDog [27], adapting it from fake-news detection to LLM-generated text detection. By leveraging an LLM to transform human-written texts into various styles [30], we construct a diverse training dataset. The model's feature extraction integrates contextual embeddings and statistical metrics. Subsequently, the training process is optimized through a joint objective comprising three loss functions: style alignment, classification, and veracity attribution supervision.

5.1 Adversarial Style Augmentation

To simulate the diverse stylistic variations that LLM-generated content may exhibit in real-world scenarios, we adopt a style transformation-based data augmentation strategy. Specifically, each human-written text $p_{\text{human}} \in D_{\text{human}}$ is converted into multiple stylistic variants using an auxiliary LLM, denoted as M_{LLM}. Let N_s represent the total number of stylistic transformations, and s_i denote a specific target style (e.g., formal, narrative, or scientific). The transformation process is formalized as follows:

$$p_{\text{gen},i} = M_{\text{LLM}}(p_{\text{human}}, \text{style} = s_i), \quad i = 1, \ldots, N_s \tag{3}$$

The augmented training dataset D_{train} is subsequently constructed by integrating these stylistic variants with the original human-written texts, thereby ensuring a comprehensive and diverse sample set for robust model training:

$$D_{\text{train}} = D_{\text{human}} \cup \{p_{\text{gen},i} \mid i = 1, 2, \ldots, N_s\} \tag{4}$$

5.2 Style-Invariant Representation Learning

To enhance the model's robustness against stylistic shifts, we introduce a style alignment objective that enforces consistent predictive distributions across all N_s stylistic variants $\{p_{\text{gen},i}\}$. We achieve this by passing the hidden representation $h_{p_{\text{gen},i}}$ of each variant through a Multi-Layer Perceptron (MLP) classifier M to yield the output probability distribution:

$$y_i = \text{Softmax}(M(h_{p_{\text{gen},i}})) \tag{5}$$

The style alignment loss L_{style} is formulated as the average pairwise Kullback-Leibler (KL) divergence between the predictive distributions of all variant pairs:

$$L_{\text{style}} = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \text{KL}(y_i \,\|\, y_j) \tag{6}$$

where $\text{KL}(y_i \,\|\, y_j)$ denotes the KL divergence from y_j to y_i [31]. This mechanism compels the classifier to produce consistent outputs independent of superficial stylistic features.

5.3 Provenance Classification Objective

The classification objective ensures the model accurately predicts the fundamental source of the input text. For a given input sequence p with its ground-truth label y and predicted probability distribution $\hat{y}$, the cross-entropy loss is computed as:

$$L_{\text{class}} = -\sum_{k=1}^{K} y_k \log \hat{y}_k \tag{7}$$

where $K = 2$ corresponds to the binary classification task (human-written versus LLM-generated).

5.4 Content-Driven Attribution Supervision

Veracity attributions are generated by querying an LLM to identify structural and logical patterns intrinsically characteristic of machine-generated text. We utilize the following prompt template for this inquiry:

> **Input Texts**: *[LLMs-generated texts]*
> **Question**: Which of the following problems does this text have? Single language style, too structured logical structure, Lack of background knowledge and personal experience, Repetitive or patterned expression, Data biases, and errors. If multiple options are applicable, provide a comma-separated list ordered from most to least relevant. Answer "No" if none of the options apply.

For a predefined veracity attribution space $C = \{c_1, c_2, \ldots, c_m\}$, the k-th element of the attribution vector $\mathbf{A}_i$ for a generated text $p_{\text{gen},i}$ is assigned as:

$$A_{i,k} = \begin{cases} 1, & \text{if } p_{\text{gen},i} \text{ satisfies } c_k, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We then define an attribution loss based on the Binary Cross-Entropy (BCE) between the predicted attribution vector $\hat{\mathbf{A}}_i$ and the ground-truth vector $\mathbf{A}_i$, averaged over N_a samples:

$$L_{\text{attr}} = \frac{1}{N_a} \sum_{i=1}^{N_a} \text{BCE}(\mathbf{A}_i, \hat{\mathbf{A}}_i) \tag{9}$$

5.5 Overall Optimization Objective

The final training objective seamlessly integrates the style alignment loss, classification loss, and veracity attribution supervision loss into a unified framework:

$$L = L_{\text{style}} + L_{\text{class}} + L_{\text{attr}} \tag{10}$$

Scenarios	Description
A	Introduce minor errors and colloquial expressions
B	Increase personalization and subjectivity
C	Mix topics and cite diverse resources
D	Use specialized terminology and cultural references

Table 2: The strategies for text stylization based on different writing styles.

Table 3: VERITAS demonstrates superior performance compared to competitive baselines across four adversarial test scenarios under LLM-powered style attacks, evaluated in terms of F1 Score (%). The strategies for text stylization are in Table 2.

Method	Story				PolitiFact			
	A	B	C	D	A	B	C	D
GLTR (ACL 2019)	52.09 ± 1.43	52.98 ± 0.69	50.37 ± 1.55	49.40 ± 1.30	56.17 ± 1.20	54.05 ± 1.86	53.41 ± 0.98	52.85 ± 0.70
Detectllm (Emnlp 2023)	65.57 ± 0.35	62.81 ± 0.61	65.26 ± 0.63	62.27 ± 1.01	64.60 ± 1.64	66.91 ± 1.17	63.40 ± 1.28	61.04 ± 1.00
LLMDet (Emnlp 2023)	61.94 ± 0.28	60.31 ± 0.76	58.83 ± 1.84	59.91 ± 1.27	61.81 ± 1.64	63.82 ± 0.90	59.38 ± 0.44	58.14 ± 1.93
DetectGPT (ICML 2023)	68.92± 5.67	67.85± 4.42	62.81± 0.82	66.35± 2.19	65.83± 1.58	68.74± 2.71	72.49± 0.38	65.52± 1.58
SeqXGPT (Emnlp 2023)	59.63 ± 0.61	57.41 ± 1.47	59.81 ± 1.01	64.75 ± 0.97	69.58 ± 0.61	61.98 ± 0.49	55.55 ± 1.07	57.90 ± 1.66
COCO (Emnlp 2023)	70.28 ± 0.86	69.11 ± 0.52	69.90 ± 0.96	68.84 ± 1.26	71.75 ± 0.48	71.86 ± 1.33	68.39 ± 1.24	69.55 ± 1.19
BERT-finetuned (ACL 2024)	51.25 ± 0.49	54.28 ± 1.67	55.75 ± 0.93	52.22 ± 2.24	55.19 ± 1.01	56.58 ± 1.28	57.43 ± 1.17	55.34 ± 1.92
RoBERTa-finetuned (ACL 2024)	65.28 ± 1.35	66.69 ± 2.51	67.47 ± 1.95	79.00 ± 1.88	70.32 ± 1.28	77.00 ± 2.35	68.37 ± 1.11	66.52 ± 1.47
T5-Sentinel (Emnlp 2024)	62.93 ± 1.49	71.84 ± 2.03	72.37 ± 1.57	71.76 ± 1.02	72.39 ± 1.80	78.23 ± 2.41	73.07 ± 2.41	70.59 ± 1.26
OUTFOX (AAAI 2024)	78.03 ± 0.87	72.91 ± 1.26	70.99 ± 1.23	72.58 ± 2.01	80.55 ± 1.02	79.84 ± 1.58	70.05 ± 1.84	76.45 ± 1.10
GECScore (ACL 2025)	70.13 ± 0.48	77.44 ± 0.58	76.99 ± 1.16	67.04 ± 1.09	63.34 ± 0.95	83.29 ± 1.53	67.60 ± 0.32	67.42 ± 1.56
GPT-2 (2019)	55.20± 2.42	56.96± 2.45	50.51± 1.01	49.40± 2.06	50.86± 1.93	52.03± 5.52	55.32± 3.47	55.89± 5.18
Neo-2.7B (2020)	61.46 ± 0.95	55.77 ± 1.84	58.82 ± 0.40	69.48 ± 1.27	71.01 ± 1.61	62.49 ± 2.08	56.99 ± 1.80	59.13 ± 0.82
OPT-2.7B (2022)	54.30± 0.99	57.29± 6.56	52.51± 5.36	51.27± 5.02	52.85± 3.72	51.53± 1.72	49.24± 0.58	48.68± 0.50
LLaMA-7B (2023)	57.56± 2.75	51.93± 2.63	51.45± 6.89	52.26± 0.71	53.12± 3.82	56.10± 3.12	56.64± 0.94	54.39± 4.17
LLaMA2-13B (2023)	62.60 ± 1.38	60.61 ± 0.60	61.69 ± 1.04	57.14 ± 0.22	61.64 ± 1.91	63.60 ± 0.23	59.07 ± 1.00	59.92 ± 1.14
GPT-3.5-turbo (2023)	72.92± 0.58	76.41± 0.75	68.68± 4.85	69.85± 2.63	69.77± 0.05	71.08± 0.40	65.50± 0.87	68.62± 2.53
GPT-NeoX (2024)	71.83 ± 1.46	71.64 ± 1.17	69.52 ± 0.33	69.30 ± 0.44	75.90 ± 1.96	74.48 ± 1.78	67.73 ± 2.00	67.90 ± 1.37
GPT-4 (2024)	73.86± 2.06	70.53± 2.32	75.75± 1.46	78.55± 0.39	77.34± 2.33	79.88± 2.20	68.78± 0.18	71.80± 1.65
Gemma (2025)	74.18± 8.32	74.06± 1.17	66.78± 0.74	69.60± 5.88	69.62± 8.61	73.74± 1.56	69.28± 9.08	72.92± 5.77
Qwen-2 (2025)	76.05± 2.12	75.95± 5.69	70.52± 0.08	74.87± 1.85	78.08± 5.38	79.66± 4.08	65.04± 2.94	72.54± 0.88
Deepseek-R1 (2025)	73.74± 0.77	73.08± 0.56	72.00± 3.74	67.77± 6.86	75.73± 3.66	82.83± 0.31	69.10± 1.24	75.95± 0.29
VERITAS	**79.56 ± 1.28**	**78.37 ± 1.99**	**77.60 ± 1.24**	**79.26 ± 1.08**	**81.82 ± 0.68**	**84.69 ± 1.35**	**79.18 ± 0.95**	**79.32 ± 1.18**

6 Experiments

In this section, we describe the experimental setup, including dataset details and implementation specifics of VERITAS. Eventually, we will assess the performance of our approach, including accuracy and F1 score, etc., on multiple datasets and compare these metrics against state-of-the-art algorithms.

6.1 Experiment setting

Datasets. Our experiment employs a meticulously curated collection of three datasets to appraise the capabilities of generated text detection across a spectrum of domains. The Story category features narrative-rich texts, including literary works from the Gutenberg dataset [32] and story-focused articles from the X-Sum dataset [33]. This category assesses the model's capability to comprehend long texts and generate coherent narratives. The PolitiFact category utilizes a combined dataset from the LIAR [34] and FakeNewsNet datasets [35], encompassing political statements with truthfulness ratings and additional social media context. This category assesses the model's performance in information accuracy and factual consistency. The science category encompasses scientific and domain-specific texts, such as medical records from the MedNLI dataset [36] or scientific literature from the Gutenberg dataset, evaluating the model's proficiency in handling specialized terminology, logical reasoning, and domain knowledge.

Metrics. To evaluate the detector's capability to distinguish between texts generated by LLMs and humans, especially under style-based attacks from LLMs-powered adversaries, we employ two primary performance metrics, e.g., Accuracy (A), and the F1 score (F1).

Baselines. We undertook a comparative analysis of our proposed methodology against several state-of-the-art approaches dedicated to detecting text generated by LLMs. DetectLLM [9] assesses text origin via log perplexity, indicating predictability. GLTR [29] combines statistical methods with visual analytics to highlight anomalous token probabilities, aiding in the identification of machine-generated text. DetectGPT [28] leverages the curvature properties of language model probability functions within a probabilistic framework to identify synthetic text. LLMDet [3] uses surrogate perplexity calculations tailored to each LLM, offering a model-agnostic solution for text provenance. SeqXGPT [4] represents sentences as waveforms, utilizing convolutional networks and self-attention mechanisms for sentence-level detection. GECScore [37] evaluates text similarity using a grammar error correction model, providing a robust metric for LLM-origin detection. The OUTFOX method [38] bolsters the robustness of detecting text generated by LLMs by implementing an iterative in-context learning framework. COCO [5] enhances detection through contrastive learning.

Table 4: Ablation Study of VERITAS Loss Components under Different attack Scenarios (F1 Score %). The strategies for text stylization are in Table 2.

Experiment Setting	Dataset	Attack Scenario A	Attack Scenario B	Attack Scenario C	Attack Scenario D
Baseline (L_{class} Only)	Story	64.10 ± 1.63	63.50 ± 1.86	67.00 ± 1.53	65.90 ± 1.71
	PolitiFact	66.80 ± 1.95	66.20 ± 2.01	69.56 ± 1.76	68.30 ± 1.83
	Science	62.04 ± 1.47	61.57 ± 1.59	64.87 ± 1.34	63.54 ± 1.24
$L_{class} + L_{style}$	Story	71.51 ± 1.33	73.22 ± 1.49	73.08 ± 1.27	72.39 ± 1.39
	PolitiFact	74.15 ± 1.63	76.59 ± 1.61	76.37 ± 1.42	75.47 ± 1.68
	Science	69.48 ± 1.16	71.84 ± 1.12	70.79 ± 1.53	69.80 ± 1.10
$L_{class} + L_{attr}$	Story	73.29 ± 1.58	72.37 ± 1.14	75.06 ± 1.70	74.34 ± 1.91
	PolitiFact	76.68 ± 1.25	75.51 ± 1.67	79.28 ± 1.49	78.33 ± 1.96
	Science	70.86 ± 1.13	69.89 ± 1.40	73.75 ± 1.62	72.20 ± 1.67
Full Model (VERITAS) ($L_{class} + L_{style} + L_{attr}$)	Story	78.45 ± 1.12	78.00 ± 1.91	79.50 ± 1.50	79.08 ± 1.34
	PolitiFact	80.69 ± 1.79	82.58 ± 1.35	81.88 ± 1.02	81.43 ± 1.08
	Science	84.92 ± 1.15	80.47 ± 1.62	78.57 ± 1.49	83.31 ± 1.54

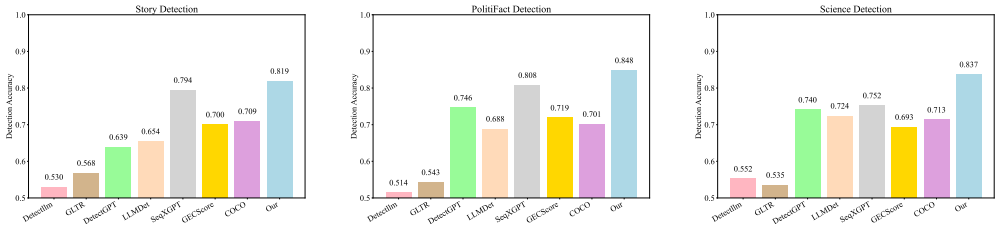


Figure 2: The visual data presented in the graphs clearly indicates that our methodology excels in detection accuracy across multiple categories,e.g., Story, PolitiFact, and Science.

T5-Sentinel [39] employs a supervised learning approach, reframing LLM-generated text detection as a token prediction task.

General-purpose LLMs. LLMs perform zero-shot veracity prediction, enabling the evaluation of truthfulness without requiring task-specific fine-tuning. We use some representative baseline LLMs for analysis: GPT-2 [40], OPT-2.7B [41], Neo-2.7B [42], LLaMA-7B [43], LLaMA-13B, GPT-NeoX [44], GPT-3.5-turbo, GPT-4, Gemma [45], Qwen-2 [46], Deepseek-R1 [47]. These models serve as benchmarks to assess the capabilities and limitations of LLMs in zero-shot detecting content-generated tasks.

6.2 Performance Evaluation

F1-Score.Table 3 illustrates the performance of different methods in addressing four distinct adversarial attack styles. The results clearly show that VERITAS consistently exhibits advantages across all test scenarios. In the Story scenario, VERITAS achieves F1 of 79.56%, 78.37%, 77.60%, and 79.26%, respectively, with improvements of over 6.69% compared to DetectGPT, demonstrating outstanding adversarial handling capabilities. In the PolitiFact scenario, VERITAS's performance is particularly remarkable, especially under colloquial adversarial attacks, where it achieves an F1 of 84.69%, surpassing COCO's 71.86% with a performance gain of nearly 12.83%. This result underscores VERITAS's excellent adaptability to complex politics-related text. This further highlights VERITAS's robust capability to recognize adversarial features in complex scientific texts.

Accuracy. As illustrated in Figure 2, our method demonstrates improvements across multiple categories, including Story, PolitiFact, and Science. Specifically, VERITAS achieves detection accuracies of 0.819, 0.848, and 0.837, respectively. When compared to other advanced methods, VERITAS consistently outperforms them by margins ranging from 6.5% to 33.9%. These results highlight VERITAS's superior capability in analyzing multi-dimensional textual features, effectively resisting adversarial attacks, and maintaining high precision. This notable

Method	Story	PolitiFact	Science
OPT-2.7B	49.74 ± 1.19	52.42 ± 2.15	55.77 ± 1.97
VERITAS-OPT-2.7B	**64.03 ± 2.55**	**58.92 ± 2.60**	**58.58 ± 5.90**
Neo-2.7B	59.95 ± 3.39	62.65 ± 3.95	60.43 ± 1.72
VERITAS-Neo-2.7B	**82.71 ± 2.04**	**80.32 ± 3.19**	**77.55 ± 1.64**
LLaMA2-13B	59.99 ± 2.45	60.17 ± 1.43	58.27 ± 1.28
VERITAS-LLaMA2-13B	**80.23 ± 2.69**	**84.19 ± 2.38**	**86.31 ± 1.45**
GPT-NeoX	61.29 ± 1.62	64.00 ± 2.93	67.38 ± 2.26
VERITAS-GPT-NeoX	**77.38 ± 1.66**	**83.20 ± 1.21**	**79.05 ± 2.27**
Qwen-2	66.43 ± 0.59	61.32 ± 0.61	67.49 ± 2.39
VERITAS-Qwen-2	**85.23 ± 6.44**	**87.24 ± 3.14**	**85.57 ± 0.39**

Table 5: On different LLM backbones, VERITAS demonstrates stable improvements on accuracy.

enhancement underscores VERITAS's reliability and versatility, positioning it as a leading solution for high-fidelity content discrimination. This can be used to check academic papers.

Different Backbones. The VERITAS method demonstrates performance improvements across multiple models and datasets. For instance, VERITAS-Neo-2.7B achieves the accuracy of 82.71% on the Story dataset, compared to 59.95% for the baseline, while VERITAS-LLaMA2-13B reaches 84.19% on PolitiFact, up from 60.17%. Using the Qwen-2 backbone, the accuracy improved from 66.43% to 85.23% for Story datasets. These results highlight VERITAS's ability to enhance detection accuracy by leveraging style-agnostic feature extraction and adversarial data augmentation, which ensures robustness against style-based adversarial attacks. Its style-agnostic feature extraction forces the model to learn the intrinsic, content-centric artifacts of LLM generation, thereby mitigating the risk of overfitting to superficial and easily manipulated stylistic cues. The adversarial data augmentation proactively hardens the detector by exposing it to a diverse array of synthesized edge cases that mimic sophisticated evasion attempts.

Ablation study. As shown in Figure 4, comparing ($L_{class} + L_{style}$) with Baseline, there's a clear increase in F1 scores across all datasets and all attack scenarios. On the Story dataset under scenario A, F1 improves from 64.10% to 71.51%. This strongly validates the effectiveness of the style alignment loss (L_{style}). By enforcing consistent predictions for stylistically varied content, it demonstrably enhances the detector's robustness against various rewriting techniques (colloquialism, subjectivity, topic mixing, specialized terminology). The result demonstrates a synergistic effect between L_{style} and L_{attr}. While each auxiliary loss provides individual performance gains, their combination within the full VERITAS framework achieves the optimal outcome, confirming that both components are integral to its state-of-the-art performance. Addressing both style invariance and content features simultaneously proves more effective than focusing on either aspect alone.

Table 6: Cross-Domain evaluation on the science datasets. The best number is highlighted in bold. Our approach consistently outperforms other methods, achieving the highest F1 in each dataset.

Method	Science			
	A	B	C	D
GLTR (ACL 2019)	48.57 ± 1.06	51.35 ± 1.81	50.83 ± 1.65	53.20 ± 2.41
Detectllm (EMNLP 2023)	59.62 ± 1.91	63.16 ± 2.34	60.22 ± 1.14	64.14 ± 3.16
LLMDet (EMNLP 2023)	57.86 ± 0.92	54.54 ± 1.41	58.06 ± 0.03	58.95 ± 1.57
DetectGPT (ICML 2023)	54.46 ± 1.12	59.72 ± 0.71	54.34 ± 2.68	56.29 ± 0.66
SeqXGPT (EMNLP 2023)	49.99 ± 1.81	55.56 ± 4.24	56.64 ± 1.16	53.85 ± 2.47
COCO (EMNLP 2023)	61.39 ± 0.23	66.14 ± 0.38	68.66 ± 2.18	65.10 ± 4.11
BERT-finetuned	54.39 ± 1.23	56.86 ± 0.77	57.08 ± 1.90	56.69 ± 1.08
RoBERTa-finetuned	67.56 ± 0.18	70.52 ± 0.23	73.27 ± 2.52	70.43 ± 1.06
T5-Sentinel	67.53 ± 1.20	71.03 ± 1.10	72.55 ± 0.15	71.56 ± 0.99
OUTFOX (AAAI 2024)	71.21 ± 0.75	71.28 ± 2.12	75.19 ± 1.03	73.93 ± 1.26
GECScore (ACL 2025)	67.93 ± 0.52	72.05 ± 2.06	68.91 ± 0.28	68.93 ± 0.84
GPT-2 (2019)	50.88 ± 0.75	52.06 ± 3.53	54.72 ± 4.46	50.23 ± 2.20
Neo-2.7B (2020)	59.28 ± 0.72	57.75 ± 1.04	58.19 ± 0.32	59.32 ± 0.90
OPT-2.7B (2022)	45.73 ± 2.83	54.82 ± 2.61	50.83 ± 2.33	54.64 ± 1.72
LLaMA-7B (2023)	51.52 ± 3.60	51.21 ± 2.89	59.45 ± 2.57	52.10 ± 3.25
LLaMA2-13B (2023)	57.85 ± 0.29	59.41 ± 0.31	60.22 ± 0.29	58.50 ± 0.98
GPT-3.5-turbo	68.98 ± 1.79	74.82 ± 2.75	71.90 ± 0.08	71.15 ± 3.92
GPT-NeoX (2024)	71.51 ± 1.33	74.16 ± 1.30	71.30 ± 1.95	68.62 ± 1.84
GPT-4 (2024)	75.22 ± 3.82	74.12 ± 3.26	79.15 ± 2.81	71.24 ± 4.44
Gemma (2025)	66.80 ± 3.91	75.65 ± 4.88	73.77 ± 4.31	74.34 ± 0.07
Qwen-2 (2025)	73.49 ± 3.53	74.72 ± 0.59	74.83 ± 1.97	74.88 ± 1.24
Deepseek-R1 (2025)	73.57 ± 0.54	74.11 ± 6.20	73.57 ± 0.44	70.28 ± 4.72
VERITAS	**82.26** ± 1.40	**83.30** ± 2.07	**81.14** ± 4.36	**80.16** ± 2.54

Cross-Domain Evaluation. To rigorously assess the generalization capabilities of VERITAS under severe distributional shifts, we conducted a cross-domain evaluation where VERITAS was trained exclusively on the narrative-centric Story dataset and evaluated on the stylistically distinct Science dataset. As detailed in Table 6, VERITAS demonstrates exceptional out-of-distribution robustness, consistently achieving the highest F1 scores (80.16%–83.30%) across all adversarial scenarios. This performance significantly surpasses both fine-tuned baselines (e.g., RoBERTa-finetuned) and powerful general-purpose LLMs such as GPT-4 and Qwen-2, which exhibit notable performance degradation when transferring learned patterns from narrative to technical discourse. These results validate that VERITAS successfully disentangles semantic content from superficial stylistic attributes, enabling it to identify intrinsic LLM-generated artifacts that persist regardless of the substantial domain gap between training and inference.

Table 7: Across different sets of reframing prompts, VERITAS demonstrates stable and significant improvements over the most competitive baseline on accuracy.

Method	**Story**	**PolitiFact**	**Science**
Baseline (Best)	79.14 ± 1.83	83.20 ± 1.48	75.37 ± 3.18
VERITAS	**82.24 ± 3.09**	**80.99 ± 2.17**	**83.50 ± 1.06**
w/P_1	79.95 ± 1.51	82.08 ± 2.13	83.30 ± 2.32
w/P_2	78.95 ± 2.16	83.55 ± 1.54	81.66 ± 3.51
w/P_3	81.95 ± 2.79	79.98 ± 1.29	81.83 ± 1.66
w/P_4	80.21 ± 2.28	83.71 ± 1.31	81.97 ± 3.14

Different Prompts. The VERITAS method demonstrates superior precision across multiple datasets (Story, PolitiFact, Science) compared to the best baseline, as evidenced by the experimental results in Table 7. VERITAS not only delivers a notable precision of 82.24% on the Story dataset, eclipsing the baseline's 79.14%, but also demonstrates a commanding lead on the more challenging Science dataset, achieving 83.50% precision against a mere 75.37% for the baseline. The underlying drivers of this performance leap are twofold. Its style-agnostic feature extraction paradigm allows the model to transcend superficial stylistic fingerprints, which often confound conventional detectors, and instead learn the fundamental, intrinsic signatures of synthetic text. Our adversarial data augmentation strategy proactively immunizes the model against evasive maneuvers by training it on a curated corpus of hard-to-classify, style-manipulated examples. This synergy culminates in exceptional robustness against a broad spectrum of style-based adversarial attacks. Crucially, VERITAS's high precision is consistently maintained across varied generative prompts (e.g., $w/P_1, w/P_2, w/P_3$), affirming its operational reliability and positioning it as a highly effective solution for identifying LLM-generated content. For instance, on the Story dataset, VERITAS achieves a precision of 82.24, significantly outperforming the baseline's 79.14, while on the Science dataset, VERITAS attains a precision of 83.50, compared to the baseline's 75.37. This improvement is attributed to VERITAS's style-agnostic feature extraction and adversarial data augmentation, which enhances its robustness against style-based adversarial attacks.VERITAS maintains high precision across various prompts and configurations (e.g., w/P_1, w/P_2, w/P_3), showcasing its adaptability and reliability in detecting LLM-generated content.

Computational Overhead and Deployment Efficiency. While the proposed framework relies on an auxiliary LLM for generating multiple stylistic variants and veracity attribution pseudo-labels, which inevitably introduces offline computational costs, this data preparation phase is highly scalable and represents merely a one-time overhead prior to training. By leveraging modern high-throughput inference engines and quantization techniques, specifically an efficiently deployed open-weight LLM such as LLaMA-3-8B optimized with vLLM and 4-bit AWQ on a single NVIDIA A6000 GPU, our offline data construction achieves a robust throughput of ~ 65 samples/sec.Crucially, this architectural design ensures that the heavy generative workloads are entirely decoupled from the deployment phase. During online inference, the auxiliary LLM is completely discarded. VERITAS relies solely on the fine-tuned lightweight backbone to extract style-agnostic features and compute the final classification probabilities. Consequently, VERITAS evaluates input text at an exceptional speed of ~ 200 samples/sec. It introduces strictly zero additional computational or memory overhead during real-time deployment compared to standard lightweight baselines, making the framework highly practical for large-scale, latency-sensitive applications.

7 Conclusion

In this work, we address critical shortcomings in detecting LLM-generated text, particularly against style-based adversarial attacks powered by advanced language models. Existing detectors exhibit significant performance degradation when confronted with stylistic manipulations, highlighting the fundamental limitations of their reliance on surface-level stylistic cues. To overcome these challenges, we propose VERITAS, a robust detection framework that prioritizes content-driven features and employs a style-agnostic training paradigm. By leveraging adversarially enriched datasets and advanced representation learning techniques, our approach effectively disentangles core semantic content from superficial stylistic variations. This ensures enhanced robustness against diverse adversarial attacks while maintaining high detection accuracy across multiple domains.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China under Grant 2024YFC3307900; the National Natural Science Foundation of China under Grants 62376103, 62302184, 62436003, 62472332, 62672202 and 62672205; the Major Science and Technology Project of Hubei Province under Grants 2025BAB011, 2024BAA008, and 2024BAB039; the Hubei Science and Technology Talent Service Project under Grant 2024DJC078; the Science and Technology Innovation Talent Program under grant 2023CB005; and Wuhan Dameng Database Co., Ltd. The authors declare no competing interests beyond the funding disclosed above.

References

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20(12), May 2026.

[2] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[3] Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*, 2023.

[4] Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore, December 2023. Association for Computational Linguistics.

[5] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 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*, pages 16167–16188, 2023.

[6] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3833–3843, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*, 2023.

[8] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

[9] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*, 2023.

[10] Kaito Taguchi, Yujie Gu, and Kouichi Sakurai. The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*, 2024.

[11] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

[12] Hetvi Waghela, Jaydip Sen, Sneha Rakshit, and Subhasis Dasgupta. Adversarial text generation with dynamic contextual perturbation. *Proc. of IEEE CALCON*, 2024.

[13] Haoran Fu, Chundong Wang, Jiaqi Sun, Yumeng Zhao, Hao Lin, Junqing Sun, and Baixue Zhang. Wordillusion: An adversarial text generation algorithm based on human cognitive system. *Cognitive Systems Research*, 83:101179, 2024.

[14] Lanjun Wang, Zehao Wang, Le Wu, and An-An Liu. Bots shield fake news: Adversarial attack on user engagement based fake news detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2369–2378, 2024.

[15] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

[16] Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, page 102701, 2024.

[17] Aminul Huq, Mst Pervin, et al. Adversarial attacks and defense on texts: A survey. *arXiv preprint arXiv:2005.14108*, 2020.

[18] Ahmed K. Kadhim, Lei Jiao, Rishad Shafik, and Ole-Christoffer Granmo. Adversarial attacks on ai-generated text detection models: A token probability-based approach using embeddings, 2025.

[19] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

[20] Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. *arXiv preprint arXiv:1912.10375*, 2019.

[21] Bing He, Mustaque Ahamad, and Srijan Kumar. Petgen: Personalized text generation attack on deep sequence embedding-based classification models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 575–584, 2021.

[22] Yasaman Boreshban, Seyed Morteza Mirbostani, Seyedeh Fatemeh Ahmadi, Gita Shojaee, Fatemeh Kamani, Gholamreza Ghassem-Sani, and Seyed Abolghasem Mirroshandel. RobustQA: A framework for adversarial text generation analysis on question answering systems. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 274–285, Singapore, December 2023. Association for Computational Linguistics.

[23] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*, 2021.

[24] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.

[25] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024.

[26] Peican Zhu, Zechen Pan, Yang Liu, Jiwei Tian, Keke Tang, and Zhen Wang. A general black-box adversarial attack on graph-based fake news detectors. *arXiv preprint arXiv:2404.15744*, 2024.

[27] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378, 2024.

[28] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

[29] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[30] Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[31] Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024.

[32] Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126, 2020.

[33] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[34] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[35] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[36] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*, 2018.

[37] Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. Who wrote this? the key to zero-shot llm-generated text detection is gecscore. *arXiv preprint arXiv:2405.04286*, 2024.

[38] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266, 2024.

[39] Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. Token prediction as implicit classification to identify llm-generated text. *arXiv preprint arXiv:2311.08723*, 2023.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[42] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[44] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

[45] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[47] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[48] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. *The Thirty-Fifth IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, USA, June 17-21*, 2024.

A Limitations

Although VERITAS demonstrates robust performance against style-based adversarial attacks, this study has several limitations that provide avenues for future research:

Dependency on the diversity of augmented styles. While VERITAS is designed to learn style-agnostic, content-centric features, its training paradigm inherently relies on the diversity of the stylistic variants generated by the auxiliary LLM. Consequently, the model may still experience performance degradation if confronted with completely out of distribution adversarial attacks, such as extreme novel writing styles, structural obfuscation adversarial rewriting that fundamentally alters the semantic content itself rather than just the stylistic attributes.

Effectiveness on very short texts. Our framework effectively leverages representations such as semantic coherence, factual consistency, and logical structure. However, these content-driven features typically require a sufficient length of text to be accurately extracted. The performance of VERITAS might degrade when applied to extremely short texts (e.g., single-sentence social media posts, short product reviews, or brief comments), where logical and factual contexts are too sparse to distinguish LLM-generated artifacts from human writing reliably.

Computational Overhead in Data Construction. The training pipeline of VERITAS requires generating multiple stylistic variants (N_s) for each human-written sample and querying an LLM to generate pseudo-labels for veracity attribution supervision. This data augmentation and attribution-labeling process introduces higher computational overhead and time costs prior to the actual model training, compared to standard fine-tuning approaches.

B Broader Impacts

This work presents VERITAS, a robust and style-agnostic framework for detecting text generated by large language models. The deployment of such technologies carries significant societal implications, encompassing both positive benefits and potential negative risks.

Potential Positive Societal Impacts: The deployment of our framework offers substantial benefits for digital information ecosystems. By identifying generated text even when malicious actors apply sophisticated stylistic obfuscation, this work provides a robust defense against the dissemination of deceptive content. In the context of political discourse and public news, the capacity to reliably detect manipulated content helps curb the spread of computational propaganda and fake news. Applying this technology to specialized domains like scientific literature supports educational and publishing institutions in maintaining academic integrity and preventing the pollution of scientific records with fabricated data. Reliable detection mechanisms ultimately foster public trust in digital platforms by ensuring the authenticity of the information consumed by society.

Potential Negative Societal Impacts: Malicious actors might analyze our content-driven approach to engineer more advanced evasion techniques that directly manipulate logical or semantic structures.

C Implementation Details

Data Augmentation and Preprocessing. To construct the augmented training dataset, we employ GPT-4 as the auxiliary language model M_{LLM} to perform the style-based reframing. For each human-written text in the training set, we prompt the model to generate $N_s = 4$ stylistic variants corresponding to the attack scenarios described in our methodology (e.g., introducing colloquial expressions, increasing subjectivity, mixing topics, and embedding specialized terminology). To obtain the ground-truth attribution vectors for the veracity attribution loss (L_{attr}), we query the same LLM using the predefined multiple-choice prompt, extracting a multi-hot encoded vector for the predefined attribution space ($N_a = 5$). All input texts are tokenized using the corresponding tokenizer of the backbone model, and the maximum sequence length is strictly truncated or padded to 512 tokens to ensure consistent memory usage during training [48].

Model Configuration. Our proposed framework VERITAS is designed to be compatible with various foundational architectures. For mid-sized pre-trained language models (such as RoBERTa or OPT-2.7B), we fine-tune all model parameters. For large language models (such as LLaMA2-13B and Qwen-2), we apply LoRA to mitigate the massive computational overhead while preserving the

generation capabilities. Specifically, LoRA is applied to the query and value projection matrices within the self-attention modules, with a rank of $r = 8$ and an alpha scaling factor of $\alpha = 16$. The representation h_p is extracted from the last hidden state of the final layer (or the output of the `CLS` token for encoder-only models) and fed into a two-layer Multi-Layer Perceptron (MLP) with a GELU activation function to yield the final classification distributions.

Training Setup and Hyperparameters. The overall objective function is a weighted sum of the classification loss, style alignment loss, and attribution loss, formulated as $L = L_{\text{class}} + \lambda_1 L_{\text{style}} + \lambda_2 L_{\text{attr}}$. Based on our empirical observations on the validation set, we set the penalty coefficients to $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$ to balance the magnitude of gradients from the auxiliary tasks. All models are trained using the AdamW optimizer with a weight decay of 0.01. The learning rate is set to 2e-5 for full fine-tuning and 1e-4 for LoRA-based configurations, accompanied by a linear learning rate scheduler with a 10% warmup proportion. We train the models for 3 epochs with an effective batch size of 32, utilizing gradient accumulation where necessary.

Hardware and Software Environment. All experiments are implemented using the PyTorch framework and the HuggingFace Transformers library. The training and evaluation processes are executed on a unified computing cluster equipped with four NVIDIA A6000 48GB GPUs. To ensure rigorous evaluation and reproducibility, we set a fixed random seed of 42 across all data splitting, shuffling, and model initialization procedures.

D Use Of AI Assistants

We utilized AI assistants (specifically ChatGPT-5.1) solely for the purpose of grammatical error correction, sentence polishing, and improving the fluency of the writing. All scientific claims, experimental designs, code implementation, and logical arguments are the original work of the authors.

Algorithm 1 Training Pipeline of `VERITAS`

Require: Human-written training set D_{human}, Auxiliary LLM M_{LLM}, Backbone Detector f_θ.
Ensure: Trained detector parameters θ^*.

Offline Data Preparation:
for each $p_{\text{human}} \in D_{\text{human}}$ **do**
  Generate stylistic variants $\{p_{\text{gen},i}\}_{i=1}^{N_s}$ using M_{LLM} (Eq. 3).
  Query M_{LLM} to generate multi-hot attribution pseudo-labels A_i.
end for
Construct augmented dataset D_{train}.
Training Phase:
while not converged **do**
  Sample a batch of texts and their augmented variants from D_{train}.
  Extract hidden representations h_p via backbone f_θ.
  Compute classification probability distributions $\hat{y}$ (Eq. 6).
  Compute Classification Loss L_{class} using standard cross-entropy.
  Compute Style Alignment Loss L_{style} via pair-wise KL-divergence across stylistic variants to enforce consistency (Eq. 7).
  Compute Attribution Loss L_{attr} via Binary Cross-Entropy against pseudo-labels A_i.
  Compute total loss $L = L_{\text{class}} + \lambda_1 L_{\text{style}} + \lambda_2 L_{\text{attr}}$.
  Update θ via backpropagation.
end while
return θ^*

E Formal Training Algorithm

To provide a clearer understanding of the training mechanism, we present the formal training pipeline of `VERITAS` in Algorithm 1. Our framework utilizes large language models purely as an offline data engine during the training phase. The final deployed detector is a specialized, smaller, and highly efficient machine learning model, optimizing computational efficiency during inference while benefiting from the deep semantic supervision extracted from LLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the contributions and scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental details necessary for reproduction are included in Appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and preprocessed data are publicly available at `https://anonymous.4open.science/status/A-Style-Agnostic-Framework-for-Detecting-LLM-Generated-Text-90B7` with detailed instructions in the README.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training/test splits, hyperparameters, and optimizer settings are specified in experiment section.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results report mean ± standard deviation over 5 random seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Experiments were run on NVIDIA A6000 GPUs, with 48GB of memory, and the total compute time was approximately 100 GPU hours.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work complies with the NeurIPS Code of Ethics; no human subjects or sensitive data were involved.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive and negative societal impacts of the work performed.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release high-risk models or scraped datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and libraries are properly cited with licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our new dataset/model is documented in appendix with license and usage terms.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve human subjects or crowdsourcing.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We utilized AI assistants (specifically ChatGPT-5.1) solely for the purpose of grammatical error correction, sentence polishing, and improving the fluency of the writing. All scientific claims, experimental designs, code implementation, and logical arguments are the original work of the authors.