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

Anonymous Author(s)

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 vulnerabilities in existing detection systems:(1) State-of-the-art detectors suffer a substantial accuracy decline, reaching up to 16.45% when exposed to style-based adversarial attacks generated by LLMs. (2) While general-purpose LLMs exhibit remarkable zero-shot capabilities, their performance in detecting adversarially manipulated text is lower than specialized detectors fine-tuned for robustness. To address these vulnerabilities, 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. The dataset and code are in the link [1].

Keywords

Large language model, Style-Agnostic, content-driven

ACM Reference Format:

Anonymous Author(s). 2025. VERITAS:Veracity-Enhanced Robust Identification of LLM-generated Text Against Style-shifts. In *Proceedings of Proceedings of the ACM Web Conference 2026 (WWW'26).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have recently demonstrated remarkable capabilities in generating textual data [21, 48], 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 [43]. Considering this, 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 [20, 37, 43], 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 style while preserving semantic coherence. 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 vulnerabilities: (1) *state-of-the-art detectors for generated text are significantly compromised by LLM-induced stylistic variations*, and (2) *general-purpose LLMs exhibit markedly inferior performance in detecting adversarially manipulated text.* Motivated by two critical vulnerabilities, 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. Then, 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 our knowledge, *this is the first approach to forgo stylistic artifacts in favor of content-driven analysis*, ensuring its broad applicability.
- We construct an adversarially enriched dataset by leveraging LLMs fine-tuned for style-based attacks. *Extensive*

[1] https://anonymous.4open.science/status/A-Style-Agnostic-Framework-for-Detecting-LLM-Generated-Text-90B7

WWW'26, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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%.

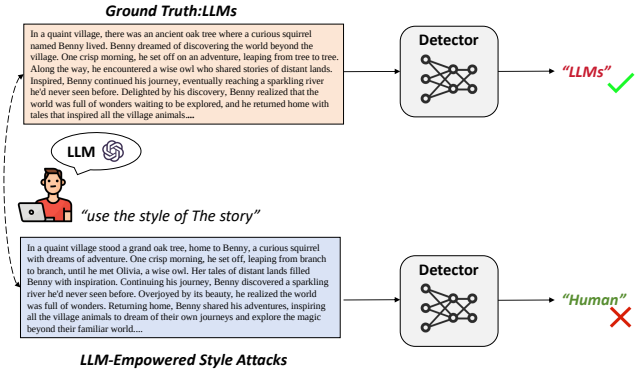


Figure 1: We have found that when the content generated by LLMs undergoes style transformation, the existing approaches and detection tools prove to be inadequate. They are unable to precisely distinguish whether the content is crafted by humans or generated by LLMs. This situation highlights that modifying the style of the generated content enables the circumvention of current detection mechanisms.

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 [2, 3, 29, 30, 39]. 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, 27, 30, 39]. 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 [9, 33, 35]. 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 [1, 46]. 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 [16, 17]. Many approaches subtly perturb parts of the text or embed feature-specific text snippets [5, 15, 19, 34], causing NLP models to produce targeted incorrect outputs while preserving semantic and syntactic integrity. Although adversarial examples in other domains, i.e., image recognition [7, 24, 40, 45] and fake news detection [8, 41, 49], 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 $f_\theta(x) : \mathcal{X} \rightarrow \{0, 1\}$ 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 employ LLMs to generate texts with diverse stylistic variations. We conduct a preliminary analysis to assess the performance of state-of-the-art detections in identifying this generated content.

4.1 Diverse Stylistic Variations

The advanced capabilities of LLMs enable users to transform text styles through tailored prompts, challenging the robustness of detection systems against such stylistic alterations. In this study, we investigate a direct style-based attack by employing distinctive writing styles characteristic of texts such as Andersen's fairy tales and the scientific prose found in prestigious academic journals like Nature and Science as prompts. These writing styles are marked by distinctive narrative elements and tonal qualities, making them viable options for adversarial manipulations. For instance, a modern narrative might be rewritten with whimsical language, moral undertones, and vivid imagery characteristic of fairy tales. Our general prompt format for these transformations is structured as follows:

> **Rewrite the following text using the style of [publisher/book]:** [input text]

For narrative texts, we employ the writing style of Andersen's fairy tales to transform stories generated by LLMs. For political texts, we adopt the style of CNN, while for scientific texts, we utilize the writing style of Nature and Science. We employ these stylistically transformed test samples to systematically evaluate the

performance of detection systems when subjected to style-oriented adversarial attacks.

4.2 Style-Related Detector Vulnerability

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.

As shown in Table 1, the original accuracy (O) reflects the detection methods' performance on unaltered texts, while the adversarial accuracy (A) represents their robustness when faced with style-based adversarial attacks, with a drop in accuracy indicated by (↓). DetectGPT [22] demonstrates the highest original and adversarial accuracies across all datasets, particularly excelling in Story (O: 78.24%, A: 70.13%) and PolitiFact (O: 75.91%, A: 71.96%). However, its performance drops more significantly in the Science dataset (O: 69.72%, A: 60.43%), highlighting the challenge of adversarial attacks in more structured or technical domains. GLTR [12] and LLMDet [43] exhibit weaker performance, with GLTR demonstrating notable vulnerability under adversarial conditions. For instance, GLTR's accuracy drops on the PolitiFact dataset (O: 59.24%, A: 52.32%). LLMDet performs moderately better, particularly on the PolitiFact dataset (O: 70.08%, A: 62.57%). The LLaMA-7B and Neo-2.7B models exhibit relatively lower original and adversarial accuracies overall, particularly in the Science dataset where Neo-2.7B struggles the most (O: 62.25%, A: 48.33%), underlining the challenges of handling adversarial attacks with smaller or less robust LLMs. Existing detectors and general large language models perform inadequately under adversarial attacks, particularly when handling specialized or technical texts, where their performance degradation is especially pronounced.

Observation 1 (Style-related vulnerability of LLMs generated text detectors). State-of-the-art detectors for generated text are found to be impacted by LLM-driven stylistic variations. This impact results in the performance drop, as evidenced by an accuracy decline of up to 16.45% when evaluated on stylistically altered test sets.

Observation 2 (Limitations of LLMs in text robustness detection). While LLMs demonstrate remarkable zero-shot capabilities as general-purpose foundational models, their performance in detecting adversarially manipulated text is notably inferior compared to specialized LLMs-generated text detection systems and pre-trained language models fine-tuned for specialized or technical tasks.

5 Methodology

In this section, we propose a method to detect LLM-generated content by transforming human texts into various styles using an LLM for diverse training data. Features combine contextual embeddings and statistical metrics. The training uses three loss functions: style alignment, classification, and pseudo-label supervision. This ensures robustness to stylistic variations and adaptability to different generation patterns.

Algorithm 1 VERITAS

Require: Dataset D, LLMs M_{LLM}, Styles N_s.
Ensure: Classifier M to predict whether a text is human-written or LLMs-generated.

Initialize: Augmented dataset $D_{\text{train}} \leftarrow \emptyset$.
Data Augmentation:
for each $p_{\text{human}} \in D_{\text{human}}$ **do**
  for each style s_i, $i = 1, 2, \ldots, N_s$ **do**
    Generate style-augmented text:
    $p_{\text{gen},i} \leftarrow M_{\text{LLM}}(p_{\text{human}}, \text{style} = s_i)$
    Add $p_{\text{gen},i}$ to D_{train}.
  end for
end for$D_{\text{train}} \leftarrow D_{\text{human}} \cup \{p_{\text{gen},i} \mid i = 1, 2, \ldots, N_s\}$
Training the Classifier:
for each batch $B \subset D_{\text{train}}$ **do**
  Compute predictions for each $p \in B$.
  Compute losses:
  $L_{\text{style}} = \frac{1}{N_s^2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \text{KL}(y_i, y_j)$
  $L_{\text{class}} = -\sum_{k=1}^{K} y_k \log \hat{y}_k$
  $L_{\text{attr}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \text{BCE}(s_i, \hat{s}_i)$
  $L = L_{\text{style}} + L_{\text{class}} + L_{\text{attr}}$
  Update M using backpropagation.
end for
Output: Trained classifier M.

5.1 Style-Based Reframing

To simulate the diverse styles that LLM-generated content may exhibit in real-world scenarios, we adopt a data augmentation strategy based on style transformation. Specifically, each human-written text $p_{\text{human}} \in D_{\text{human}}$ is transformed into multiple stylistic variants by leveraging an LLM M_{LLM}. N_s is the number of stylistic transformations, s_i represents a specific style (e.g., "formal," "narrative," or "scientific"), and the equation 3 describes the transformation process.

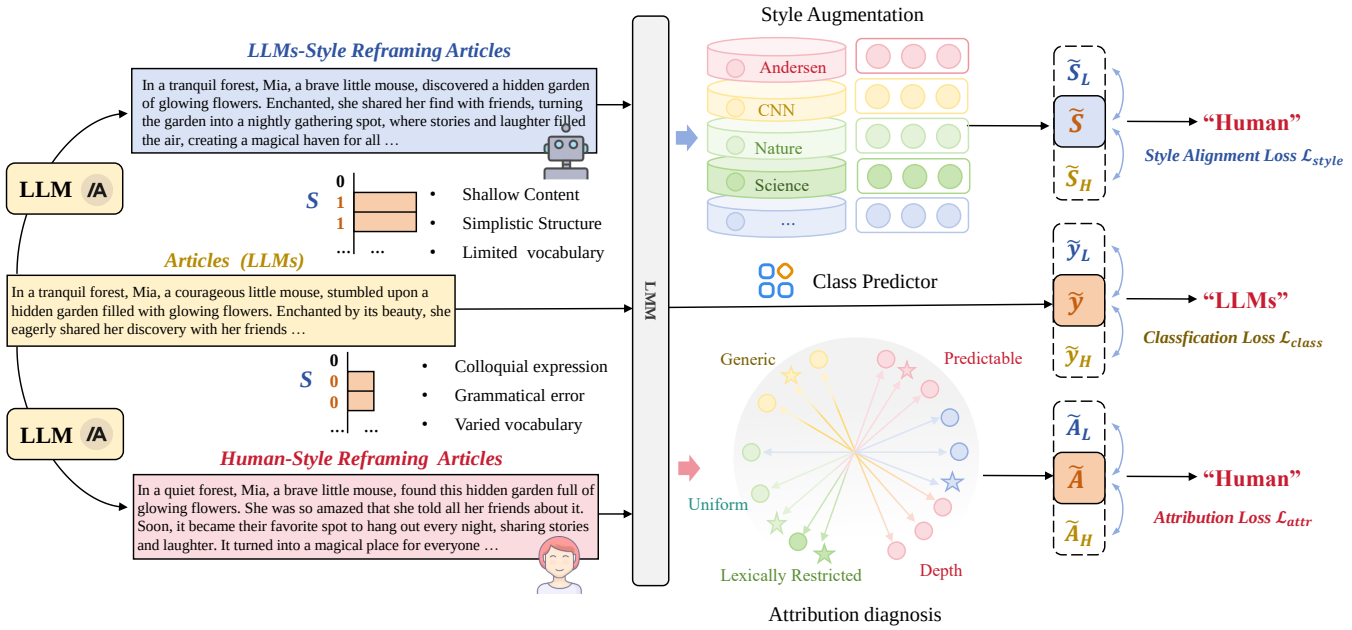
$$p_{\text{gen},i} = M_{\text{LLM}}(p_{\text{human}}, \text{style} = s_i), i = 1, \ldots, N_s \tag{3}$$

The augmented training dataset D is subsequently constructed by integrating stylistic variants derived from human-written texts with equation 4, thereby ensuring a diverse and comprehensive 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 Alignment Loss

To improve robustness to stylistic variations, our style alignment loss enforces consistent predictive distributions across N_s stylistic



$$\mathcal{L} = \mathcal{L}_{style} + \mathcal{L}_{class} + \mathcal{L}_{attr}$$

Figure 2: 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.

variants $\{p_{\text{gen},i}\}$. We achieve this by passing the hidden representation $h_{p_{\text{gen},i}}$ of each variant through an MLP classifier M to obtain the output probability distributions:

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

The style alignment loss L_{style}, is then 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 [44]. This encourages the classifier to produce a consistent output, irrespective of superficial stylistic features.

5.3 Classification Loss

The classification loss ensures that the model correctly predicts the source of the input text. For a given input p, with true label y and predicted probabilities $\hat{y}$, the loss is:

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

where $K = 2$ corresponds to the two classes (human-written and LLMs-generated)

5.4 Content-Focused Attribution Supervision

The veracity attributions are generated through a process that involves querying an LLM to identify distinctive features and patterns characteristic of text produced by LLMs. We employ the following inquiry method:

> **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 given veracity attributions space $C = \{c_1, c_2, \ldots, c_m\}$, the generated-text $p_{\text{gen},i}$ attribution A_i satisfy:

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

we define a veracity attribution loss based on the binary cross-entropy between the predicted attribution vector $\hat{\mathbf{A}}_i$ and the ground-truth vector $\mathbf{A}_i$. The attribution loss is averaged over N_a samples.

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

Table 2: 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 3. Bold denotes the overall best results.

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

5.5 Final Objective Function

The overall training objective combines style alignment loss, classification loss, and veracity attribution supervision loss:

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

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 [13] and story-focused articles from the X-Sum dataset [23]. 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 [38] and FakeNewsNet datasets [28], 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 [26] or scientific literature from the Gutenberg dataset, evaluating the model's proficiency in handling specialized terminology, logical reasoning, and domain knowledge.

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 3: Strategies for text stylization based on different writing styles

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 three primary performance metrics, e.g., Accuracy (A), Area Under the Receiver Operating Characteristic Curve (AUROC), 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 [29] assesses text

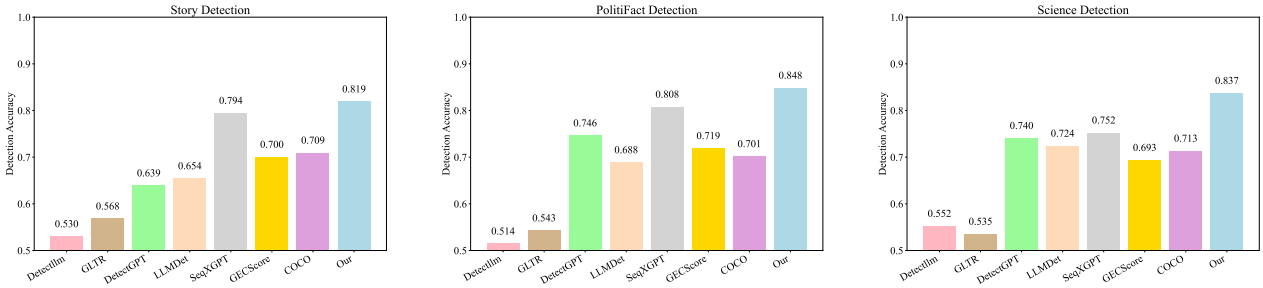


Figure 3: The visual data presented in the graphs clearly indicates that our methodology excels in detection accuracy across multiple categories,e.g., Story, PolitiFact, and Science. Our approach consistently outperforms other methods, achieving the highest accuracy in each category.

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 4: On different LMM backbones, VERITAS demonstrates stable improvements on accuracy.

origin via log perplexity, indicating predictability. GLTR [12] combines statistical methods with visual analytics to highlight anomalous token probabilities, aiding in the identification of machine-generated text. DetectGPT [22] leverages the curvature properties of language model probability functions within a probabilistic framework to identify synthetic text. LLMDet [43] uses surrogate perplexity calculations tailored to each LLM, offering a model-agnostic solution for text provenance. SeqXGPT [37] represents sentences as waveforms, utilizing convolutional networks and self-attention mechanisms for sentence-level detection. GECScore [42] evaluates text similarity using a grammar error correction model, providing a robust metric for LLM-origin detection. The OUTFOX method [18] bolsters the robustness of detecting text generated by LLMs by implementing an iterative in-context learning framework. COCO [20] enhances detection through contrastive learning. T5-Sentinel [6] 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 [25], OPT-2.7B [47], Neo-2.7B [11], LLaMA-7B [32], LLaMA-13B, GPT-NeoX [4], GPT-3.5-turbo, GPT-4, Gemma [31], Qwen-2 [36], Deepseek-R1 [14]. 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 2 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 3, 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.

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.

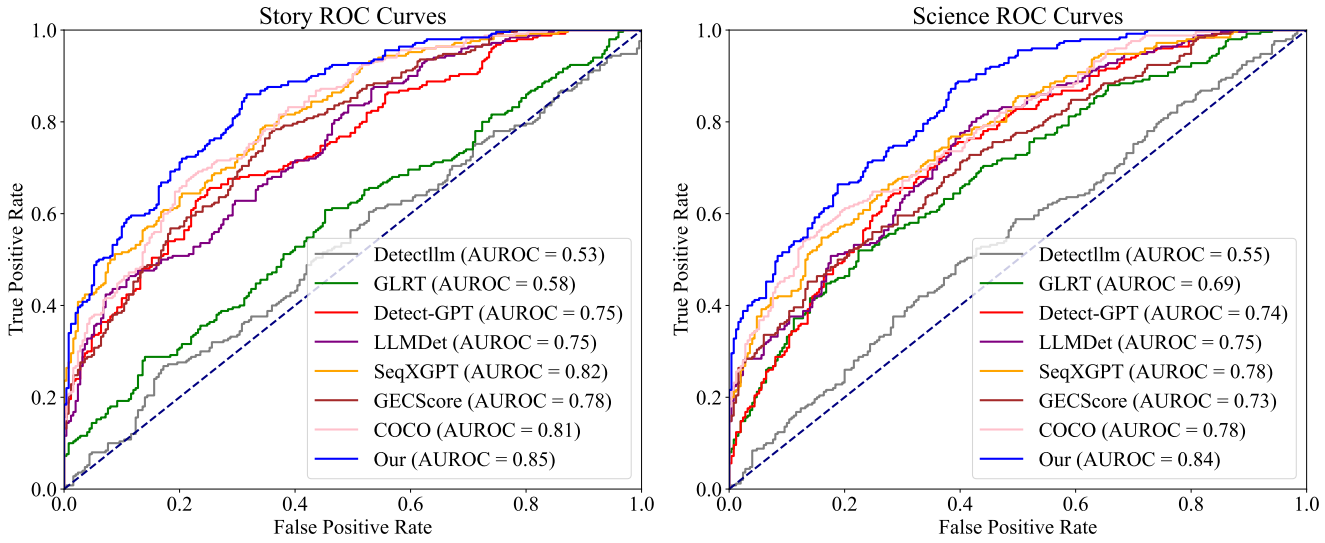


Figure 4: The figure presents the ROC curves for various methods evaluated on two categories: Story and Science. The AUROC is used to quantify each method's ability to distinguish between human and LLM-generated content.

Table 5: Ablation Study of VERITAS Loss Components under Different attack Scenarios (F1 Score %). The strategies for text stylization are in Table 3.

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

Ablation study. As shown in Figure 5, 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.

AUROC. As shown in Figure 4, our method demonstrates significant advantages in distinguishing between human-written content and content generated by LLMs. Specifically, in the Story category, VERITAS achieved an AUROC value of 0.85, ranking first among all comparative methods. This result highlights its exceptional performance in handling creative and narrative texts. Furthermore, in the Science category, VERITAS also performed impressively with an AUROC value of 0.84, which is not only significantly higher than that of COCO (0.78) but also surpasses GECScore (0.73). VERITAS exhibits

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

unique strengths when confronted with adversarial sample attacks across various text styles. Whether in literary creation, scientific discourse, or other types of texts, VERITAS effectively resists the impact of adversarial samples, maintaining high-precision discrimination capabilities. By thoroughly analyzing multiple dimensions of textual features, including intrinsic content, word preferences, and logical coherence, VERITAS can accurately identify and adapt to stylistic variations in texts, thereby providing reliable and precise judgments.

Cross-Domain Evaluation. To rigorously assess the generalization capabilities of our framework 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. First, 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. Second, 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.

7 Conclusion

We address critical vulnerabilities in detecting LLM-generated text, particularly against style-based adversarial attacks powered by LLMs. Existing detectors exhibit performance degradation when confronted with stylistic manipulations, highlighting the limitations of their reliance on stylistic cues. To overcome these challenges, we proposed 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 disentangles semantic content from stylistic variations, ensuring enhanced robustness against diverse adversarial attacks. Future work could explore integrating multi-modal signals or developing certified defense mechanisms to provide formal guarantees.

References

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

[2] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351* (2019).

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

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

[5] Yasaman Boreshban, Seyed Morteza Mirbostani, Seyedeh Fatemeh Ahmadi, Gita Shojaee, Fatemeh Kamani, Gholamreza Ghassem-Sani, and Seyed Abolghasem Mirroshandel. 2023. RobustQA: A Framework for Adversarial Text Generation Analysis on Question Answering Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 274–285.

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

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

[8] Karen M DSouza and Aaron M French. 2024. Fake news detection using machine learning: an adversarial collaboration approach. *Internet Research* 34, 5 (2024), 1664–1678.

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

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

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027* (2020).

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

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

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

[15] Bing He, Mustaque Ahamad, and Srijan Kumar. 2021. 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*. 575–584.

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

[17] Ahmed K Kadhim, Lei Jiao, Rishad Shafik, and Ole-Christoffer Granmo. 2025. Adversarial Attacks on AI-Generated Text Detection Models: A Token Probability-Based Approach Using Embeddings. *arXiv preprint arXiv:2501.18998* (2025).

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

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

[20] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 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.

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

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

[23] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium.

[24] KL Navaneet, Soroush Abbasi Koohpayegani, Essam Sleiman, and Hamed Pirsiavash. 2024. SlowFormer: Adversarial Attack on Compute and Energy Consumption of Efficient Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24786–24797.

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

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

[27] Yuhui Shi, Qiang Sheng, Juan Cao, Hao Mi, Beizhe Hu, and Danding Wang. 2024. Ten Words Only Still Help: Improving Black-Box AI-Generated Text Detection via Proxy-Guided Efficient Re-Sampling. *arXiv preprint arXiv:2402.09199* (2024).

[28] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. 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).

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

[30] Kaito Taguchi, Yujie Gu, and Kouichi Sakurai. 2024. The Impact of Prompts on Zero-Shot Detection of AI-Generated Text. *arXiv preprint arXiv:2403.20127* (2024).

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

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

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

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

[35] Lanjun Wang, Zehao Wang, Le Wu, and An-An Liu. 2024. 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*. 2369–2378.

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

[37] Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-Level AI-Generated Text Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1144–1156. https://aclanthology.org/2023.emnlp-main.73

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

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

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

[41] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. 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*. 3367–3378.

[42] Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. 2024. Who Wrote This? The Key to Zero-Shot LLM-Generated Text Detection Is GECScore. *arXiv preprint arXiv:2405.04286* (2024).

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

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

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

arXiv:2103.01498 (2021).

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

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

[48] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

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