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
[2] Division of Integrative Systems and Design, Hong Kong University of Science and Technology
[3] Mohamed bin Zayed University of Artificial Intelligence
* Corresponding author: Jinghua Zhang

Abstract & Introduction

Large language models (LLMs) can increasingly imitate diverse human writing styles, making reliable identification of LLM-generated text essential. However, existing detectors often depend on lexical choices, sentence structures, and other stylistic cues that can be manipulated through adversarial rewriting.

This work identifies two limitations: current detectors suffer substantial performance degradation under LLM-powered style attacks, and general-purpose LLMs remain less reliable than specialized, robustness-tuned detectors on adversarially modified text.

We propose VERITAS, a style-agnostic framework that prioritizes content-driven features over surface style. By combining adversarial style augmentation, style-invariant learning, and veracity attribution supervision, VERITAS learns to identify text provenance across stylistic variations. Experiments on Story, PolitiFact, and Science demonstrate improved detection performance and cross-domain robustness.

Motivation: Detection under Style Shifts

Advanced LLMs can rewrite generated text in recognizable styles while preserving its underlying meaning. We simulate these attacks using stylistic proxies such as Andersen's fairy tales, CNN news writing, and the scientific prose of Nature and Science. These transformations expose the fragility of detectors that rely on superficial language patterns.

Table 1 compares detection accuracy before and after style-based attacks. On Science, LLaMA-7B drops from 68.59% to 52.14%, a decline of 16.45 percentage points. DetectGPT also declines from 69.72% to 60.43%, illustrating that strong performance on unaltered text does not guarantee robustness.

These observations motivate a detector that remains consistent across writing styles and uses content-related evidence, including logical structure, repetitive expression, and factual errors, rather than treating style alone as a reliable signal of authorship.

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: Original (O) and adversarial (A) accuracy across three domains.

Methodology

VERITAS integrates style augmentation and content-driven supervision within a unified training framework, as shown in Figure 1. Auxiliary LLMs generate stylistic variants that broaden the detector's exposure to diverse writing patterns. The detector is optimized with three complementary objectives.

The style alignment loss minimizes the pairwise Kullback–Leibler divergence between predictions for stylistic variants, encouraging consistent outputs despite surface-level changes. The classification loss uses cross-entropy to distinguish human-written from LLM-generated text.

The veracity attribution loss uses binary cross-entropy to learn content-related diagnostic labels provided by an auxiliary LLM, including logical regularity, repetition, missing background knowledge, and data errors. The three losses are jointly optimized, while auxiliary LLMs are used offline and are not required during detector deployment.

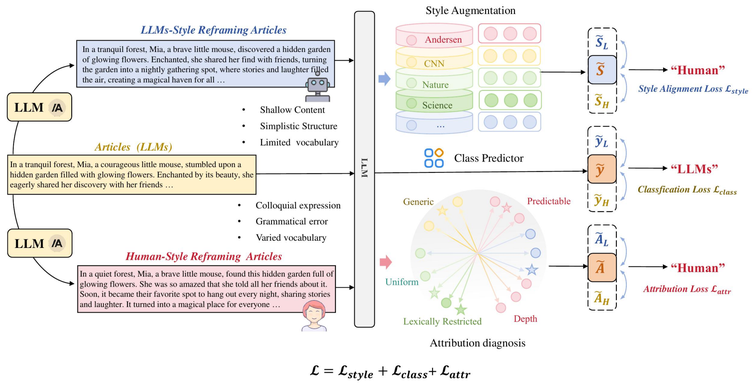


Figure 1: Overview of the VERITAS framework.

Experimental Results

We evaluate VERITAS on three content domains: Story (Gutenberg and X-Sum), PolitiFact (LIAR and FakeNewsNet), and Science (MedNLI and scientific literature). Four attack scenarios introduce colloquial language and minor errors, personalization and subjectivity, topic mixing and diverse sources, or specialized terminology and cultural references.

As shown in Table 2, VERITAS achieves the highest F1 score in all eight Story and PolitiFact attack conditions. Its Story F1 scores range from 77.60% to 79.56%. On PolitiFact, it reaches 84.69% under scenario B, compared with 71.86% for COCO.

With Qwen-2, VERITAS improves Story accuracy from 66.43% to 85.23% and PolitiFact accuracy from 61.32% to 87.24%.

When trained only on Story and tested on Science, VERITAS maintains F1 scores of 80.16%–83.30% across four attack scenarios, demonstrating cross-domain robustness.

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**

Table 2: F1 (%) under four style attacks on Story and PolitiFact.

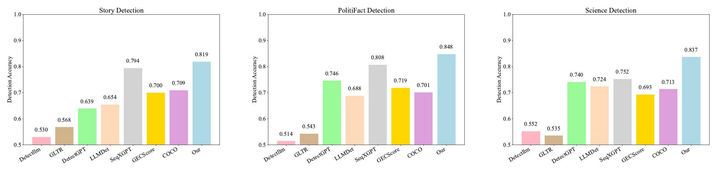


Figure 2: Detection accuracy on Story, PolitiFact, and Science.

Ablation Study

Table 3 shows that both auxiliary objectives improve F1 over classification-only training. On Story under attack A, adding style alignment increases F1 from 64.10% to 71.51%, while adding attribution supervision reaches 73.29%. The full model achieves 78.45%.

The full model performs best across all domains and attacks, showing that style invariance and content attribution are complementary.

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

Table 3: Ablation of the three training objectives.

References

Wu J, Guo J, Hooi B. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. KDD, 2024: 3367–3378.

Mitchell E, Lee Y, Khazatsky A, et al. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. ICML, 2023: 24950–24962.

Koike R, Kaneko M, Okazaki N. OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. AAAI, 2024: 21258–21266.