

RATF: Robust Adaptive Tri-Encoder Framework for Fake News Detection against Style-Confusion Attacks from LLMs

Haoda Zhang
Software College
Northeastern University
Shenyang, Liaoning, China
zhanghd4@mails.neu.edu.cn

Zhenhua Tan*
Software College
Northeastern University
Shenyang, Liaoning, China
tanzh@mail.neu.edu.cn

Ziwei Cheng
Software College
Northeastern University
Shenyang, Liaoning, China
chengzw@stumail.neu.edu.cn

Abstract

The rise of Large Language Models (LLMs) enables a new challenge for fake news detection: style-confusion attacks, where attackers rewrite news to mimic trusted styles, challenging detectors that rely on stylistic features. Current defenses like style-invariance training ignore stylistic features, but this is suboptimal for style-confusion attack datasets as it discards valuable cues in clean (non-attacked) articles. To defend against style-confusion attacks, we propose the Robust Adaptive Tri-Encoder Framework (RATF). Our framework uses an LLM to decouple news into content features and style features, and leverages conditional probability curvature to compute a “rewritten score” estimating the likelihood that the article has been rewritten by an LLM. RATF injects this score into a score-conditioned cross-attention fusion module and a learnable parametric gating module. The former uses the content representation as a reliable query to retrieve relevant semantic and stylistic evidence, while the latter continuously calibrates the contributions of original semantics, content features, and style features. We simulate these attacks by creating an equal mixture of original articles from the PolitiFact and GossipCop datasets and their four corresponding style-rewritten versions generated by an LLM. Our experiments show RATF achieves state-of-the-art performance with superior robustness against style-confusion attacks, while also proving highly adaptable across various backbones.

1 Introduction

Fake news has become a major threat to information integrity, affecting politics, public health, and social cohesion (Lazer et al., 2018; Allcott and Gentzkow, 2017).¹ Because it can spread faster than authentic information and manual verification

* Corresponding author.

¹ Code: <https://github.com/Iceldancer/RATF>

Robustness against
Style-Confusion Attacks

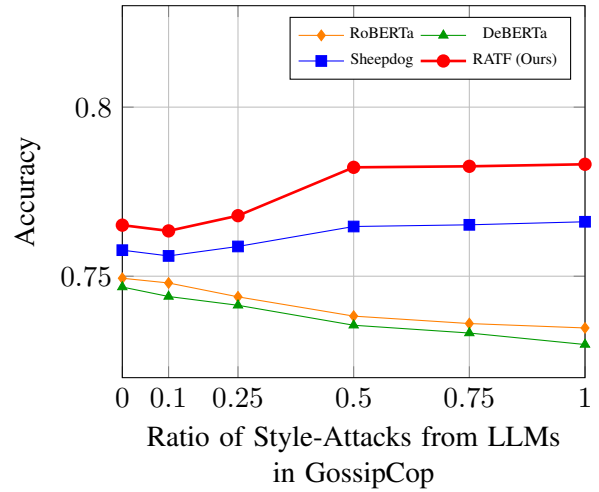


Figure 1: Observations: Performance of different models on the GossipCop dataset under varying ratios of style-confusion attacks from LLMs.

is costly (Vosoughi et al., 2018; Pérez-Rosas et al., 2018), automated fake news detection has received growing attention.

Initially, advances in detection leveraged both content-based and stylistic features, drawing on evidence that legitimate and fabricated news exhibit distinguishable linguistic patterns (Potthast et al., 2018). Stylistic features proved particularly valuable, as unreliable sources often employ sensationalist language, emotional appeals, and partisan framing, in contrast to the measured tone of reputable journalism (Przybyla, 2020). Integrating these stylistic indicators into detection frameworks led to substantial performance improvements on benchmark datasets.

However, the emergence of Large Language Models (LLMs) has fundamentally altered this landscape, shifting the creation of fake news from crude fabrication to a technologically-driven, in-

dustrialized process (Achiam et al., 2023; Brown et al., 2020). Malicious actors can now leverage these models to decouple a core false claim from its original inflammatory style and reconstruct it to mimic the linguistic style of a trusted source (Wu et al., 2024). In practice, such evasion can take the form of content attacks, which distort factual claims, or style attacks, which preserve the same false claim while changing its linguistic presentation. Even if a rewritten article becomes less sensational and may spread less explosively, it can still be more harmful to detection systems and readers because the false claim is “laundered” into a formal or journalistic tone that appears more credible. This makes style manipulation a low-cost and increasingly accessible attack path as LLMs become widely available. To evade detection by existing systems, attackers can then mix these stylistically disguised articles with original, unrewritten news, thereby confusing detectors that rely on consistent stylistic patterns. We term this strategy a **style-confusion attack**. The challenge posed by such attacks is empirically demonstrated in Figure 1. As shown on the GossipCop dataset (Shu et al., 2020), standard detectors like RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) suffer significant performance degradation as the proportion of style-attacked news increases. Conversely, while style-invariant defenses like Sheepdog (Wu et al., 2024) remain robust against the attack, their performance is inherently capped. By design, these models deliberately disregard stylistic features entirely, which, despite improving robustness, incurs a performance penalty on non-manipulated articles where style can be a valuable indicator. In the context of style-confusion attacks, this dilemma—risking vulnerability by trusting style or sacrificing performance by ignoring it—is the central challenge our work addresses.

This leads to the central, unaddressed challenge: **how can a detection system selectively trust and utilize stylistic features, while simultaneously defending against an attacker’s manipulation of stylistic features?** This paper introduces a new adaptive framework that effectively narrows this gap. Instead of treating style as a feature to be either fully trusted or completely ignored, we treat the authenticity of the style itself as a variable to be estimated.

In this paper, we propose a novel Robust Adaptive Tri-Encoder based Fake news detector, named **RATF**, to bridge the gap between robust defense

and effective feature utilization. Our framework achieves this through a tri-encoder architecture, a score-conditioned cross-attention fusion module, and a learnable parametric gating mechanism that adaptively balances the influence of content and style features based on the likelihood that the article has been rewritten by an LLM.

The RATF detector includes three core components. Firstly, we compute the rewritten score via conditional probability curvature to assess the likelihood that a text has been algorithmically manipulated. This assessment, which we term a “rewritten score”, serves as a dynamic signal for our novel RATF. This model uses the score in two complementary ways: it conditions cross-attention fusion between the reliable content representation and the other feature streams, and it drives a learnable parametric gate that softly adjusts the weights assigned to style-based features relative to content-based and semantic features. This allows our system to leverage stylistic indicators when a text’s style is deemed authentic, while dynamically discounting them when manipulation is suspected, thereby maintaining robust performance against style-confusion attacks.

In summary, our contributions are:

1. We formulate style-confusion attacks from LLMs and propose a Tri-Encoder architecture that separates original semantics, content, and style.
2. We introduce RATF, which combines rewritten-score-conditioned cross-attention with learnable gating to adaptively balance semantic, content, and style evidence.
3. Experiments on benchmark datasets show that RATF achieves strong robustness and state-of-the-art performance under LLM-driven style-confusion attacks.

2 Related Work

Automated fake news detection has been studied with neural text models (Devlin et al., 2019; Pelrine et al., 2021) and auxiliary signals such as comments, propagation environments, knowledge bases, and social graphs (Shu et al., 2019; Sheng et al., 2022; Sakthivelu and Vinoth Kumar, 2023; Nguyen et al., 2020). Prior work also shows that unreliable sources often exhibit distinctive linguistic styles (Potthast et al., 2018). While these signals are useful, they become less reliable under de-

liberate style manipulation, which is the focus of our work.

Adversarial fake news studies have examined manipulated social engagement, user behavior, fact distortion, and subject-object exchange (Wang et al., 2023; He et al., 2021; Hu et al., 2025; Kanan and Hamza, 2026). LLMs further amplify this threat by enabling scalable misinformation generation and rewriting (Hu et al., 2024; Kreps et al., 2022; Huang et al., 2023; Chen and Shu, 2024a), while also being used for auxiliary detection and reasoning-based fact checking (Hu et al., 2023; Chen and Shu, 2024b; Asai et al., 2024; He et al., 2026; Ahmad and Hongson, 2026). Existing style-invariant defenses improve robustness but discard useful stylistic evidence; RATF instead estimates style credibility and adaptively balances style with content.

3 Methodology

Our proposed framework is designed to solve a critical challenge: enabling a detection system to selectively trust and utilize stylistic features while simultaneously defending against their manipulation. Our methodology unfolds in three stages: (1) We employ a powerful Large Language Model (LLM) to perform feature engineering, extracting raw news text into focused content features and style features. (2) We assess the probability that a news article has been rewritten by an LLM by calculating its conditional probability curvature, inspired by recent AI-text detection work, to yield a “rewritten score.” (3) These three distinct information streams—the original article’s semantics (A), the content features (A_{con}), and its style features (A_{sty})—are fed into our novel RATF architecture, which performs score-conditioned cross-attention fusion and learnable score-based feature calibration to produce the final veracity prediction. Hereafter we refer to these three textual streams as the *original semantics* (A), its *content features* (A_{con}), and the *style features* (A_{sty}).

3.1 LLM-Powered Feature Engineering

Traditional feature engineering for fake news detection often relies on pre-defined word lists or manually designed rules. We bypass this limitation by tasking a general-purpose LLM (Qwen2.5-7B-Instruct) with a single, multi-part prompt to concurrently distill the raw news text into its core content and style components. This

prompt is detailed in Appendix D (Figure 5). It instructs the model to perform two tasks: first, rewrite the news in an objective tone, and second, describe its original stylistic elements.

For each original article A , this process yields two distinct textual feature streams:

- **Content Feature** (A_{con}): The objective rewritten text, serving as a representation of the article’s core factual substance.
- **Style Feature** (A_{sty}): The LLM-generated description of the original semantics’s stylistic elements.

We use Qwen2.5-7B-Instruct as the feature extraction component. When a text is style-rewritten for adversarial evasion, RATF can identify the increased rewriting signal and reduce reliance on potentially manipulated stylistic evidence while preserving the content stream for veracity prediction. When a text is modified by AI only for grammar polishing, RATF may still assign a high rewritten score and reduce reliance on style. This behavior is desirable for our threat model: after any AI rewriting, the original writing style is no longer fully reliable, whereas the content stream remains available for veracity prediction.

3.2 Rewritten Text Detection Module

With the rapid development of large language models, style attacks in the field of news detection have become increasingly accessible. Therefore, an important step in our solution is to assess the probability that a news article has been rewritten by a large model. Inspired by recent work in AI-text detection (Hu et al., 2023) and the fast-detect-gpt (Bao et al., 2024), we adopt their conditional probability curvature metric to make this assessment.

The core hypothesis is that human-written and machine-generated text exhibit different distributions of alternative word choices. Given a context, LLMs tend to select high-probability, predictable words. In contrast, human word choice is often more surprising and varied, resulting in a distribution with higher variance. This approach quantifies this difference by measuring how much the log-probability of the original semantics deviates from the average log-probability of its plausible alternatives.

Formally, given an input text x and a source language model p_θ (in our case, a model from the

GPT-Neo family (Black et al., 2021)), the process is as follows:

1. **Conditional Sampling:** For each position i in the text, a set of N alternative tokens, $\tilde{x}_i$, are sampled from the model’s conditional distribution $p_\theta(\cdot|x_{<i})$. This creates a set of alternative sequences $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \dots, \tilde{x}^{(N)}\}$.
2. **Distribution Estimation:** The log-probabilities of these sampled sequences are calculated. From these, we estimate the mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$ of the log-probability distribution of the alternatives:

$$\tilde{\mu} = \frac{1}{N} \sum_{j=1}^N \log p_\theta(\tilde{x}^{(j)}) \quad (1)$$

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{j=1}^N (\log p_\theta(\tilde{x}^{(j)}) - \tilde{\mu})^2 \quad (2)$$

3. **Curvature Calculation:** The conditional probability curvature, $d(x, p_\theta)$, is then defined as the standardized z-score of the original semantics’s log-probability with respect to the distribution of its alternatives:

$$d(x, p_\theta) = \frac{\log p_\theta(x) - \tilde{\mu}}{\tilde{\sigma}}$$

Finally, to make this metric directly usable in our framework, we convert the raw curvature score into a “rewritten score” s_{rw} via normalization and clipping, which is used by RATF as a probability-like signal in the range $[0, 1]$. GPT-Neo is used here because the curvature computation requires an autoregressive language model that exposes token-level log-probabilities.

3.3 RATF: Robust Adaptive Tri-Encoder Framework

Our proposed detection model, RATF, is designed to dynamically integrate information from the three textual streams: the original semantics (A), the content features (A_{con}), and the writing style (A_{sty}). This is achieved through a tri-encoder architecture, a score-conditioned cross-attention fusion module, and a learnable parametric gating mechanism. Both modules are driven by the estimated *rewritten score* (s_{rw}), allowing the model to adaptively calibrate semantic and stylistic evidence according to the estimated degree of rewriting. The overall architecture is depicted in Figure 2.

Tri-Encoder Module. The framework begins with three parallel, independent pre-trained deberta-v3-base encoders: an original semantics encoder (E_{sem}), a content feature encoder (E_{con}), and a style encoder (E_{sty}). Each encoder processes its corresponding tokenized input (A, A_{con}, A_{sty}) and produces both a pooled output vector and token-level hidden states. Let d_h be the hidden dimension of the encoders. The pooled outputs are:

$$\mathbf{h}_{sem} = E_{sem}(A) \in \mathbb{R}^{d_h} \quad (3)$$

$$\mathbf{h}_{con} = E_{con}(A_{con}) \in \mathbb{R}^{d_h} \quad (4)$$

$$\mathbf{h}_{sty} = E_{sty}(A_{sty}) \in \mathbb{R}^{d_h} \quad (5)$$

These vectors are then passed through separate linear projection layers, yielding $\mathbf{p}_{sem}, \mathbf{p}_{con}, \mathbf{p}_{sty} \in \mathbb{R}^{d_h}$. We denote the token-level hidden states of the original semantics and style encoders as $\mathbf{H}_{sem}$ and $\mathbf{H}_{sty}$.

Score-Conditioned Cross-Attention Fusion.

Inspired by recent work on cross-order representation fusion for denoising heterogeneous signals (Wu et al., 2026), we design a fusion module that lets the most reliable feature stream selectively integrate evidence from the others. Content features are designed to preserve the factual core of the article and are therefore treated as the most reliable view under style-confusion attacks. RATF uses the projected content vector as the query and injects the rewritten score as a condition:

$$\mathbf{q} = \text{LN}(\mathbf{p}_{con} + \phi(s_{rw})), \quad (6)$$

where $\phi(\cdot)$ is a small MLP that maps the scalar rewritten score to a d_h -dimensional bias. The conditioned query is then used to attend to token-level semantic and stylistic evidence:

$$\mathbf{h}_{sem}^{ca} = \text{MHA}(\mathbf{q}, \mathbf{H}_{sem}, \mathbf{H}_{sem}), \quad (7)$$

$$\mathbf{h}_{sty}^{ca} = \text{MHA}(\mathbf{q}, \mathbf{H}_{sty}, \mathbf{H}_{sty}). \quad (8)$$

The attention outputs are combined with their corresponding projected pooled representations through residual connections:

$$\mathbf{h}_{sem} = \mathbf{p}_{sem} + \mathbf{h}_{sem}^{ca}, \quad (9)$$

$$\mathbf{h}_{con} = \mathbf{p}_{con}, \quad (10)$$

$$\mathbf{h}_{sty} = \mathbf{p}_{sty} + \mathbf{h}_{sty}^{ca}. \quad (11)$$

This design allows the most reliable content representation to selectively retrieve useful semantic

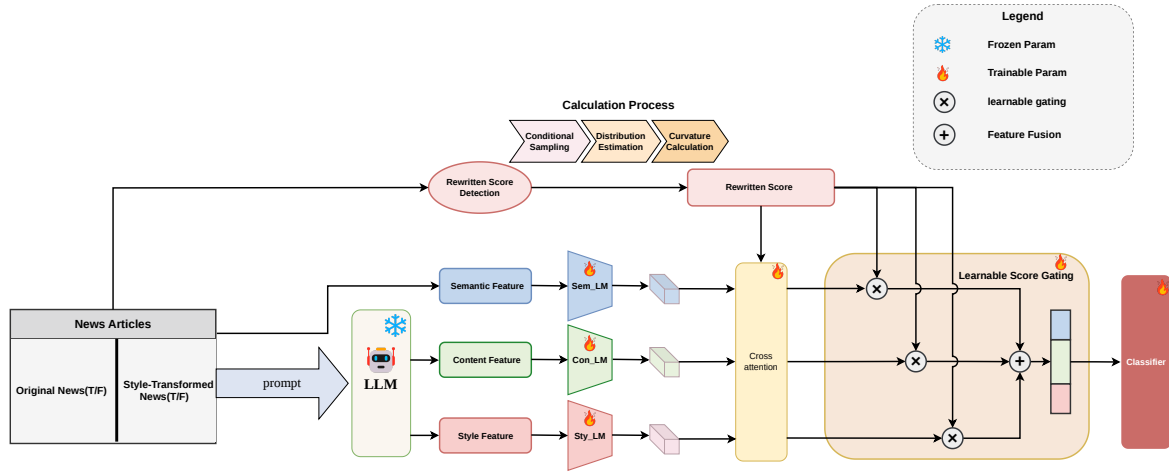


Figure 2: The architecture of the RATF detection model.

and stylistic signals while making the retrieval process aware of the degree of rewriting.

Learnable Score Gating. RATF introduces a learnable parametric gating mechanism to calibrate the contribution of each feature stream. The rewritten score is the only input to the gate, and the attenuation curves are learned end-to-end:

$$w_{sem}(s_{rw}) = \sigma(-k_{sem}(s_{rw} - b_{sem})), \quad (12)$$

$$w_{con}(s_{rw}) = 1, \quad (13)$$

$$w_{sty}(s_{rw}) = \sigma(-k_{sty}(s_{rw} - b_{sty})), \quad (14)$$

where $k_{sem}, k_{sty} > 0$ control the slope of the learned decay curves, $b_{sem}, b_{sty} \in (0, 1)$ determine the decay centers, and $\sigma(\cdot)$ denotes the sigmoid function. These parameters are optimized jointly with the detector. As a result, RATF preserves the central role of the rewritten score while enabling continuous and data-adaptive feature calibration.

Feature Fusion and Training. The calibrated feature vectors are computed as:

$$\hat{\mathbf{h}}_{sem} = w_{sem} \mathbf{h}_{sem}, \quad (15)$$

$$\hat{\mathbf{h}}_{con} = w_{con} \mathbf{h}_{con}, \quad (16)$$

$$\hat{\mathbf{h}}_{sty} = w_{sty} \mathbf{h}_{sty}. \quad (17)$$

They are concatenated ($\mathbf{h}_{merged} = [\hat{\mathbf{h}}_{sem}; \hat{\mathbf{h}}_{con}; \hat{\mathbf{h}}_{sty}]$) and fused via a fully-connected layer with LayerNorm and ReLU. The resulting $\mathbf{h}_{fused}$ is projected to logits $\mathbf{z} \in \mathbb{R}^2$ for classification. We optimize the model using Cross-Entropy loss $\mathcal{L}_{CE}$ on softmax probabilities, selecting the class with the highest probability during inference. For binary classification, this two-logit softmax

formulation is equivalent to a one-logit sigmoid formulation while remaining directly extensible to multi-class veracity prediction.

4 Experiments

In this section, we empirically evaluate our proposed model from several complementary perspectives. We first present the experimental setup (Section 4.1), followed by the main results on style-confusion attack datasets (Section 4.2). We then analyze the adaptability to different language model backbones (Section 4.3) and conduct an ablation study to isolate the contribution of each major component (Section 4.4). Additional analyses, including performance on original datasets, the sensitivity analysis of the rewritten content detector, and encoder representation analysis, are provided in Appendices A, B, and C, respectively.

4.1 Experimental Setup

4.1.1 Datasets

We use PoliFact and GossipCop from FakeNewsNet (Shu et al., 2020), two widely adopted benchmark datasets for fake news detection and for the style-confusion setting studied in (Wu et al., 2024). We follow the temporal split protocol of FakeNewsNet: articles are ordered by time, the held-out test set corresponds to the later portion, and the validation set is further held out from the training portion in the same temporal manner. Following (Wu et al., 2024), we incorporate their style-attacked dataset (four stylistic variations per article) and downsample to create a balanced mix of original and rewritten news (Table 1). The four

Table 1: Statistics of the mixed-style datasets used in our experiments. Each dataset is a 1:1 mix of original articles and their style-rewritten counterparts. All splits are also label-balanced, with a 50:50 ratio of real and fake news; the fake-article counts are shown for train/validation/test.

Dataset	Train	Validation	Test	%Rewritten	Fake
PolitiFact(confusion)	540	180	180	50%	270/90/90
GossipCop(confusion)	9496	3168	3168	50%	4748/1584/1584

Table 2: Main results on the style-confusion attack datasets. Our model (RATF) is benchmarked against LLM-Only (G1), SLM-Only (G2), and LLM+SLM (G3) baselines. The highest scores are in **bold**.

Method		PolitiFact(confusion)				GossipCop(confusion)			
		Acc	F1	Pre	Rec	Acc	F1	Pre	Rec
G1	Gemini 2.5 Flash	0.6638	0.6852	0.6476	0.7083	0.6039	0.6023	0.5803	0.6192
	GPT5.4	0.7149	0.7124	0.7155	0.7129	0.6353	0.6261	0.6292	0.6337
G2	RoBERTa	0.7528	0.7634	0.7030	0.8353	0.7361	0.7409	0.7207	0.7621
	DeBERTa	0.7921	0.7702	0.8158	0.7249	0.7554	0.7523	0.7540	0.7506
	PET	0.7536	0.7651	0.7124	0.7847	0.7374	0.7362	0.7343	0.7385
G3	Sheepdog	0.8258	0.8238	0.8380	0.8258	0.7749	0.7747	0.7753	0.7748
	ARG	0.8467	0.8465	0.8475	0.8464	0.7456	0.7657	0.7726	0.7641
	TED	0.8378	0.8379	0.8382	0.8376	0.7538	0.7540	0.7659	0.7636
Ours (RATF)		0.8820	0.8823	0.8820	0.8819	0.7813	0.7816	0.7811	0.7807

stylistic variants cover common manipulation dimensions such as sensational, objective, partisan, and other style transfers, and stronger real-world attacks can often be viewed as combinations of these dimensions.

4.1.2 Evaluation Metrics

We employ four standard metrics: **Accuracy (Acc)**, **Precision (Pre)**, **Recall (Rec)**, and **Weighted F1-score (F1)**. These metrics provide a comprehensive view of model performance, balancing overall correctness with the ability to identify fake news accurately and completely.

4.1.3 Baselines

We benchmark our model against eight representative baseline methods, which are categorized into three groups to align with our results tables. To ensure a fair comparison, all trainable baselines are retrained or fine-tuned on the same training splits as RATF. We tune key hyperparameters, including learning rate, batch size, training epochs, and model-specific settings, on the validation set and report the test results from the best validation checkpoint. For LLM-only baselines, we use the validation set to select the prompt template with-

out accessing the test labels.

- **G1 (LLM-Only)**: This group consists of large language models performing zero-shot veracity prediction. We evaluate **Gemini 2.5 Flash** and **GPT5.4**.
- **G2 (Fine-tuned SLMs)**: This category includes pre-trained language models fine-tuned for the detection task (Zhang et al., 2026). We include **RoBERTa** (Liu et al., 2019), **DeBERTa** (He et al., 2020), and **PET** (Schick and Schütze, 2021), which converts textual inputs into cloze questions that contain a task description.
- **G3 (LLM-Enhanced SLMs)**: These models employ architectures or frameworks that integrate LLM capabilities and are specifically tailored for the fake news detection task. We include **Sheepdog** (Wu et al., 2024), **ARG** (Hu et al., 2024), and **TED** (Liu et al., 2025). **Sheepdog** is a style-agnostic framework designed to resist style-confusion attacks. It uses an LLM to generate multiple “reframings” of a news article in different styles (e.g., sensational, objective) and then trains the detector to produce consistent veracity predictions across these stylis-

tically diverse versions, forcing it to focus on content rather than style. **ARG** treats an LLM as an advisor to a smaller detection model. It prompts the LLM to generate multi-perspective rationales (e.g., from textual and commonsense standpoints). The detector then analyzes both the news and these rationales, using a news-rationale collaboration module to adaptively integrate the most useful insights before making a final prediction. **TED** is a multi-agent framework that simulates a formal debate to determine news veracity. LLM-powered agents are assigned proponent and opponent roles to argue for and against the article’s truthfulness. An Analysis Agent then models the debate’s structure and interactions to make a final, informed judgment.

4.1.4 Implementation Details

Our model, RATF, uses `deberta-v3-base` for its three encoders. The content and style features are extracted using the `Qwen2.5-7B-Instruct` model. The core of our model is a score-conditioned fusion strategy consisting of cross-attention fusion and learnable parametric gating driven by the `rewritten_score`. We trained our model with a batch size of 32. We used the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01, coupled with a linear scheduler with no warmup steps. We use the checkpoint with the best validation performance for evaluation. For all baseline models, we retrained or fine-tuned them on the same training splits of `PolitiFact(confusion)` and `GossipCop(confusion)` used by RATF.

4.2 Main Results on style-confusion attack datasets

We evaluate the performance of our proposed model against the established baselines on our mixed-style attack datasets. Table 2 presents the comparative results. The findings indicate that our model, RATF, surpasses all baseline methods. While standard models like RoBERTa and DeBERTa achieve reasonable performance, their efficacy diminishes when confronted with stylistically manipulated articles. Advanced frameworks like ARG and TED are also susceptible, as they lack an explicit mechanism to first assess and then adapt to potentially manipulative writing styles; ARG’s rationales can be skewed by deceptive prose, while TED’s debate can be misguided by a convincingly styled but false narrative. Conversely, style-

invariant defenses like Sheepdog are robust but face an inherent performance cap. They achieve this robustness by disregarding stylistic features entirely. However, this comes at the cost of lower performance on non-attacked articles, where style can be a valuable indicator. Our model addresses this central challenge—risking vulnerability by trusting style or sacrificing performance by ignoring it. By leveraging score-conditioned cross-attention and learnable score gating, it navigates these challenging scenarios, thereby achieving higher overall performance and highlighting the practical utility of our approach. The stronger relative gains on `PolitiFact` are consistent with its longer and more style-sensitive political claims, whereas `GossipCop` contains shorter and more formulaic entertainment news; we provide further discussion in Appendix A.

The initialization of the learnable score-gating parameters is reported in Appendix E.

4.3 Adaptability to Different LM Backbones

To assess the flexibility of RATF, we evaluate its performance when combined with three representative LM backbones: RoBERTa, BERT, and DeBERTa. This analysis is crucial to demonstrate that our framework’s adaptive mechanism is not tied to a specific encoder architecture but can provide improvements across various models.

As demonstrated in Table 4, RATF substantially enhances the performance of each respective LM backbone. This adaptability highlights RATF’s robustness from score-conditioned fusion and learnable feature calibration, implying its practical utility in real-world scenarios where different LMs may be preferred or more readily available. The auxiliary feature extraction LLM can also be upgraded independently: stronger LLMs are expected to produce higher-quality content and style descriptions, while RATF itself remains the same adaptive fusion framework.

4.4 Ablation Study

To investigate the contribution of each of our model’s components, we performed an ablation study. We deconstructed RATF to assess each component’s contribution to the final result by creating several simpler versions: **RATF w/o Rewritten Score** (which removes the rewritten score signal), **SC-CAF Only** (which uses only score-conditioned cross-attention), **Learnable Gate Only** (which uses only the learnable

Table 3: Ablation study results on the test sets.

Method	PolitiFact(confusion)				GossipCop(confusion)			
	Acc	F1	Pre	Rec	Acc	F1	Pre	Rec
Ours (RATF)	0.8820	0.8823	0.8820	0.8819	0.7813	0.7816	0.7811	0.7807
RATF w/o Rewritten Score	0.8487	0.8464	0.8482	0.8453	0.7526	0.7524	0.7532	0.7521
SC-CAF Only	0.8743	0.8754	0.8748	0.8759	0.7566	0.7561	0.7567	0.7558
Learnable Gate Only	0.8516	0.8507	0.8511	0.8502	0.7604	0.7615	0.7626	0.7612
Content-Only	0.8315	0.8308	0.8335	0.8315	0.7431	0.7429	0.7441	0.7431
Style-Only	0.7753	0.7744	0.7862	0.7753	0.7052	0.7044	0.7067	0.7052

Table 4: Performance on different LM backbones. RATF demonstrates stable and significant improvements (in Weighted-F1 %).

Method	PolitiFact(confusion)	GossipCop(confusion)
RoBERTa	76.34	74.09
RATF-RoBERTa	87.85	77.94
BERT	75.36	73.11
RATF-BERT	86.31	75.68
DeBERTa	77.02	75.23
RATF-DeBERTa	86.72	77.89

score-gating mechanism), and two basic models that use only one type of feature: **Content-Only** and **Style-Only**. The results in Table 3 give us some clear answers.

- **RATF w/o Rewritten Score:** Removing the rewritten score tests whether explicit rewriting likelihood is necessary for adaptive fusion.
- **SC-CAF Only:** This variant keeps score-conditioned cross-attention but removes the learnable gate, isolating the effect of attention-based feature interaction.
- **Learnable Gate Only:** This variant keeps the learnable score-gating mechanism but removes SC-CAF, isolating the effect of direct score-driven feature calibration.
- **Content-Only:** This model, which only looks at the content features, performs very well on its own. This proves that our feature extraction process using the LLM is effective. Still, it is not as good as the full model, which means the semantic and style features, when used carefully, add valuable information.
- **Style-Only:** As expected, this model performs very poorly. Style alone is not a reliable way

to detect fake news, particularly when it can be easily attacked.

The results, summarized in Table 3, provide clear insights into the architectural design. The performance drops observed across these variants underscore the importance of adaptive score-aware fusion and calibration, validating our design choices.

5 Conclusion

In this paper, we addressed the critical challenge of detecting fake news in an era of sophisticated style-confusion attacks from Large Language Models. We introduced RATF, a Robust Adaptive Tri-Encoder framework that enhances detection robustness by disentangling news into original semantics, content features, and style features. The core of our contribution is to use the rewritten score as a continuous condition for cross-attention fusion and learnable parametric gating. This allows our model to dynamically calibrate stylistic and semantic evidence, anchoring its predictions in factual content when an attack is suspected and leveraging valuable style cues when the text is authentic. Our experiments confirm that RATF not only achieves state-of-the-art performance against such attacks but also adapts effectively across different language model backbones.

Limitations

This work has several limitations. First, RATF relies on external LMs to extract content and style features and to compute the rewritten score, which introduces additional computational cost compared with purely text-only detectors. This cost is acceptable for many high-stakes applications, such as batch fact-checking and content moderation queues, but it may be less suitable for

latency-critical deployment. Second, our empirical evaluation is conducted on PolitiFact and GossipCop. Although these are widely used benchmarks, they cover a limited range of domains and style-confusion attack patterns, so the cross-dataset generalizability of RATF has not yet been established at scale.

These limitations point to several directions for future work. A natural next step is a larger-scale evaluation spanning more datasets and a wider variety of rewriting attacks. Beyond broader evaluation, we plan to incorporate external knowledge and evidence retrieval to strengthen factual grounding in complex real-world news scenarios, and to investigate retrieval-augmented generation (RAG) techniques (Asai et al., 2024) for richer contextual evidence in veracity prediction. Finally, extending RATF to multilingual and multimodal fake news detection is another important direction.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2023YFC3306201.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Waqas Ahmad and Chen Hongsong. 2026. From detection to detoxification: Challenges and solutions for fake news and cyberbullying language governance in the era of large language models. *Engineering Applications of Artificial Intelligence*, 178:115115.
- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pages 24814–24836.
- Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *Zenodo*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Canyu Chen and Kai Shu. 2024a. Can llm-generated misinformation be detected? In *International Conference on Learning Representations*, volume 2024, pages 34687–34726.
- Canyu Chen and Kai Shu. 2024b. Combating misinformation in the age of llms: Opportunities and challenges. *AI magazine*, 45(3):354–368.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- 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*, pages 575–584.
- Jacqueline He, Howard Yen, Margaret Li, Stella Li, Zhiyuan Zeng, Weijia Shi, Yulia Tsvetkov, Danqi Chen, Pang Wei Koh, and Luke Zettlemoyer. 2026. Precise information control in long-form text generation. *Advances in Neural Information Processing Systems*, 38:92727–92792.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.
- Bo Hu, Zhendong Mao, and Yongdong Zhang. 2025. An overview of fake news detection: From a new perspective. *Fundamental research*, 5(1):332–346.
- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.
- Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking

- fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14571–14589.
- Tarek Ghazi Kanan and Marwa Hamza. 2026. Ai-driven fake news detection using multimodal features and hybrid deep learning models. *International Journal of Advances in Soft Computing and its Applications*, 18(1):382–395.
- Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.
- Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the web conference 2021*, pages 3432–3441.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3391–3401.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylistometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 231–240.
- Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497.
- U Sakthivelu and CNS Vinoth Kumar. 2023. Advanced persistent threat detection and mitigation using machine learning model. *Intelligent Automation & Soft Computing*, 36(3).
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 255–269.
- Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 4543–4556.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- Haoran Wang, Yingdong Dou, Canyu Chen, Lichao Sun, Philip S Yu, and Kai Shu. 2023. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023*, pages 3978–3986.
- Haoze Wu, Jian Ding, Chenghui Yu, Bingfeng Deng, Hongyu Xiong, and Kun Xu. 2026. Corf-net: Cross-order representation fusion network for denoised survey response modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 8214–8224.
- 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*, pages 3367–3378.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. 2026. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36.

Table 5: Performance (%) on the original, non-attacked test sets. Our model is benchmarked against LLM-Only (G1), SLM-Only (G2), and LLM+SLM (G3) baselines.

Method		PolitiFact		GossipCop	
		Acc	F1	Acc	F1
G1	Gemini 2.5 Flash	70.23	70.15	65.71	64.74
	GPT5.4	75.21	75.27	71.11	71.15
G2	RoBERTa	88.00	87.40	74.94	74.85
	DeBERTa	86.33	86.30	74.68	74.62
	PET	85.56	85.51	74.75	74.63
G3	Sheepdog	88.44	88.39	75.77	75.75
	ARG	88.89	88.80	75.88	75.85
	TED	90.25	90.37	76.12	76.17
Ours (RATF)		92.17	92.26	76.51	76.51

A Performance on Original Datasets

We assess our model’s performance on the original, unperturbed PolitiFact and GossipCop test sets. This experiment is crucial to ensure that our defense mechanism against style attacks does not compromise the model’s effectiveness on clean data. The results, presented in Table 5, demonstrate that our model remains highly competitive. This suggests that the adaptive nature of our architecture is beneficial not only for adversarial robustness but also for general-purpose detection, affirming its utility as a versatile and effective fake news detector.

The difference between the two datasets is also informative. PolitiFact contains fewer but longer and more style-sensitive political claims, so style manipulation creates a larger shift between the mixed-style attack setting and the original clean setting. GossipCop is much larger and consists of shorter, more formulaic celebrity and entertainment news, where the distribution shift introduced by style rewriting is smaller. This pattern also explains why RATF’s adaptive gating provides a larger relative advantage on PolitiFact, where the style-content trade-off is more pronounced.

B Sensitivity Analysis of the Rewritten Content Detector

We evaluate the effectiveness of our rewritten content detection module. Figure 3 visualizes the distribution of the *curvature-based detector score* for original and rewritten articles from the GossipCop test set. A distinct separation is evident: original articles (blue curve) score low, while all four

adversarial versions score significantly higher. In RATF, this detector score is subsequently normalized into the rewritten score $s_{rw} \in [0, 1]$, which is used as a continuous conditioning signal for both cross-attention fusion and learnable feature gating.

We also follow the temperature setting of (Wu et al., 2024). Higher rewriting temperatures may increase surface diversity, but they also introduce a practical trade-off: excessive temperature can cause semantic drift and invalidate the original veracity label, while very low temperature yields insufficient stylistic diversity and weakens the attack. Therefore, the adopted temperature setting represents a realistic balance for preserving content while altering writing style.

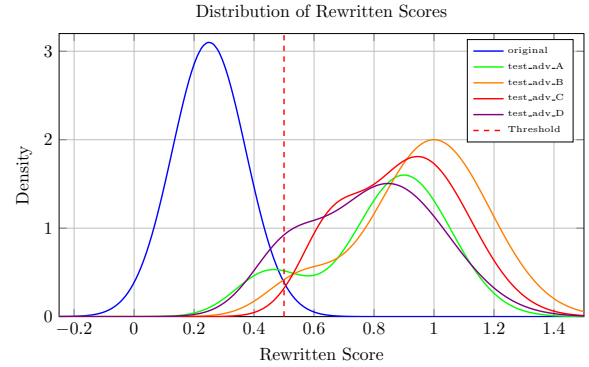


Figure 3: Distribution of curvature-based detector scores for original and rewritten articles on the GossipCop dataset. The clear separation between the scores for the original semantics (blue curve) and the rewritten texts demonstrates the effectiveness of our detection module.

C Encoder Representation Analysis

Our model’s architecture is predicated on the hypothesis that the three parallel encoders learn distinct, non-overlapping representations. We use t-SNE (Maaten and Hinton, 2008) to qualitatively investigate whether the original semantics, content features, and style encoders learn disentangled feature spaces. The t-SNE plot in Figure 4 projects the high-dimensional encoder outputs into a two-dimensional space and shows three distinct, well-separated clusters corresponding to the original semantics, content features, and style encoders. This visual separation provides evidence that each encoder specializes in its designated feature type, supporting the premise of our adaptive gating mechanism.

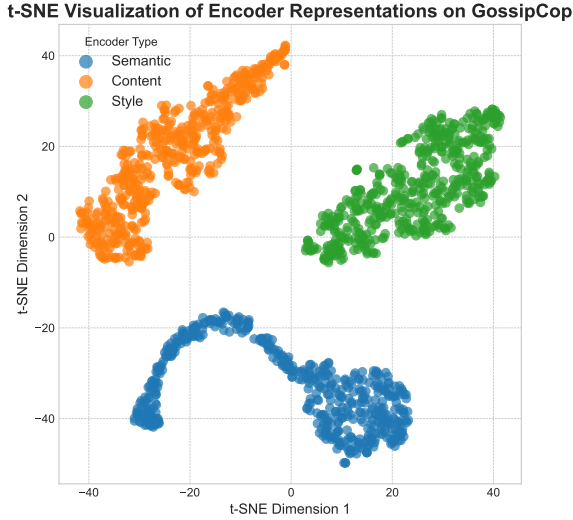


Figure 4: t-SNE visualization of the semantic, content, and style encoder output representations on the GossipCop dataset. The clear separation demonstrates that the encoders learn distinct, disentangled features.

D Prompt Details

I need you to analyze the following news article.
Your task is to:

1. Identify and describe the stylistic features of this news article (writing style, tone, word choice, etc.)
2. Rewrite the news in an objective descriptive style to eliminate its stylistic features.

News article:
“{news_text}”

Return your analysis in exactly this format:
Content features: [objective rewritten news]
Stylistic features: [description of stylistic elements]

Figure 5: Prompt used to extract content and stylistic features for each news article.

E Learnable Gate Initialization

The score gate is parameterized by learnable decay slopes and centers, which are optimized jointly with the detector. Table 6 lists the initialization values used in our experiments. These values provide a stable starting point but do not remain fixed during training.

Table 6: Initialization of the learnable score-gating parameters.

Gate	Initial slope	Initial center
Semantic gate w_{sem}	$k_{sem} = 4.0$	$b_{sem} = 0.6$
Style gate w_{sty}	$k_{sty} = 15.0$	$b_{sty} = 0.5$

F Robustness to the Choice of LLM and Inference Efficiency

Robustness to the Choice of LLM. By design, RATF is not tied to the specific LLM used for feature extraction or attack generation. The auxiliary LLM interacts with our framework only through natural language: given a fixed prompt (Appendix D), it returns a content rewrite and a style description that are re-encoded by the tri-encoder. Since these features are passed as text rather than model-specific embeddings, the extraction LLM can be replaced with any general-purpose, instruction-following LLM without re-training, as the task is a generic summarization-style capability rather than a behavior unique to Qwen2.5-7B-Instruct. Attack generation is likewise model-agnostic: RATF relies on the rewritten score, a distributional property (conditional probability curvature) of the text itself, rather than artifacts of any particular rewriting model. The framework is therefore expected to remain effective when the extraction or rewriting model changes.

Inference Efficiency. The main cost that RATF adds over a standard encoder-based detector is the external LLM used for feature extraction. This is a single-pass, summarization-style call per article—restating the article in an objective tone and describing its stylistic elements within one prompt (Appendix D)—and involves neither iterative reasoning nor multi-round interaction. It is run once as offline preprocessing rather than repeatedly at query time. In contrast to methods that query an LLM for multi-perspective rationales at inference (e.g., ARG), RATF keeps the per-article LLM cost to a single generation pass.