Synergizing Stylometrics with Semantics: Dual-Path Framework for LLM Detection and Attribution

Xingyu Lu[1] **Yumeng Ma**[1] **Xiang Zhou**[1] **Shengli Gan**[1] **Guiying Deng**[1]
Yang Wen[1] **Yanbing Liu**[1,2*]

[1]Chongqing Key Laboratory of Image Cognition,
School of Artificial Intelligence and School of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]School of Medical Information, Chongqing Medical University, Chongqing 400016, China
Correspondence: liuyb@cqupt.edu.cn

Abstract

The widespread application of Large Language Models (LLMs) has made Machine-Generated Text (MGT) detection increasingly important in cyberspace security and governance. The existing detection paradigms mainly focus on statistical likelihood or deep embeddings. However, in complex applications such as short texts, derivative works, and cross-domain content, the discriminative capabilities fragility of these conventional methods increases significantly with the development of LLMs. Conversely, our research reveals that LLMs exhibit inherent style inertia. To address these limitations, this study attempts to synergize stylometrics and semantics for identifying MGT. This approach draws from the forensic perspective of experts who detect human imitation by focusing on stylistic nuances. Based on the above inspiration, we propose Stylometric-Semantic LLM Attribution (SSLA), a framework that extracts model-specific stylistic fingerprints across lexical, syntactic, and structural dimensions. SSLA employs a dual-path attention fusion architecture to dynamically integrate explicit stylistic signals with implicit semantic encodings. Extensive experiments across six LLM families demonstrate that our method achieves state-of-the-art performance. Notably, SSLA achieves a Macro-F1 score of 95.6% on the challenging Wikipedia dataset, demonstrating exceptional robustness and surpassing state-of-the-art baselines like OTBDetector.

1 Introduction

Large language models (LLMs) have rapidly become the dominant paradigm in text generation, powering applications from summarization and rewriting to conversational assistance (He et al., 2023). As machine-generated text (MGT) becomes pervasive, the ability to attribute a piece of text to its generating model—or determine whether it was produced by a human—has emerged as a critical research challenge (Li et al., 2023a; Abburi et al., 2025; Kumarage et al., 2024).

Existing attribution methods largely rely on statistical likelihood signals (e.g., perplexity scores, curvature) (Mitchell et al., 2023; Tang et al., 2023) or supervised semantic embeddings derived from fine-tuned classifiers. While effective in controlled settings, statistical signals exhibit inherent instability in complex scenarios such as short-text or cross-domain transfers, while semantic models often suffer from heavy data dependency and opaque decision-making. Consequently, relying solely on these cues is insufficient to distinguish model-specific generative characteristics. This limitation stems partly from the underutilization of stylistic nuances. Recent studies suggest that, analogous to human writers, LLMs demonstrate stable model-specific "style inertia" independent of topic (Bitton et al., 2025). However, such deep, multi-layered stylistic structures remain insufficiently leveraged in current attribution paradigms, which typically treat style merely as an auxiliary or shallow feature (Kumarage and Liu, 2023a; Posadas-Durán et al., 2025; Wu et al., 2025a).

The core intuition behind our study is illustrated in Figure 1. When distinguishing an original author (e.g., Ernest Hemingway) from a skilled mimic—whether a human fan or an LLM—relying solely on surface content is often insufficient. Instead, a robust attribution process should parallel a human expert's multilevel text analysis, evaluating text across lexical patterns, syntactic structures, and semantic intentions. Combining traditional stylometrics with deep learning-based semantic models has been established as a highly effective and mature paradigm in literary authorship attribution (korić et al., 2022), inspiring our extension of this approach to machine-generated text detection. To investigate whether LLMs indeed leave behind detectable traces across these layers, we conducted

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 37252–37265

an exploratory analysis comparing human-written sentences and their counterparts generated by multiple LLM families under identical writing prompts. The results reveal consistent and model-dependent deviations that emerge even when semantic content is preserved. These deviations form structured clusters across three core stylistic dimensions: (1) lexical style, characterized by model-specific vocabulary preferences and function-word distributions; (2) syntactic structure, revealing latent habits in dependency organization; and (3) semantic intention, manifesting as subtle variations in abstraction and framing. By synthesizing these layers, we can construct a robust "style fingerprint" that transcends simple semantic cues.

This observation led to the key intuition of our work: LLMs leave persistent multi-level "style fingerprints" that remain stable under content constraint and are thus indicative of their model provenance (Wu et al., 2025b; Yu et al., 2024).

Based on this insight, we propose Stylometric-Semantic LLM Attribution (SSLA), a framework that reformulates attribution as recognizing a model's stylistic signature. SSLA operationalizes the three stylistic dimensions using a combination of contrastive stylistic analysis, multi-granularity similarity metrics, and interpretable syntax-semantic n-grams (SN-Grams) (Posadas-Durán et al., 2025). To integrate these explicit stylistic features with contextual embeddings from RoBERTa (Liu et al., 2019), we design a dual-path fusion architecture that effectively synthesizes explicit stylistic signals with implicit semantic representations.

Our contributions are threefold:

- We propose SSLA, a modular attribution framework that synergizes explicit multi-level stylistic signals (lexical, syntactic, and structural) with implicit semantic representations via a dual-path fusion architecture to achieve robust fine-grained attribution.
- We discover and formalize multi-level stylistic fingerprints of LLMs, grounded in empirical observations derived from controlled prompt-based generation.
- We perform comprehensive experiments across six LLM families. Results demonstrate that stylistic profiling significantly outperforms state-of-the-art baselines. Notably, SSLA achieves a Macro-F1 score of 95.6% on the challenging Wikipedia dataset, showing exceptional robustness in cross-domain and short-text scenarios while providing interpretable feature-level evidence.

2 Related Work

Fine-grained attribution of machine-generated text (MGT) has become a critical factor limiting the effectiveness of cyberspace governance and forensic analysis (Wu et al., 2025a). To address this challenge, existing research attempts to distinguish text sources by exploiting various discriminative signals ranging from surface statistics to deep representations. Methodologically, related studies can be broadly categorized into three paradigms: (1) profiling probabilistic biases, (2) analyzing explicit linguistic fingerprints, and (3) learning implicit semantic representations. While advancements in these areas have established strong baselines, they often treat stylistic form and semantic content in isolation, failing to capture the cognitive signature of the generator as a unified whole.

Probabilistic and linguistic methods operate on the surface level, relying on unstable statistical cues or coarse handcrafted features that lack deep structural anchoring (Kumarage and Liu, 2023b); conversely, semantic methods leverage deep neural networks to learn decision boundaries but often function as opaque "black boxes". Our proposed SSLA framework integrates the strengths of these paradigms, bridging the gap between interpretability and performance. By dynamically fusing explicit multi-level stylistic fingerprints with implicit deep semantic encodings, SSLA constructs a robust "cognitive profile" for each model. This approach fundamentally enhances attribution reliability in short-text and cross-domain scenarios, rather than relying solely on brittle likelihoods or unexplainable embeddings.

2.1 Zero-Shot Probabilistic Signals

This paradigm posits that an LLM leaves a unique probabilistic signature on its generated text. The core assumption is that a text sequence will yield the lowest perplexity or highest curvature when evaluated by its source model compared to others (Wu et al., 2025a). Pioneering techniques like DetectGPT (Mitchell et al., 2023) utilize the curvature of the log probability function to distinguish source models based on their specific likelihood landscapes. Similarly, methods leveraging log-rank

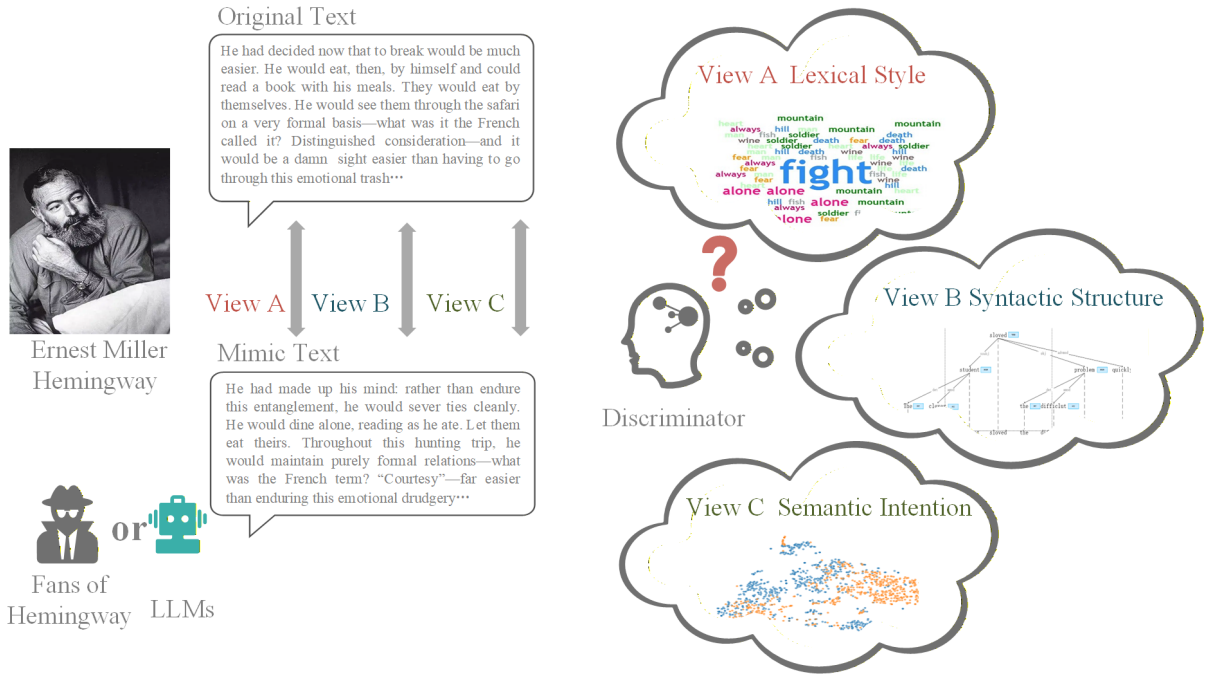


Figure 1: The core intuition of SSLA: Cognitive Attribution. Just as humans distinguish an original author (e.g., Hemingway) from mimics via multi-dimensional stylometric cues, our framework identifies LLMs based on (A) Lexical Style, (B) Syntactic Structure, and (C) Semantic Intention. These stylistic fingerprints remain distinguishable even when the semantic content is similar.

signals (Su et al., 2023) and Uniform Information Density (UID) (Venkatraman et al., 2023) identify the source by analyzing whether the token selection aligns with a specific model's vocabulary ranking preferences. However, these likelihood-based signals face critical applicability barriers. First, they require white-box access to calculate probabilities for every candidate model, which is impossible for proprietary API-based models (Bao et al., 2023). Second, in short-text scenarios, the statistical differences between models become indistinguishable due to high variance (Li et al., 2023b), limiting their effectiveness for fine-grained sourcing.

2.2 Explicit Linguistic Fingerprints

The search for content-invariant signals leads to Stylometry, which utilizes stable linguistic patterns to profile generators. This approach assumes that different models exhibit distinct "style inertia"—unique habits in lexical choice or syntactic structuring that persist across topics (Bitton et al., 2025). Early efforts relied on macro-level statistics, such as POS tag distributions and lexical diversity, to differentiate model families (Li et al., 2023a). To capture deeper signatures, recent research has employed Syntactic N-Grams (SN-Grams) (Sidorov et al., 2014; Posadas-Durán et al., 2025), which traverse dependency trees to reveal non-linear structural habits specific to certain architectures. Nevertheless, despite their interpretability, these traditional features often prove too coarse to distinguish between sophisticated modern LLMs that mimic human nuances closely. Furthermore, most prior works treat these features in isolation, failing to capture how stylistic habits interact with semantic contexts for fine-grained attribution (Kumarage et al., 2024).

2.3 Deep Semantic Representations

In parallel with statistical and stylometric approaches, the dominant paradigm treats attribution as a multi-class supervised classification task. These methods leverage the deep semantic understanding of Pre-trained Language Models (PLMs), such as RoBERTa or BERT, to learn discriminative decision boundaries between different source models (Liu et al., 2019; He et al., 2023). By fine-tuning on labeled datasets, these classifiers can capture subtle semantic deviations specific to each generator. To improve robustness against domain shifts, recent state-of-the-art methods like ConDA(Bhattacharjee et al., 2023) and OTB-D(La Cava and Tagarelli, 2025) have incorporated contrastive adaptation and optimal transport geom-

etry to align feature spaces across different distributions (He et al., 2023). However, these semantic classifiers suffer from critical flaws regarding data inefficiency—requiring massive labeled data for each new model to prevent overfitting—despite achieving high accuracy. Furthermore, they function largely as "black boxes" and lack the interpretability required to explain why a text is attributed to a specific model, highlighting the need for a framework that synergizes interpretability with performance.

3 Methodology

In this section, we introduce SSLA, a framework designed to attribute texts to their source LLMs by profiling stable stylistic fingerprints. We outline the problem formulation and our core motivation in Section 3.1. We then elaborate on the rationale and construction of our multi-dimensional stylistic features in Section 3.2, and finally describe the dual-path interaction architecture in Section 3.3.

3.1 Problem Definition and Motivation

Problem Formulation. We define fine-grained LLM attribution as a multi-class classification task. Given a query text x, the goal is to predict the source generator $y \in \mathcal{Y} = \{y_1, ..., y_N\}$ (e.g.,Claude, GPT-turbo, Human). Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$ be the dataset. We aim to learn a mapping $\mathcal{F}_\theta : \mathcal{X} \to [0, 1]^N$ that minimizes the prediction error.

Motivation: The Stylistic Stability Hypothesis. Existing detectors often rely on brittle signals: semantic keywords (which change with topics) or token probabilities (which are inaccessible for black-box APIs). Our approach is grounded in the Stylistic Stability Hypothesis. Recent literature has demonstrated that LLMs possess distinct and consistent stylistic fingerprints that persist even when they are prompted to write in different writing styles (Bitton et al., 2025). While the generated content varies significantly across topics, its stylistic signature—manifested in syntactic complexity, lexical preferences, and structural rigidity—remains consistent across domains. SSLA mimics a forensic linguistic approach: (1) extracting invariant stylistic fingerprints (s, g) that persist despite topic shifts; and (2) interacting these signals with semantic context via a self-attention fusion mechanism to achieve robust attribution.

3.2 Multi-Dimensional Stylometric Profiling

We construct a multi-dimensional feature vector designed to capture the "cognitive signature" of the generator. We categorize these features into static linguistic complexity, dynamic stylistic rigidity, and structural preferences.

Static Complexity: Linguistic Metrics (s_{ling}). LLMs, constrained by their training objectives, specifically Reinforcement Learning from Human Feedback (RLHF), often exhibit specific statistical biases. We extract a 6-dimensional vector to quantify the information density and syntactic elaboration:

- Syntactic Structure: We compute Average Dependency Depth and Average Dependency Distance using dependency parsing. These metrics reveal whether a model prefers simple, flat structures or complex, nested recursions.
- Lexical & Part-of-Speech Distribution: We calculate Lexical Diversity, measured by Type-Token Ratio (TTR), and Sentence Length to measure vocabulary richness. Additionally, we explicitly monitor part-of-speech preferences via Noun Ratio and Verb Ratio, capturing the generator's tendency towards descriptive (noun-heavy) or action-oriented (verb-heavy) phrasing.

Dynamic Rigidity: Differential Analysis via Rewriting (s_{pres}). Static features may fail when models mimic human styles. To address this, we introduce a novel Differential Stylistic Analysis. The core idea is to employ a general Large Language Model as a rewriting probe to stress-test the input x by generating a semantic-preserving rewrite x'. We measure the "stylistic drift" between x and x' using a 6-dimensional vector covering semantic, embedding, and surface levels:

$$s_{pres} = [\mathcal{M}_{sem}; \mathcal{M}_{emb}; \mathcal{M}_{surf}] \in R^6 \quad (1)$$

where $[;]$ denotes the concatenation operator. Specifically, each of the three components is formulated as a 2-dimensional sub-vector:

- $\mathcal{M}_{sem} \in R^2$ (Semantic Consistency): We use BLEURT and BERTScore to capture deep semantic preservation;
- $\mathcal{M}_{emb} \in R^2$ (Latent Alignment): We compute the Cosine Similarity of Sentence-BERT (SBERT) and RoBERTa embeddings, measuring the shift in high-dimensional contextual space;

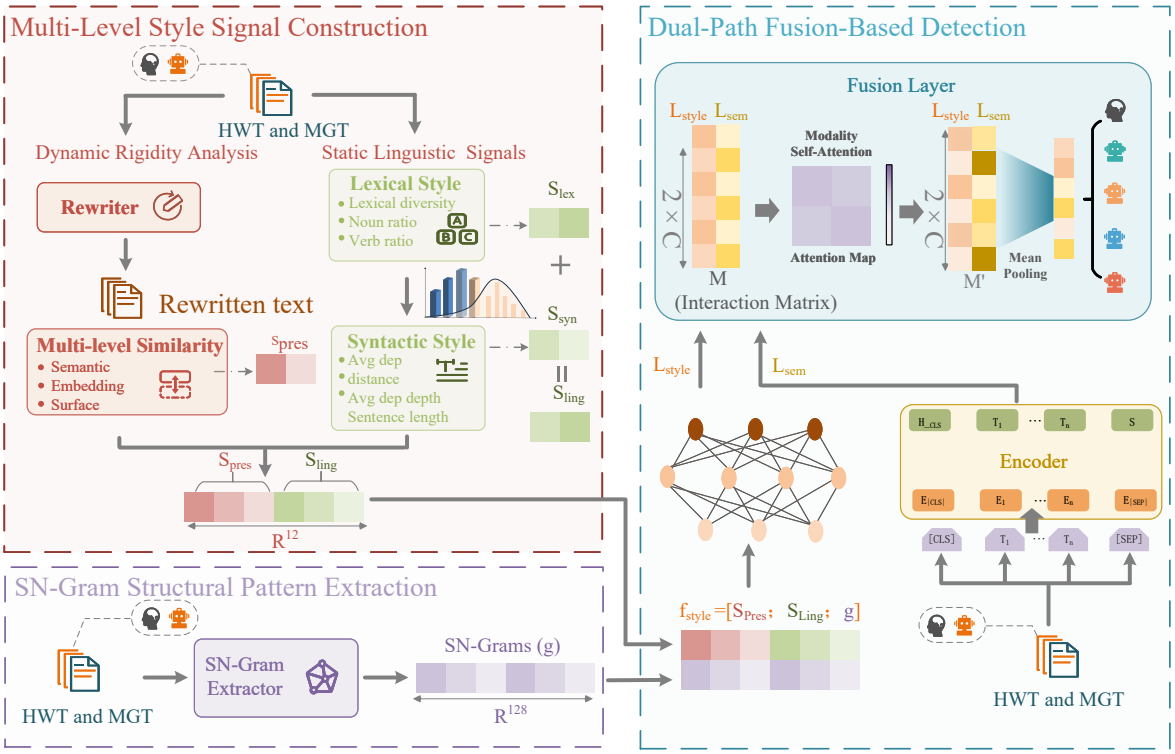


Figure 2: The overall architecture of SSLA. The framework constructs multi-level stylometric signals (Left) and synergizes them with RoBERTa embeddings via a dual-path attention fusion (Right) for robust attribution.

- $\mathcal{M}_{surf} \in R^2$ (Surface Similarity): We utilize CHRF++ and Surface Cosine Similarity to quantify character-level and lexical retention.

Rationale: This differential vector acts as a measure of Stylistic Rigidity. If x is human-written, the standardized rewriting process induces a significant drift (low similarity scores); if x is generated by an aligned LLM, the style remains rigid (high similarity), serving as a powerful discriminative signal.

Structural Preferences: SN-Grams (g). To capture long-range dependencies, we employ Syntax-Semantic N-Grams. We extract dependency triples defined as $(Relation, Head, Dependent)$ and vectorize these patterns using TF-IDF weighting with variance-based selection, resulting in a dense vector $g \in R^{128}$.

The final stylistic input is the concatenation: $f_{style} = [s_{ling}; s_{pres}; g] \in R^{140}$.

3.3 Dual-Path Interaction and Optimization

We propose a Dual-Path architecture that fuses the interpretability of stylistic profiles with the semantic power of PLMs, utilizing an attention-based mechanism to dynamically synthesize explicit and implicit signals.

Dual-Encoder Architecture. The framework consists of two parallel encoders: (1) Style Path: A Multi-Layer Perceptron (MLP) projects the handcrafted feature vector f_{style} to logits $L_{style} \in R^C$. (2) Semantic Path: A pre-trained RoBERTa encoder projects the text embedding to logits $L_{sem} \in R^C$, where C is the number of LLM classes.

Self-Attention Fusion. We stack the projected logits to form $H = [L_{style}; L_{sem}] \in R^{2\times C}$ and apply Multi-Head Self-Attention (MHSA) to capture non-linear interactions. We simplify the attention operation as:

$$\tilde{H} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V are obtained by projecting H with learnable matrices $W_Q, W_K, W_V \in R^{C\times d_{model}}$, and d_k is the scaling factor derived from the head dimension. This layer weighs the confidence of structural vs. semantic cues. The final prediction $\hat{y}$ is obtained via mean pooling and a linear projection parameterized by $W_o \in R^{C\times C}$ and $b_o \in R^C$:

$$\hat{y} = W_o \cdot \text{Mean}(\tilde{H}) + b_o \tag{3}$$

Optimization. We employ a joint optimization strategy. To ensure both paths learn discrimina-

tive features independently while synergizing effectively, the total loss $\mathcal{L}_{total}$ is computed as a weighted combination of the fused output loss and the auxiliary losses from both paths:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_f \mathcal{L}(\hat{y}, y) + \lambda_s \mathcal{L}(\hat{y}_s, y) \\ &+ \lambda_m \mathcal{L}(\hat{y}_m, y)\end{aligned} \tag{4}$$

where y is the ground-truth label, $\hat{y}$ is the final fused prediction, and $\hat{y}_s$ and $\hat{y}_m$ denote the independent predictions derived from the Style Path (L_{style}) and Semantic Path (L_{sem}), respectively. Following empirical tuning, we set the hyperparameters to $\lambda_f = 0.4$, $\lambda_s = 0.3$, and $\lambda_m = 0.3$.

4 Experiments

In this section, we first introduce the experimental setup, including datasets, baselines, and implementation details in Section 4.1. We then present the main experimental results on both binary detection and fine-grained attribution in Section 4.2. Subsequently, we conduct a deep-dive analysis into the robustness, efficiency, and generalization capabilities of our framework in Section 4.3, followed by ablation studies in Section 4.4.

This section is organized around the following research questions:

- **RQ1 (General Effectiveness):** Can the proposed SSLA framework, by synergizing explicit stylistic fingerprints with implicit semantic embeddings, achieve state-of-the-art performance in both binary detection and fine-grained LLM attribution tasks?
- **RQ2 (Robustness & Stability):** Does SSLA demonstrate superior robustness compared to semantic-heavy baselines under resource-constrained and distributional-shift scenarios, specifically including short text lengths, data scarcity, and cross-domain transfers?
- **RQ3 (Mechanism & Interpretability):** How much does each stylistic component contribute to the final decision, and does the dual-path architecture provide transparent, interpretable evidence that addresses the "black box" limitation of traditional detectors?

4.1 Experimental Setup

Datasets. We employ three representative datasets from MGTBench(He et al., 2023) to evaluate our proposed method: **Essay**, **Reuters**, and **Wikipedia Pages (WP)**. Each dataset comprises human-written texts and machine-generated texts from 6 different LLMs (e.g., ChatGPT, Claude, LLaMA), forming a challenging 7-way attribution task (and binary detection task). These datasets cover diverse domains, ensuring a comprehensive evaluation of stylistic generalizability.

Baseline Methods. To validate the performance of SSLA, we compare it against a comprehensive set of baselines. These include zero-shot statistical methods such as DetectGPT(Mitchell et al., 2023) and Entropy(Gehrmann et al., 2019); semantic-based classifiers like RoBERTa-based detectors(Liu et al., 2019; Solaiman et al., 2019) and LM-D(He et al., 2023); and specialized attribution models such as ConDA(Bhattacharjee et al., 2023) and OTB-D(La Cava and Tagarelli, 2025).

Implementation Details. For the semantic encoder, we utilize the pre-trained RoBERTa-base model. We freeze the embedding layers and fine-tune the remaining parameters to prevent overfitting in low-resource settings. The style encoder utilizes a multi-layer perceptron (MLP) to process the constructed stylistic vectors (s and g). To extract the differential stylistic features (s_{pres}), we employ DeepSeek-V3 as the default rewriting probe to generate semantic-preserving texts. The structural stylistic features are extracted through deep dependency parsing implemented via the Stanza parser. For model training, we employ the AdamW optimizer with a learning rate of 2e-5 and a linear warmup scheduler to ensure stable convergence.

4.2 Main Results

We evaluate the performance of SSLA with a primary focus on the challenging fine-grained 7-way LLM attribution task, where distinguishing between sophisticated models requires capturing subtle stylistic nuances.

Qualitative Analysis. As visualized in Figure 5 in the Appendix, the t-SNE plot (a) demonstrates that SSLA learns highly discriminative stylistic representations. Unlike semantic embeddings that may overlap due to similar topics, our stylistic features drive the formation of clear, compact clusters for each specific LLM family. The confusion matrix (b) further corroborates this robustness. It shows that SSLA effectively minimizes off-diagonal errors, successfully distinguishing even closely related models (which typically share similar training data) with high confidence.

Attribution Performance. The quantitative re-

sults in Table 1 reveal distinct performance patterns. First, traditional zero-shot statistical methods (e.g., DetectGPT, Entropy) fail completely on this fine-grained task, yielding Macro-F1 scores below 0.40. This confirms that simple likelihood signals are insufficient for distinguishing sophisticated LLMs.

Comparison with supervised baselines highlights the superiority of our approach. While semantic-heavy methods like OTBDetector(OTB-D) perform competitively on the Essay dataset (where semantic variance is high), they exhibit notable performance degradation on domains characterized by high semantic homogeneity. In contrast, SSLA demonstrates exceptional robustness. Specifically, on the Reuters dataset, SSLA achieves an F1 of 0.966, surpassing OTB-D (0.945) by 2.1%. Even more notably, on the Wikipedia (WP) dataset—characterized by neutral tone and factual constraints—SSLA establishes a clear lead of 1.7% (0.956 vs 0.939). This performance gap offers a critical insight: when semantic boundaries become blurred due to factual standardization, the stylistic fingerprints profiled by SSLA—such as syntactic rigidity and functional word usage—become the decisive discriminative signals.

Additionally, we verify SSLA's effectiveness on the fundamental binary detection task (Human vs. Machine). Detailed results provided in the Appendix show that SSLA achieves state-of-the-art performance on the Reuters dataset while maintaining highly competitive capabilities across other diverse domains. This confirms that explicitly modeling fine-grained stylistic signals effectively captures general machine-generated artifacts without compromising detection stability.

4.3 Robustness and Efficiency Analysis

Beyond standard benchmarks, we investigate the robustness of our method under challenging conditions, specifically focusing on short text lengths, data scarcity, and cross-domain generalization.

Sensitivity to Text Length. Standard attribution methods often degrade on short texts due to insufficient statistical signals. To evaluate robustness, we test performance across four length intervals on the Reuters dataset. As illustrated in Figure 6, while baselines like OTB-D suffer a catastrophic failure in short-text scenarios—dropping to an F1 score below 0.30 for texts under 50 tokens—SSLA exhibits exceptional resilience, maintaining an F1 score of over 0.80. Notably, SSLA creates a significant performance gap of approximately 55%

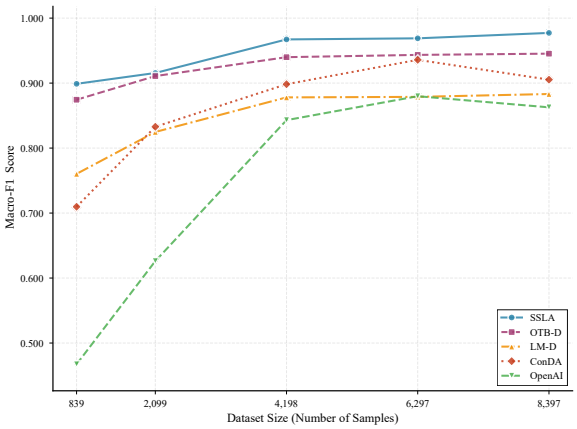


Figure 3: Data Efficiency Analysis. SSLA consistently outperforms all baseline methods across varying training set sizes, demonstrating robust capabilities particularly in few-shot settings.

against OTB-D in this challenging low-density setting. Even compared to the semantic-heavy LM-D, SSLA offers a better balance between Precision and Recall, and scales effectively as text length increases.

Data Efficiency. We further explore the data efficiency of SSLA by training on subsets of varying sizes (from 839 to 8397 samples). As shown in Figure 3, SSLA exhibits superior few-shot learning capabilities. With only 839 samples, our method achieves an accuracy of ∼90%, significantly outperforming strong baselines like ConDA and LM-D by margins of roughly 19% and 13% respectively. This finding supports our hypothesis that stylistic features are more informative and data-efficient than purely semantic embeddings, allowing the model to capture distinguishing patterns with significantly fewer examples.

Cross-Domain Generalization. Finally, we evaluate the model's generalization capability under severe distribution shifts by training on the Reuters dataset and testing on the Essay dataset. This setting is particularly challenging as it requires the model to ignore domain-specific semantics and focus on intrinsic generative signatures.

As illustrated in Figure 4, most baselines struggle with this shift. Purely semantic-based methods like LM-D and ConDA achieve F1 scores of only 0.4663 and 0.4543 respectively, indicating a heavy reliance on in-domain content. While the state-of-the-art baseline OTB-D shows improved robustness with a score of 0.6719, our proposed SSLA still outperforms it.

Specifically, SSLA achieves a Macro-F1 score

Table 1: Main Benchmark Results on Essay, Reuters, and WP Datasets (7-way Attribution). Macro-F1 is reported. Best results in each column are bolded.

Test Task Detector	Essay				Reuters				WP			
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Likelihood	0.334	0.304	0.308	0.334	0.334	0.304	0.308	0.334	0.382	0.334	0.337	0.382
Rank	0.409	0.378	0.361	0.408	0.259	0.213	0.226	0.259	0.255	0.255	0.216	0.208
Log_Rank	0.442	0.412	0.407	0.442	0.343	0.319	0.317	0.343	0.400	0.360	0.348	0.400
Entropy	0.409	0.378	0.378	0.409	0.249	0.221	0.229	0.249	0.209	0.163	0.182	0.209
Rank_GLTR	0.476	0.437	0.445	0.476	0.411	0.392	0.389	0.411	0.433	0.385	0.390	0.433
DetectGPT	0.409	0.378	0.407	0.442	0.241	0.197	0.197	0.241	0.237	0.192	0.201	0.237
NPR	0.409	0.378	0.407	0.442	0.292	0.211	0.208	0.292	0.308	0.249	0.246	0.308
LRR	0.432	0.403	0.396	0.432	0.386	0.363	0.365	0.386	0.403	0.378	0.377	0.403
OpenAI-Detector	0.808	0.802	0.836	0.808	0.893	0.890	0.901	0.893	0.725	0.723	0.781	0.725
ConDA(23)	0.935	0.935	0.935	0.935	0.946	0.946	0.948	0.946	0.918	0.915	0.925	0.918
LM-D(24)	0.890	0.882	0.901	0.890	0.921	0.921	0.918	0.921	0.890	0.890	0.898	0.890
OTB-D(25)	0.948	**0.949**	**0.952**	**0.949**	0.946	0.945	0.948	0.946	0.940	0.939	0.941	0.940
SSLA (Ours)	**0.968**	0.945	0.949	0.943	**0.978**	**0.966**	**0.964**	**0.969**	**0.971**	**0.956**	**0.955**	**0.958**

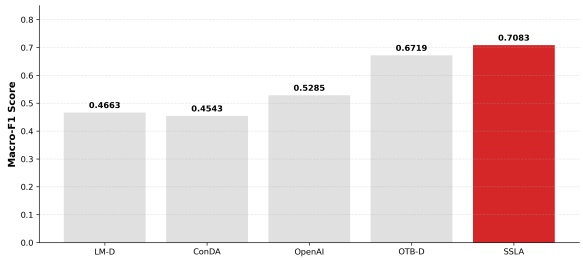


Figure 4: Cross-Domain Performance (Reuters → Essay). SSLA achieves the highest F1 (0.7083), surpassing OTB-D (0.6719) and semantic baselines, validating superior robustness against distribution shifts.

of 0.7083, surpassing OTB-D by a notable margin. This result confirms that our dual-path architecture effectively synthesizes stylistic rigidity with semantic context. Even without explicit domain adaptation techniques, the attention-based fusion mechanism allows SSLA to adaptively leverage stable stylistic fingerprints (such as SN-Grams and functional word distributions) that persist across genres, thereby achieving superior generalization.

Robustness Against Persona-Steering. To evaluate SSLA's resilience against deliberate style-masking, we conducted an extreme adversarial test on the XSum dataset. We prompted 4 SOTA LLMs (GPT-4o, DeepSeek-V3, Gemini-1.5-Flash, and GLM-4) to rewrite articles under a strict identity constraint (the full prompt template is detailed in **Appendix A.7**). As shown in Table 2, even when these highly aligned models are forced into an identical stylistic persona, SSLA maintains a Macro-F1 of 0.952, closely matching the unconstrained baseline. This confirms that our framework captures intrinsic "style inertia" that is resistant to surface-level persona alignment.

Table 2: Performance under Extreme Adversarial Persona Constraints.

Setting	Models	Macro-F1
Standard Attribution	6 LLMs	0.956 ± 0.011
Extreme Adversarial	**4 SOTA LLMs**	**0.952**

4.4 Ablation Studies and Analysis

Component Analysis. To systematically evaluate the contribution of each component, we conduct ablation studies as detailed in Table 5 in the Appendix. The results confirm the necessity of multi-view modeling: variants relying on a single source of evidence consistently underperform the full SSLA framework across domains. In addition, to isolate the contribution of the rewrite-based rigidity feature, we include two targeted variants:"s_{pres} only" and "SSLA w/o s_{pres}". The results show that s_{pres} alone retains meaningful discriminative power, while removing it from the complete system leads to a consistent performance drop. This indicates that s_{pres} provides complementary attribution evidence beyond static linguistic metrics and structural SN-Grams. It is worth noting that while the simple concatenation variant ("RoBERTa + SN_Gram") performs competitively on the Essay dataset, it struggles to generalize to the more diverse Wikipedia (WP) domain. In contrast, our proposed SSLA (Full Fusion) maintains consistently high performance across all domains, achieving the highest Average F1. This indicates that the attention-based fusion mechanism effectively mitigates the risk of overfitting to specific structural patterns.

Probe Independence Analysis. A critical con-

cern in our differential framework is whether the effectiveness of s_{pres} stems from the intrinsic "style inertia" of the source model or merely from the specific rewriting bias of the chosen probe. To verify that our stylistic signal is probe-independent, we replaced the default rewriter (DeepSeek-V3) with a structurally distinct model, Gemini-1.5-Flash. As shown in Table 4 in the Appendix, swapping the probe does not materially degrade attribution performance, with the Average F1 remaining highly stable (0.954 vs. 0.956). In addition, the ablation results in Table 5 show that removing s_{pres} from the full system causes a consistent performance drop, while s_{pres} only still retains non-trivial attribution ability. Together, these findings suggest that s_{pres} captures a meaningful and complementary stylistic signal rather than merely reflecting idiosyncrasies of a specific rewriter.

Interpretability and Qualitative Insights. We further analyze the decision-making process using SHAP values, as illustrated in Figure A.7 in the Appendix. The analysis reveals that stylistic features, particularly specific SN-Gram structures and lexical density metrics, account for a significant portion of the decision weight. This provides interpretable supporting evidence for attribution, addressing the "black box" limitation of traditional neural classifiers. To further validate these findings, we provide a qualitative case study in **Appendix A.8**. These cases demonstrate that when the Stylistic Path encounters ambiguity due to standardized narrative structures (e.g., conventional fairytale openings), the Semantic Path can identify latent fingerprints to perform corrective fusion; conversely, the semantic path may introduce noise in highly stylized human narratives. While performance on ultra-short texts (< 50 tokens) remains a challenge, the overall results confirm that integrating multi-level stylistic fingerprints creates a robust and interpretable attribution framework.

5 Conclusion

In this paper, we presented SSLA, a framework that synergizes stylometrics with semantics designed to address the fragility of existing semantic-based LLM attribution methods. By modeling the cognitive signature of generators—spanning lexical habits, syntactic structures, and semantic intentions—SSLA extracts inherent style inertia that persists even when surface content varies. To effectively integrate these signals, we introduced a dual-path fusion architecture that dynamically synthesizes explicit stylistic features with implicit semantic embeddings, ensuring robust attribution across diverse scenarios.

Comprehensive experiments validate that SSLA achieves state-of-the-art performance while demonstrating exceptional robustness in challenging conditions such as short texts and cross-domain transfers. Furthermore, the explicit modeling of stylistic signatures enables effective generalization in few-shot settings, significantly outperforming deep semantic classifiers when training data is scarce. Feature analysis further confirms that SSLA offers superior interpretability, allowing users to trace decisions back to specific linguistic patterns rather than opaque probability scores.

Limitations

Despite its effectiveness, SSLA has limitations. First, the construction of multi-dimensional stylistic profiles incurs additional computational overhead compared to end-to-end semantic models. This trade-off between inference efficiency and robustness suggests that the current framework is optimally suited for offline forensic analysis, rather than latency-sensitive real-time streaming detection. Second, the framework relies on high-quality linguistic parsers, which may limit its applicability to low-resource languages lacking robust NLP tools. Finally, our evaluation focuses on closed-set attribution; the framework's capability to detect "unknown" models (open-world setting) remains to be explored in future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62272074, 62201112, and 62402077), the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202400607 and KJQN202400654).

References

Harika Abburi, Sanmitra Bhattacharya, Edward Bowen, and Nirmala Pudota. 2025. Ai-generated text detection: A multifaceted approach to binary and multiclass classification. *arXiv preprint arXiv:2505.11550*.

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*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 598–610.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. 2025. Detecting stylistic fingerprints of large language models. *arXiv preprint arXiv:2503.01659*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Annual Meeting of the Association for Computational Linguistics*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*.

Mihailo korić, Ranka Stanković, Milica Ikonić Neić, Joanna Byszuk, and Maciej Eder. 2022. Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution. *Mathematics*.

Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Amanat Chadha, Joshua Garland, and Huan Liu. 2024. A survey of ai-generated text forensic systems: Detection, attribution, and characterization. *arXiv preprint arXiv:2403.01152*.

Tharindu Kumarage and Huan Liu. 2023a. Neural authorship attribution: Stylometric analysis on large language models. *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54.

Tharindu Kumarage and Huan Liu. 2023b. Neural authorship attribution: Stylometric analysis on large language models. *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54.

Lucio La Cava and Andrea Tagarelli. 2025. Openturingbench: an open-model-based benchmark and framework for machine-generated text detection and attribution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26666–26682.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. *arXiv preprint arXiv:2303.11666*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023b. Mage: Machine-generated text detection in the wild. In *Annual Meeting of the Association for Computational Linguistics*.

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*.

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*.

Pablo Francisco Posadas-Durán, Germán Ríos-Toledo, Erick Velázquez-Lozada, A De Jesús Osuna-Coutiño, Madaín Pérez-Patricio, and Fernando Pech May. 2025. Learning the style via mixed sn-grams: An evaluation in authorship attribution. *AI*, 6(5):104.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41:853–860.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Conference on Empirical Methods in Natural Language Processing*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated text. *Communications of the ACM*, 67:50 – 59.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. In *NAACL-HLT*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. 2025a. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51:275–338.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. 2025b. Gradient-based model fingerprinting for llm similarity detection and family classification. *arXiv preprint arXiv:2506.01631*.

Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Neng H. Yu. 2024. Text fluoroscopy: Detecting llm-generated text through intrinsic features. In *Conference on Empirical Methods in Natural Language Processing*.

A Appendix

In this appendix, we provide supplementary evaluations to support the main findings presented in

Section 4.2. Specifically, we present detailed quantitative results for the binary detection task, qualitative visualizations of the stylistic feature space, and an in-depth analysis of feature contributions.

A.1 Binary Detection Performance

Table 3 presents the comprehensive performance metrics—including Accuracy, Macro-F1, Precision, and Recall—across the Essay, Reuters, and Wikipedia (WP) datasets.

Consistent with our observations in fine-grained attribution, SSLA demonstrates exceptional robustness in the binary classification setting. Notably, on the **Reuters** dataset, our method achieves near-perfect scores (F1 > 0.99), effectively matching the performance of specialized binary detectors like ConDA and OTB-D. This confirms that the fine-grained stylistic fingerprints extracted by SSLA also serve as highly effective discriminators for general machine-generated artifacts.

A.2 Probe Independence Analysis

To further support the discussion in the main text regarding the probe-independent nature of our stylistic signals, Table 4 presents the detailed performance comparison when replacing the default rewriter (DeepSeek-V3) with a structurally distinct model (Gemini-1.5-Flash). The results demonstrate that SSLA maintains highly stable attribution performance across different rewriting probes.

A.3 Qualitative Analysis Visualization

Figure 5 visualizes the discriminative capability of SSLA through feature clustering and classification confusion matrices. Specifically, the t-SNE plot (a) is generated using the 128-dimensional aligned hidden representation $h_{align} = [h_{style}; h_{sem\_proj}]$. Here, $h_{style} \in R^{64}$ is the output of the style MLP, and $h_{sem\_proj} \in R^{64}$ is obtained by projecting RoBERTa's 768-dimensional pooled semantic features through a linear dimensionality reduction layer.

In Table 5, "Feature" denotes the handcrafted feature block composed of static linguistic metrics (s_{ling}) and rewrite-based rigidity features (s_{pres}), excluding SN-Grams.

A.4 Text Length Sensitivity Analysis

We further investigate the robustness of SSLA under resource-constrained settings by analyzing its performance across different text length intervals. Figure 6 presents the detailed curves for F1-score, Recall, and Precision on the Reuters dataset. The results indicate that SSLA maintains superior stability compared to baselines, particularly in the challenging short-text regime (< 50 tokens), where traditional methods typically suffer from insufficient discriminative signals.

A.5 Detailed Ablation Analysis

To systematically evaluate the contribution of each component within the SSLA framework, we conduct a comprehensive ablation study. Table 5 details the performance of various model variants across the Essay, Reuters, and Wikipedia (WP) datasets. The results confirm the necessity of multi-view modeling: variants relying on a single source of evidence consistently underperform the full SSLA framework across domains. To further isolate the contribution of the rewrite-based rigidity signal, we additionally include two targeted variants, "s_{pres} only" and "SSLA w/o s_{pres}". The results show that s_{pres} alone retains meaningful discriminative power, while removing it from the complete system leads to a consistent performance drop. Notably, our proposed Full Fusion strategy achieves the best stability and performance, confirming that the attention-based fusion mechanism effectively mitigates the risk of overfitting.

A.6 Feature Contribution Analysis

We conduct a SHAP (SHapley Additive exPlanations) analysis to interpret the decision-making process of our dual-path framework. As illustrated in Figure 7, specific stylistic components, such as SN-Grams and lexical density metrics, play a decisive role in the model's predictions, complementing the semantic backbone. This provides transparency to the attribution process, validating that SSLA relies on consistent linguistic patterns rather than opaque artifacts.

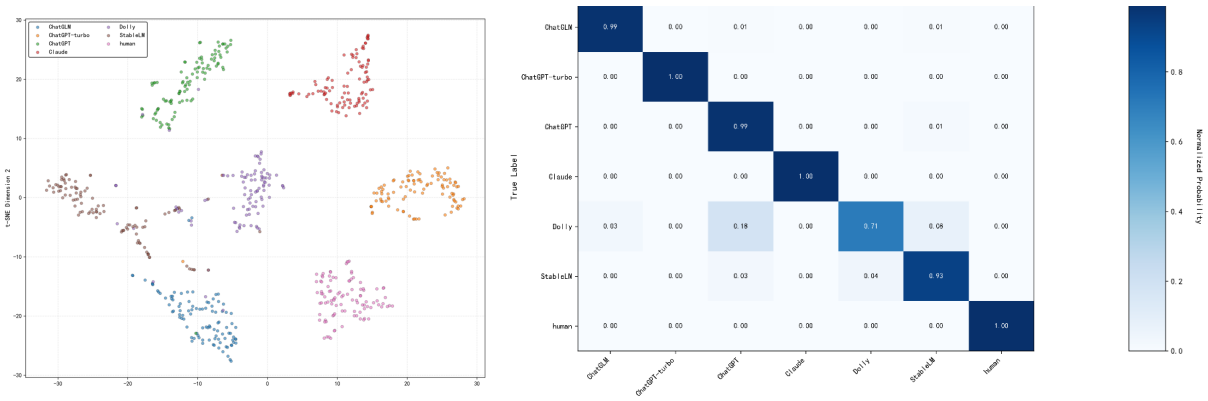
A.7 Adversarial Persona Prompting Details

In Section 4.3, we evaluate the robustness of SSLA against extreme style-masking attacks (persona-steering). To ensure all evaluated SOTA models generate text under an identical stylistic constraint, we employed the following standardized prompt template:

> *"You are a BBC news editor. Rewrite the following lead sentence into a full news article (∼300 words).*
>
> ***Rules:** (1) MUST start with: "[prefix]"; (2) Use a neutral tone; (3) Include realistic details; (4) Do*

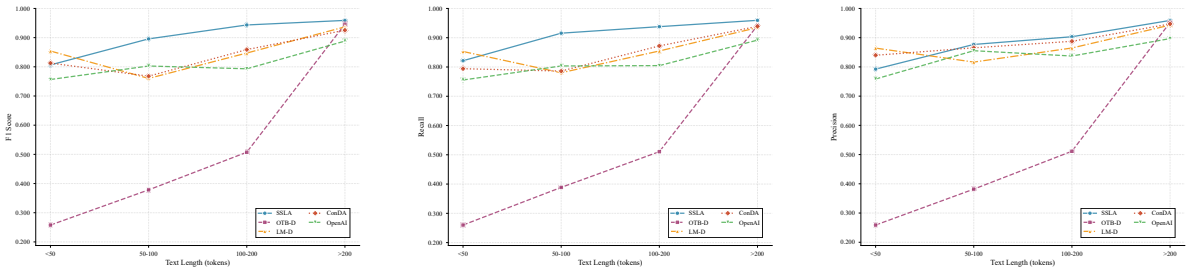
Table 3: Binary Detection Performance (HWT vs. MGT). Best results in each column are bolded.

Test Task Detector	Essay				Reuters				WP			
	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
Likelihood	0.893	0.894	0.887	0.901	0.765	0.750	0.801	0.705	0.830	0.829	0.835	0.823
Rank	0.784	0.800	0.745	0.865	0.617	0.659	0.594	0.740	0.727	0.747	0.695	0.808
Log_Rank	0.898	0.900	0.888	0.911	0.762	0.742	0.810	0.685	0.820	0.819	0.825	0.813
Entropy	0.761	0.766	0.751	0.782	0.455	0.426	0.450	0.405	0.724	0.741	0.699	0.787
Rank_GLTR	0.901	0.902	0.897	0.906	0.767	0.747	0.816	0.690	0.803	0.794	0.829	0.762
DetectGPT	0.759	0.742	0.797	0.694	0.750	0.729	0.794	0.675	0.674	0.683	0.665	0.702
NPR	0.681	0.657	0.710	0.611	0.735	0.715	0.773	0.665	0.717	0.718	0.715	0.722
LRR	0.888	0.885	0.907	0.865	0.782	0.746	0.895	0.640	0.797	0.783	0.843	0.732
OpenAI-Detector	0.722	0.703	0.755	0.658	0.817	0.780	0.977	0.650	0.790	0.741	0.967	0.601
ConDA	0.992	0.992	0.992	0.992	**0.997**	**0.997**	**0.997**	0.995	0.891	0.902	0.821	0.892
OTB-D	**0.998**	**0.996**	**0.998**	0.993	**0.997**	0.995	**0.998**	0.992	**0.989**	**0.977**	**0.994**	0.963
LM-D	0.979	0.979	0.964	0.964	0.972	0.972	0.960	0.985	0.904	0.911	0.844	**0.989**
SSLA (Ours)	0.993	0.993	0.995	**0.994**	**0.997**	**0.997**	0.993	**0.997**	0.974	0.974	0.975	0.974



(a) (b)

Figure 5: Qualitative visualization of SSLA's discriminative capability. (a) t-SNE plot showing clear cluster separation based on the 128D aligned hidden representations. (b) Confusion matrix demonstrating high classification accuracy on the Reuters dataset.



(a) (b) (c)

Figure 6: Text Length Sensitivity Analysis. SSLA demonstrates consistent robustness across different text lengths. From left to right: F1, Recall, and Precision, maintaining stable performance even in short-text scenarios.

NOT include phrases such as 'As an AI model' or 'I am an assistant'."

By enforcing a strict professional persona ("BBC news editor") and explicit constraints (neutral tone, specific prefix), this prompt forces the generated texts from different LLMs to exhibit massive semantic, lexical, and structural overlap. The results presented in the main text demonstrate that SSLA's deep stylometric fingerprints remain effective discriminators even under such extreme semantic homogenization.

A.8 Qualitative Case Study and Error Analysis

To provide a transparent view of the decision-making process, we present the full text and cor-

Table 4: Performance comparison of different rewriting probes. The results demonstrate that SSLA maintains robust attribution performance regardless of whether DeepSeek-V3 or Gemini-1.5-Flash is used as the rewriter.

Dataset	Rewriter Probe	Acc	F1	Prec	Rec
2*Essay	DeepSeek-V3 (Default)	0.968	0.945	0.949	0.943
	Gemini-1.5-Flash (New)	**0.974**	**0.953**	**0.954**	**0.952**
2*Reuters	DeepSeek-V3 (Default)	**0.978**	**0.966**	**0.964**	**0.969**
	Gemini-1.5-Flash (New)	0.954	0.952	0.954	0.952
2*WP	DeepSeek-V3 (Default)	**0.971**	0.956	**0.955**	**0.958**
	Gemini-1.5-Flash (New)	0.958	**0.957**	0.953	0.952

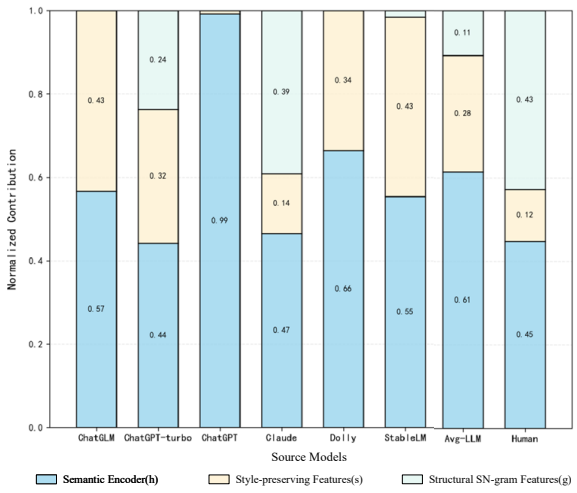


Figure 7: **Relative Contribution of Semantic vs. Stylistic Paths.** SHAP analysis reveals that stylistic features (Features + SN-Grams) play a crucial role, complementing the semantic backbone.

responding metrics for two representative cases. These examples highlight the interaction between the stylistic and semantic paths in SSLA.

Case 1: Semantic Overpowering (Error Analysis)

Original Text (Human):

> "Pikachu was bleeding, but Ash didn't care. He needed this badge to move on and he wasn't stopping until he had crushed everyone in his path. He yelled to Pikachu, 'Alright that's enough, stop wasting time and Hit him again with a thunderbolt, Pikachu!' The small red dots of Pikachu's cheeks began to pulse, The air began to tense and lightning arced from the sky down to the battlefield. The opponent was nervous now, because at once a large bolt of thunder burst from the clouds, through the air as it went and catching the opponent's Arbok by surprise. It shook with the force of the electricity coursing through its body and within seconds, collapsed on the field. It attempted to inch slowly towards its master in one final act, as the life drained slowly from its eyes. Ash was declared the winner, but he already knew that. Just another victim on the road to his destiny, there would be many, many more."

Predictions & Metrics:

Path	Prediction
Ground Truth	Human
Style Path	Human
Semantic Path	StableLM
Final Fusion	**StableLM (Incorrect)**

Analysis: This sample features a highly stylized, dark-themed fan-fiction narrative. Supported by our multi-dimensional profiling, the **Style Path** correctly detected human-like stylistic traits, evidenced by a high Average Dependency Distance (2.99) and a specific rewriting rigidity (BERTScore F1: 0.869). However, the **Semantic Path** (RoBERTa) misclassified the text as StableLM. This is likely due to the semantic overlap between creative writing prompts and the data distribution found in specific LLM training sets. The fusion layer incorrectly assigned higher attention weight to the semantic embeddings, failing to leverage the accurate stylistic signal in this instance.

Case 2: Synergetic Correction (Success Case)

Original Text (ChatGLM):

> "Once upon a time, in a small town nestled in the heart of a bustling city, there was a young girl named Sarah. Sarah was a kind and gentle soul, with a kind heart and a gentle nature. She was the only child in her family, and she loved nothing more than spending time with her siblings and her parents. One day, Sarah's parents took her to the dentist to get her first set of teeth. As they were cleaning her teeth, the dentist told Sarah that she had a rare condition that caused her teeth to grow at an unusual rate. She needed to have all of her teeth removed and replaced with artificial ones, which would cost a fortune. Sarah's parents were shocked and horrified by the news. They tried to find a way to pay for the surgery, but the price of baby teeth was skyrocketing, and there was no way they could afford it. As the days went by, Sarah's parents became more and more desperate. They started selling their valuable property and scrimping on their expenses to try and make ends meet."

Predictions & Metrics:

Path	Prediction
Ground Truth	ChatGLM
Style Path	ChatGPT
Semantic Path	ChatGLM
Final Fusion	**ChatGLM (Correct)**

Analysis: The generator employs a standard fairytale opening ("Once upon a time..."). Supported by our explicit stylistic profiling, this structure aligns closely with the polished and standardized output

Table 5: Ablation study on core components of SSLA. The table additionally includes two targeted variants, "s_{pres} only" and "SSLA w/o s_{pres}", to isolate the contribution of the rewrite-based rigidity signal. The proposed Full Fusion strategy achieves the best overall stability and performance across diverse domains, especially on the challenging WP dataset.

Model Variant	**Essay**		**Reuters**		**WP**		**Avg**
	Acc	F1	Acc	F1	Acc	F1	**F1**
feature	0.620	0.534	0.377	0.350	0.537	0.654	0.513
s_{pres} only	0.435	0.446	0.516	0.317	0.474	0.561	0.441
SN_Gram	0.783	0.654	0.800	0.679	0.757	0.654	0.662
RoBERTa Baseline	0.941	0.899	0.975	0.963	0.959	0.943	0.935
SN_Gram + Feature	0.904	0.904	0.921	0.921	0.850	0.850	0.892
RoBERTa + Feature	0.904	0.829	0.937	0.899	0.937	0.901	0.876
RoBERTa + SN_Gram	**0.968**	**0.952**	0.975	0.963	0.959	0.943	0.953
SSLA w/o s_{pres}	0.954	0.944	0.958	0.951	0.946	0.947	0.947
SSLA (Full Fusion)	**0.968**	0.945	**0.978**	**0.966**	**0.971**	**0.956**	**0.956**

patterns of ChatGPT (evidenced by a Lexical Diversity of 0.615, an Avg. Dep. Depth of 2.87, and a high rewriting rigidity BERTScore F1 of 0.898), leading the **Style Path** to a misclassification. However, the **Semantic Path** (RoBERTa) accurately identified the latent distributional fingerprints of ChatGLM based on deep semantic embeddings. The attention-based fusion successfully prioritized the semantic confidence over the stylistic ambiguity, resulting in a correct attribution.