

Degradation of Stylometric Attribution Accuracy for AI-Generated Text [†]

Kalin Kopanov *  and Tatiana Atanasova 

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Bl. 2, 1113 Sofia, Bulgaria; tatiana.atanasova@iict.bas.bg

* Correspondence: kalin.kopanov@iict.bas.bg

[†] Presented at the 15th International Scientific Conference TechSys 2026—Engineering, Technologies and Systems, Plovdiv, Bulgaria, 14–16 May 2026.

Abstract

The growing use of large language models (LLMs) makes it essential to identify which system produced a given text. Existing stylometric detectors perform well on raw outputs; however, users routinely translate, paraphrase, or edit content before release. These transformations can erase the lexical and syntactic cues that support attribution. This study measures the resulting performance drop on a purpose-built benchmark of synthetic responses from two instruction-tuned models, Qwen 2.5 (32B) and Gemma 3 (27B). Each passage is subjected to common editing workflows ranging from machine translation to paraphrasing and grammar-focused rewriting. The edited variants are evaluated using a custom machine-learning classifier. Overall, attribution remains robust, but obfuscation leads to degradation that often renders the text unusable for professional purposes. Although the analysis is limited to data from two LLMs, the controlled setting shows how specific post-editing actions alter stylometric signals. Recognising this sensitivity is essential for academic integrity checks, platform-moderation tools, and the ongoing debate between watermarking and stylometry. The results provide practical guidance on developing attribution methods that remain reliable after routine text updates.

Keywords: large language models (LLMs); stylometry; SVM; round-trip translation (RTT); Qwen 2.5; Gemma 3; COMET-Kiwi; xCOMET

1. Introduction

Attribution of authorship through writing style has long been a subject of study in forensic and computational linguistics. Stylometric pipelines assume that writers manifest stable lexical, syntactic, and structural patterns. Early work extracted word and character n-grams, function-word frequencies, vocabulary richness, and sentence-length statistics, and fed them to linear or margin-based classifiers, such as Naïve Bayes, logistic regression, k-nearest neighbours, and support vector machines (SVM) [1,2]. A meta-analysis shows that SVM consistently outperforms simpler algorithms on sparse text, a result confirmed by [3] on constructive journalism metadata and by the authors' own baseline with Qwen 2.5 versus Gemma 3 [4]. These findings establish SVM as an interpretable and high-accuracy baseline for the robustness tests that follow.

Building on these classical methods, deep learning extends stylometry by learning features directly from raw text. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can capture both local and sequential patterns. For instance,



Academic Editors: Nikola G. Shakev,
Valyo N. Nikolov, Nikolay
R. Kakanakov and Sevil
A. Ahmed-Shieva

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

character-level CNNs handle short or noisy inputs [5], whereas syntactic RNNs trained on part-of-speech sequences generalise across domains [6]. Transformer encoder models are rapidly becoming state-of-the-art in authorship attribution. Recent work demonstrates that BERT-family models can achieve over 95% accuracy on domain-specific datasets [7] and exceed 90% on established benchmarks, such as Legal and IMDb62, although their performance on shorter-text sets, like Blog10/50 and Reuters50, remains lower [8]. Over time, however, the topic of authorship obfuscation has been well-examined, with studies highlighting the fragility of both linear and neural detectors, particularly their sharp degradation after paraphrasing or translation [9]. Despite subsequent gains in detection performance, this vulnerability persists, motivating ongoing research on transformation-robust stylometry to develop more resilient methods.

This fragility is often tested through text transformations, such as round-trip translation (RTT), a canonical pipeline in which text is translated out of and then back into the source language. The process alters lexical and syntactic cues while largely preserving semantics [10]. Study [11] demonstrates that single-hop RTT remains highly correlated with forward-translation quality, even for typologically distant languages; however, copy-based adversarial systems can reduce these correlations by approximately 50 percentage points. The survey in [12], therefore, calls for benchmarks that gauge robustness across a spectrum of real-world edits, a gap that our study aims to address.

Advances in machine translation evaluation further support these efforts by helping to measure the preservation of meaning after editing. Contrastive Perturbation Enhancement (CoPE) fine-tunes medium-sized machine translation (MT) models with entailment (positive) and contradiction (negative) sentence pairs, raising BLEU, COMET-22, and xCOMET scores on WMT'21/'22 language pairs [13]. Reference-free COMET-Kiwi [14] and its multilingual successor xCOMET [15] output a sentence-level quality score together with token- and span-level error labels, enabling fine-grained analysis of post-edited text. New judge models such as M-PROMETHEUS [16] push sentence-level accuracy higher, although they still lack structured error spans. These metrics allow us to link stylometric decay to underlying semantic drift, an aspect that earlier work largely ignored.

Machine-generated text under post-editing (such as translation, paraphrasing, or rewriting) represents a less-explored domain in stylometric attribution. When one LLM's output is modified by a translation service, paraphrasing tool, or another LLM, the resulting text is effectively co-produced by multiple systems, namely the original generator and the editor(s). This scenario parallels human author–translator or author–editor dynamics but introduces the complexity of potentially layered authorship. While there has been interest in identifying translation or paraphrasing systems, our study uniquely examines how these transformations affect the attribution of the original source model, with a focus on both multilingual (e.g., round-trip translation) and monolingual (e.g., grammar rephrasing) edits.

In our previous comparative pattern analysis of Qwen 2.5 and Gemma 3 text generation [4], we demonstrated that machine learning classifiers could achieve 99% accuracy in distinguishing between these models using stylometric features. However, we acknowledged a critical limitation: “The main difficulty stems from the expected decline in attribution accuracy when texts experience extensive editing or paraphrasing. Human editing processes can eliminate or modify the distinctive linguistic patterns of Qwen 2.5 and Gemma 3, which would decrease the classifier's performance” [4].

We leverage insights from prior work with specific hypotheses. Post-editing will dilute some original stylometric signals, but sufficient distinctive patterns should remain for attribution, especially when the editor is not deliberately obfuscating style. We also expect that different methods (such as service-based versus LLM-based translation, or surface-

level grammar fixes versus deep neural paraphrasing) may have differential impacts on stylometric preservation, with multilingual workflows introducing cross-language concerns. This investigation thus combines neural text attribution with robustness testing across various editing types. It addresses a critical gap in the resilience of LLM attribution systems to real-world modifications, including multilingual round-trip translation (RTT) alongside monolingual paraphrasing approaches delivered by tools like *Grammarly* and *Undetectable.ai*, as well as fine-tuned models such as *Parrot* (T5-based) [17] and *Pegasus-PAWS* [18].

These hypotheses guide our research questions, including those on the impacts of paraphrasing, semantic preservation, and model differences. Furthermore, our empirical findings reveal that while post-editing indeed attenuates stylometric markers (e.g., lexical diversity and syntactic complexity), attribution accuracy remains robust across most tested scenarios. By quantifying stylometric degradation, this study establishes the foundation for more robust detection methods in an era of rapidly increasing AI-generated content.

2. Materials and Methods

Building on the framework introduced in [4], we focus on the 200 synthetic English passages (200–250 words each) that formed that study’s held-out test set, split evenly between Qwen 2.5 and Gemma 3. We reuse the “all-features” SVM detector, which combines Term Frequency–Inverse Document Frequency (TF-IDF) with syntactic and stylistic cues. The detector was trained and validated on the remaining 5800 passages (4640/1160 split), and achieved 99% accuracy on our SVM benchmark. Any performance changes reported here are solely due to the editing workflows.

Our study asks five focused research questions (RQ) that link stylometric erosion to text quality:

- RQ1: Does paraphrasing hide the source-model signature?
- RQ2: How much meaning is preserved after editing?
- RQ3: Is there a trade-off between hiding and textual quality?
- RQ4: Which model’s generated text preserves semantics better under paraphrasing workflows?
- RQ5: What specific factors cause very low-quality outputs?

To answer them, we apply five everyday transformation workflows: RTT, grammar correction, neural paraphrasing, summarization-based rewriting, and commercial “humanization” services. We then measure how far each workflow deviates from its 99% baseline.

RTT marks the most aggressive end of the editing spectrum. A companion paper [19] evaluated six language paths. Here, we retain four that typify single versus multi-hop behaviour without overwhelming the analysis. The single-hop chains are EN → BG → EN processed by *googletrans* (v 4.0.2) [20] and EN → RU → EN processed by OpenAI’s *4o-mini* (API model identifier gpt-4o-mini-2024-07-18) accessed via API. The multi-hop chains extend those paths to EN → BG → RU → EN with *googletrans* and EN → RU → ZH → EN with *4o-mini*. We use the tenth RTT iteration for both single- and multi-hop chains because earlier experiments showed that BLEU drift stabilises after three to six passes, depending on the pipeline. Pass-10 texts therefore provide a clear worst-case stylometric challenge while still retaining interpretable meaning.

The lighter transformations reflect scenarios that writers deploy. Grammarly (<https://www.grammarly.com>, accessed on 16 May 2025) runs in both basic grammar-fix and advanced rephrase modes, separating surface polishing from style shifting. Neural paraphrasing utilises *Parrot* (prithivida/parrot_paraphraser_on_T5, commit 9f32aa1), a T5 variant tuned for semantic preservation, and *Pegasus-PAWS* (domenicrosati/pegasus-xsum-finetuned-paws, commit 1b13d84), originally a summariser but now repurposed for rewriting. *Undetectable.ai*

(<https://undetected.ai>, accessed on 27 May 2025) is a commercial tool that promises to “humanise” AI text, aiming to evade AI detectors.

We evaluate each edited passage on two dimensions. Stylometric detectability is measured by our SVM’s accuracy after editing. Text quality and semantic fidelity are assessed using COMET-Kiwi, a reference-free tool, and xCOMET, a 10.7 billion-parameter explainable model that provides segment-level scores and error spans. Both metrics originate from translation evaluation and were trained on human adequacy judgments, in which any divergence from the reference is treated as an error. When applied to paraphrased text, where synonym substitution and clause restructuring are desirable rather than erroneous, xCOMET in particular may assign lower scores than the actual meaning loss warrants. We retain it as the most discriminating metric available, but interpret absolute scores conservatively. To ensure that paraphrasing introduces genuine change and meaningful paraphrastic variation, we compute the Sørensen–Dice coefficient on word bigrams (using the same tokenisation that xCOMET produces) and discard any pair whose score exceeds an adaptive threshold—the 95th percentile of each workflow’s similarity distribution with a floor of 0.90—resulting in 0–5% of pairs being discarded across workflows (e.g., 5% for *Grammarly* ‘fix’, whose minimal edits pushed the adaptive threshold to 0.98; the threshold remained 0.90 for all other workflows). This filtering ensures that we analyse only cases in which transformations produce substantive textual variation rather than near-identical copies.

Together, these elements form a continuous spectrum of editing intensity, from simple grammar adjustments to controlled paraphrasing and extensive multi-hop RTT. By monitoring SVM accuracy and quality metrics along this spectrum, we can identify when stylometric cues degrade and guide practitioners on editing choices that conceal model provenance without compromising readability.

3. Results

Unless noted, all reported results are averaged over the 200 source passages and their edited variants. Accuracy is reported with Wilson 95% confidence intervals (CIs); percentage-point changes (Δ) are measured relative to the unedited baseline for the same model. McNemar tests (continuity-corrected, complemented by the exact binomial test for small discordant counts; Table A2) assess paired differences; an asterisk (*) marks results significant at $p < 0.01$. Full CIs for all entries are provided in Table A1 in Appendix A, together with complete McNemar test statistics in Table A2.

3.1. Stylometric Classification

Cross-transformation analysis reveals systematic model-specific vulnerability patterns that persist regardless of editing methodology. As described in Table 1, the detector performs almost perfectly on raw outputs, achieving 98% (95% CI: 93–99) for Qwen 2.5 and 100% (95% CI: 96–100) for Gemma 3. Post-editing, the two models diverge sharply.

RTT illustrates the contrast. *4o-mini* keeps most stylometric cues intact: Qwen 2.5 still scores 96–98% ($\Delta \leq 2$ pp, n.s.), while Gemma 3 falls to 86% (95% CI: 78–91; $\Delta -14$ pp*) after a single-hop chain from pass 10 (e.g., EN $\rightarrow$ RU $\rightarrow$ EN) and to 78% (95% CI: 69–85; $\Delta -22$ pp*) after a multi-hop chain. *Googletrans*, however, inflicts heavy damage. A multi-hop chain from pass 10 (e.g., EN $\rightarrow$ BG $\rightarrow$ RU $\rightarrow$ EN) drives Gemma 3 down to 66% (95% CI: 56–75; $\Delta -34$ pp*), while Qwen 2.5 remains unaffected at 99% (95% CI: 95–100; $\Delta +1$ pp, n.s.).

Table 1. Performance metrics for Qwen 2.5 and Gemma 3 models under various obfuscation workflows, including SVM accuracy, AI % (probability of AI-generated text; lower values indicate human classification by *Undetectable.ai*), delta differences (AI – SVM), and quality estimation scores (COMET-Kiwi and xCOMET means $\pm$ standard deviations).

Workflow	Model	SVM Acc %	AI %	Δ (AI – SVM)	COMET-Kiwi (Mean $\pm$ σ)	xCOMET (Mean $\pm$ σ)
Baseline	Qwen 2.5	98.00	92.49	−5.51	–	–
	Gemma 3	100.00	82.65	−17.35	–	–
RTT (googletrans) single-hop chains	Qwen 2.5	99.00	80.86	−18.14	0.776 $\pm$ 0.050	0.331 $\pm$ 0.134
	Gemma 3	76.00	60.68	−15.32	0.737 $\pm$ 0.067	0.199 $\pm$ 0.115
RTT (4o-mini) single-hop chains	Qwen 2.5	98.00	92.55	−5.45	0.884 $\pm$ 0.018	0.646 $\pm$ 0.174
	Gemma 3	86.00	89.17	+3.17	0.879 $\pm$ 0.021	0.445 $\pm$ 0.188
RTT (googletrans) multi-hop chains	Qwen 2.5	99.00	72.80	−26.20	0.706 $\pm$ 0.059	0.235 $\pm$ 0.114
	Gemma 3	66.00	46.41	−19.59	0.659 $\pm$ 0.060	0.135 $\pm$ 0.076
RTT (4o-mini) multi-hop chains	Qwen 2.5	96.00	93.19	−2.81	0.873 $\pm$ 0.021	0.616 $\pm$ 0.173
	Gemma 3	78.00	82.96	+4.96	0.868 $\pm$ 0.025	0.398 $\pm$ 0.192
Grammarly ‘fix’	Qwen 2.5	96.00	93.02	−2.98	0.898 $\pm$ 0.013	0.625 $\pm$ 0.181
	Gemma 3	100.00	84.87	−15.13	0.896 $\pm$ 0.017	0.440 $\pm$ 0.177
Grammarly ‘rephrase’	Qwen 2.5	98.00	93.54	−4.46	0.879 $\pm$ 0.017	0.658 $\pm$ 0.158
	Gemma 3	84.00	86.67	+2.67	0.880 $\pm$ 0.020	0.449 $\pm$ 0.184
Parrot	Qwen 2.5	97.00	75.36	−21.64	0.842 $\pm$ 0.030	0.464 $\pm$ 0.201
	Gemma 3	95.00	56.49	−38.51	0.822 $\pm$ 0.038	0.241 $\pm$ 0.141
Pegasus-PAWS	Qwen 2.5	99.00	63.11	−35.89	0.773 $\pm$ 0.048	0.279 $\pm$ 0.148
	Gemma 3	56.00	32.15	−23.85	0.725 $\pm$ 0.062	0.123 $\pm$ 0.098
<i>Undetectable.ai</i> ‘humanize’ option	Qwen 2.5	100.00	1.00	−99.00	0.801 $\pm$ 0.031	0.492 $\pm$ 0.167
	Gemma 3	73.33	1.00	−72.33	0.812 $\pm$ 0.030	0.296 $\pm$ 0.155

Light grammar polishing is largely invisible to the detector. *Grammarly’s* ‘fix’ mode leaves Qwen 2.5 at 96% (95% CI: 90–98) and Gemma 3 at 100% (95% CI: 96–100). Switching to ‘rephrase’ begins to erode surface-driven signatures: Gemma 3 drops to 84% (95% CI: 76–90; Δ −16 pp*), whereas Qwen 2.5 still posts 98% (95% CI: 93–99; Δ 0 pp, n.s.).

Paraphrasing tools behave very differently. *Parrot* causes only modest losses, Gemma 3 95% (Δ −5 pp, n.s.), Qwen 2.5 97% (Δ −1 pp, n.s.), whereas *Pegasus-PAWS* virtually collapses Gemma 3’s footprint to 56% (95% CI: 46–65; Δ −44 pp*) while Qwen 2.5 stays at 99% (95% CI: 95–100; Δ +1 pp, n.s.).

Commercial “humanisation” produces the strangest asymmetry. *Undetectable.ai* pushes Gemma 3 down to 73% (95% CI: 56–86; Δ −27 pp *) yet leaves Qwen 2.5 at 100% (95% CI: 89–100), statistically indistinguishable from its baseline (Table A2 records no discordant pairs). Generic obfuscation can blur lexical cues while leaving rhythm-based fingerprints intact.

To provide context, all workflows in this study are evaluated using both our custom SVM classifier and the *Undetectable.ai* detection tool as a commercial reference point. For the ‘humanize’ option specifically, *Undetectable.ai* returns low AI-detection scores (around 1%), as expected, since the tool should evaluate its own edited text. In contrast, the independent

SVM assessment reveals persistent detectability, particularly for Qwen 2.5. However, *Undetectable.ai* operates as a proprietary black-box system without public documentation on its training or updates, and we observed algorithmic instability during our May 2025 testing (e.g., scores shifting to near-uniform 0.99 after an unannounced update, as acknowledged by the service). We therefore treat its results as an illustrative reference rather than rigorous scientific evidence, focusing primarily on the consistent SVM outcomes. These results reflect *Undetectable.ai*'s intended obfuscation success rather than a direct comparison to other workflows.

Across the 18 workflow $\times$ model configurations, SVM accuracy correlates strongly with mean semantic preservation (Pearson $r = 0.61$, $p < 0.01$), while showing a weaker correlation with surface fluency ($r = 0.46$), confirming that meaning preservation, not surface readability, governs detectability (addressing RQ2 and RQ3). The relationship is carried by Gemma 3 ($r = 0.69$ across its nine workflows); for Qwen 2.5 no meaningful correlation exists, a ceiling effect of SVM accuracy remaining near-maximal (96–100%) regardless of the level of semantic degradation.

3.2. Quality Assessment

The dual-metric evaluation reveals systematic divergence between surface fluency (COMET-Kiwi) and semantic preservation (xCOMET), exposing critical limitations in transformation quality assessment (further addressing RQ2 and RQ5). As shown in Table 1, aggressive multilingual edits degrade both fluency and semantics; our companion RTT analysis [19] established that the two remain strongly correlated (Spearman $\rho \geq 0.95$ across pipelines and models), with Qwen 2.5 outperforming Gemma 3 in preservation, though LLM-based methods introduce higher per-passage variability for both models. These patterns confirm that RTT pipelines yielding better fluency retain more meaning, though with sharp absolute divergences and per-passage risks highlighted by standard deviations.

Here, we extend this framework to focus on non-RTT workflows (*Grammarly* 'fix' and 'rephrase', *Parrot*, *Pegasus-PAWS*, and *Undetectable.ai* 'humanize'; $n = 5$ aggregated, $n = 10$ per model), which span lighter monolingual edits and offer diverse insights into controlled versus aggressive transformations. Correlations remain robust, with $\rho \approx 0.92$ at the workflow level ($n = 5$) and $\rho \approx 0.89$ per model ($n = 10$), slightly lower than RTT due to varied editing intensities but still indicate a monotonic link: higher fluency predicts better semantics. For instance, lighter workflows like *Grammarly* (mean COMET-Kiwi ≈ 0.88 – 0.90 , xCOMET ≈ 0.44 – 0.66) and *Parrot* (COMET-Kiwi ≈ 0.83 , xCOMET ≈ 0.24 – 0.46) cluster at the high end, while aggressive ones like *Pegasus-PAWS* (COMET-Kiwi ≈ 0.75 , xCOMET ≈ 0.12 – 0.28) and *Undetectable.ai* 'humanize' (COMET-Kiwi ≈ 0.81 , xCOMET ≈ 0.30 – 0.49) fall lower, demonstrating how more profound restructuring amplifies divergence.

Model-specific patterns underscore source-text influence: Qwen 2.5 consistently achieves 10–20% higher xCOMET than Gemma 3 within workflows, independent of fluency, suggesting greater semantic resilience in Qwen 2.5's outputs. Standard deviation analysis provides further nuance. Semantic dispersion is broadly similar across tools ($\sigma \approx 0.10$ – 0.20 for xCOMET) and does not follow a simple lighter-versus-aggressive gradient: *Parrot* shows the widest spread for Qwen 2.5 ($\sigma = 0.201$), while *Pegasus-PAWS*, the most aggressive rewriter, is comparatively uniform for Gemma 3 ($\sigma = 0.098$). Fluency, by contrast, is far more homogeneous (e.g., *Undetectable.ai* preserves fluency at COMET-Kiwi $\sigma \approx 0.03$ while disrupting semantics variably at xCOMET $\sigma \approx 0.16$). This gap between stable fluency and volatile semantics explains the strong but imperfect correlations: passages can maintain polish while deviating widely in meaning, highlighting why fluency alone is insufficient for assessing obfuscation costs.

Quality distribution analysis employs standardised thresholds for systematic comparison. xCOMET categories span Excellent (≥ 0.85), Good (0.70–0.84), Moderate (0.55–0.69), Poor (0.40–0.54), and Very Poor (< 0.40); COMET-Kiwi uses Excellent (≥ 0.80), Good (0.60–0.79), Moderate (0.40–0.59), Poor (0.20–0.39), and Very Poor (< 0.20). These frameworks establish practical thresholds: xCOMET ≥ 0.55 maintains pooled SVM accuracy at approximately 95% (86% for Gemma 3), whereas xCOMET ≤ 0.30 corresponds to substantial attribution degradation. As noted in Section 2, these thresholds should be read as conservative boundaries. In practical terms, mean scores indicate safer workflows (e.g., *Grammarly* at xCOMET ≈ 0.44 –0.66 with COMET-Kiwi ≈ 0.88 –0.90), while variability quantifies risks, confirming that deeper metrics are essential for evaluating real-world edits.

3.3. Error-Span and Feature Patterns

Token-level span counts supplied by xCOMET reveal the fingerprint of each tool. *Grammarly* ‘fix’ alters only punctuation and agreement, averaging six spans per sentence. *4o-mini* edits 6 to 11 spans while leaving clause skeletons intact. *Pegasus-PAWS* and *googletrans* exceed 13 spans and often rebuild entire sentences. Not all flagged spans necessarily represent semantic failures; the translation-oriented error model can treat legitimate synonym choices and structural reorganisations as errors, so these counts are best interpreted as upper bounds on disruption.

Feature-level analysis clarifies the models’ opposite fates. Gemma 3’s stylometry is tied to sparse vocabulary and balanced punctuation—easily broken by synonym swaps or clause shuffles. Qwen 2.5 embeds its signature in long, regular sentence rhythms (mean sentence length ≈ 20 tokens, versus ≈ 14 for Gemma 3) and function-word ratios, patterns that survive surface edits even when summarisers compress the prose.

Error-span analysis through xCOMET reveals differential transformation strategies across methods and provides architectural insights into model-specific vulnerabilities (addressing RQ5). RTT generates 87–100% error coverage with high span counts (5.84–22.61 average spans), indicating comprehensive lexical and syntactic modification. *4o-mini* produces focused errors (5.84–11.27 average spans) with strategic preservation of semantic core structures, while *googletrans* creates extensive disruption (14.04–22.61 spans) through aggressive cross-linguistic transfer effects.

The Dice coefficient correlates only weakly to moderately with semantic preservation across workflows, confirming that surface-level lexical similarity inadequately captures deeper stylometric disruption patterns. Tool-specific Dice patterns reveal transformation strategies: *Grammarly* ‘fix’ shows high lexical preservation (mean Dice 0.91) with focused modifications, while *Pegasus-PAWS*, as the most aggressive restructuring tool, shows moderate preservation (0.46–0.52).

3.4. A Data-Driven Trade-Off Curve

To formalise the trade-off frontier between semantic quality and stylometric detectability, we swept a quality threshold τ across the full $[0, 1]$ range of xCOMET scores (step = 0.01). For each τ , we calculated coverage $C(\tau)$ as the proportion of passages with xCOMET $\geq \tau$, and accuracy $A(\tau)$ as SVM performance on that retained subset (with 95% Wilson CIs). These metrics were derived from aggregated xCOMET distributions across all edited passages (1650 texts across the nine reported workflows), producing a Pareto curve that quantifies how quality constraints limit the effectiveness of obfuscation.

The curve shows a gradual decline in accuracy as the quality threshold relaxes, from the high-quality regime down to the full dataset ($\tau = 0$, $A = 89\%$). Two clear knees emerge: at $\tau \approx 0.55$ (high-quality regime), 24% of texts survive (95% CI: 22–26) with SVM accuracy at 95% (95% CI: 93–97), Qwen 2.5 at 98% and Gemma 3 at 86%; at $\tau \approx 0.30$, 62% remain

(95% CI: 59–64) with accuracy still at 94% (95% CI: 92–95). Below 0.30, in the range xCOMET classifies as “poor”, overall accuracy falls to 81%, driven entirely by Gemma 3 (73%), while Qwen 2.5 holds at 98%, though some passages in this range may retain more usable meaning than the score implies, given the metric’s tendency to penalise paraphrastic variation. Even so, the associated SVM accuracy drop confirms that heavy obfuscation and quality loss go hand in hand. Model trends diverge sharply: Qwen 2.5’s curve is essentially flat, while Gemma 3’s carries the entire decline.

These empirically derived thresholds supersede rule-of-thumb estimates, confirming that substantial attribution loss necessitates a collapse in semantic fidelity (e.g., xCOMET < 0.30 enables Gemma degradation but at a severe cost in meaning). Qwen 2.5’s rhythm-based fingerprint proves extraordinarily robust, surviving even low-quality edits, highlighting the need for model-tailored obfuscation strategies (addressing RQ4 and RQ5).

4. Discussion

Qwen 2.5 and Gemma 3 behave very differently once their outputs are edited, addressing RQ1 and RQ4 by revealing model-specific vulnerabilities in stylometric signals. Qwen 2.5 retains at least 93% SVM accuracy across all transformations we tried. Its stylometric fingerprint rests on sentence rhythm, regular clause length, and stable function-word ratios, features that survive surface edits and even become sharper after aggressive compression. *Pegasus-PAWS*, for example, reduces semantic fidelity to roughly 28% but increases Qwen 2.5 accuracy to 99%, rather than decreasing it. Gemma 3, in contrast, is highly likely to embed its signature in sparse vocabulary, balanced punctuation, and shallow syntax. Simple synonym swaps or clause shuffles remove these cues, so accuracy slips to 84% with a *Grammarly* ‘rephrase’, falls to 56% under *Pegasus-PAWS*, and reaches the low sixties after the harsh *googletrans* round-trip chains.

Every workflow shows the same quality-obfuscation trade-off (RQ3). When xCOMET stays above 0.55, the meaning is largely preserved, and pooled SVM accuracy remains at 94–95%, though Gemma 3 dips to 86–87%. To push Gemma 3 below 80%, we must let xCOMET decline to about 0.3, at which point the text is barely usable. None of the tested tools crosses that line for Qwen 2.5, so its stylometric signal appears to be distributed across multiple feature families rather than just surface tokens. Even if xCOMET’s absolute scores underestimate the true semantic preservation of some workflows, the fact that Qwen 2.5 remains attributable at 96–100% across all conditions (including those with the lowest quality scores) confirms that its fingerprint is genuinely resilient rather than an artefact of metric strictness. COMET-Kiwi is misleading here because it never drops below 0.65, even when xCOMET shows a severe loss of meaning. Fluency, therefore, hides the damage that deeper metrics reveal.

Conversely, xCOMET may overstate the damage. Because the metric was trained to flag divergence from a reference as an error, it does not distinguish a creative synonym that preserves intent from a mistranslation that distorts meaning. Workflows such as *Parrot* and *Grammarly* ‘rephrase’, which introduce controlled lexical variation, are likely penalised more heavily than their actual semantic cost warrants. The quality-obfuscation frontier we report is therefore conservative: usable text probably extends somewhat beyond the $\text{xCOMET} \leq 0.30$ boundary we label “barely usable.” The relative ordering of workflows and the persistent asymmetry between Qwen 2.5 and Gemma 3 remain robust to this calibration issue, as the rankings hold across every metric and condition we tested.

Transformation tools follow clear optimisation philosophies (further addressing RQ5). *Grammarly* ‘fix’ corrects grammar while leaving style intact. *Grammarly* ‘rephrase’ edits words with only a slight loss of meaning but a noticeable hit to Gemma 3’s attribution. *Parrot* balances moderate loss of meaning with modest stylometric change, whereas *Pegasus-PAWS*

sacrifices meaning to disrupt style. RTT via *4o-mini* retains 40–66% of the original semantics and has a limited effect on detectability, whereas *googletrans* destroys both meaning and style. The commercial service *Undetectable.ai* behaves unpredictably: it reduces Gemma 3's accuracy to 73% while nudging Qwen 2.5 up to a perfect 100%.

5. Conclusions

This study examines the intersection of text quality and the evasion of attribution. Across every workflow, we find the same frontier: when semantic fidelity (xCOMET) remains above 0.55, pooled SVM accuracy stays near 95 percent (86% for Gemma 3); once detection falls to 80 percent or lower, xCOMET scores have typically fallen below 0.30, a region where even allowing for the metric's strictness toward paraphrastic variation, semantic degradation is substantial enough to compromise professional use. This Pareto limit holds regardless of tool or model, giving stylometric detectors a structural advantage.

The two source models illustrate why 'universal' attribution is a myth. Qwen 2.5 encodes its signature in clause rhythm and function-word balance, so even aggressive summarisation or commercial "humanisation" leaves it detectable at 96–100%. Gemma 3 relies on surface vocabulary and punctuation, and its accuracy can drop by 44 points under *Pegasus-PAWS* or by a third after harsh RTT. Detection, therefore, needs model-aware features rather than one-size-fits-all rules.

Feature-engineered academic classifiers retain far more sensitivity in the hardest cases than the commercial service *Undetectable.ai*, whose AI-probability scores collapse to about 1% on its own humanised texts while the SVM still attributes them at 73–100%. As noted in Section 3.1, the algorithmic instability we observed during testing underscores a broader concern: most consumer AI detectors lack transparency policies, version checkpoints, or reproducible models, making them unreliable for scientific or institutional use.

Quality metrics also diverge: COMET-Kiwi fluency rarely falls below 0.65, giving a false sense of safety when xCOMET shows that semantic quality can drop to as low as 0.12. An evaluation pipeline that reports only fluency will miss the true cost of obfuscation. Future work should explore paraphrase-specific evaluation that tolerates synonym substitution and structural reorganisation while remaining sensitive to factual and domain-critical distortions.

Transformation tools follow predictable trade-offs. *Grammarly* 'fix' mode polishes grammar with no obfuscation; its rephrase mode trades a small drop in meaning for a moderate fall in detectability. *Parrot* balances both goals, *Pegasus-PAWS* sacrifices meaning for stronger hiding, and *googletrans* RTT destroys both. Commercial humanisation masks Gemma 3 while leaving Qwen 2.5 fully detectable. Users can select a tool once they know which model they need to protect and how much quality they are willing to sacrifice.

Stakeholders can act on these findings. Universities can combine stylometry with an xCOMET-style quality gate where a reference text is available: texts with scores above 0.55 remain reliably attributable in aggregate, while those with scores below 0.30 and low SVM scores require manual review. Platforms should recognise that commercial detectors lag behind academic ones and upgrade accordingly. Researchers and developers of new language models must decide whether to harden or soften stylometric footprints, as robustness and anonymity pull in opposite directions.

Future work should explore editing methods that reduce detectability without compromising the xCOMET 0.30 quality threshold, investigate which architectural components generate rhythm-based fingerprints, and develop open benchmarks that track the progress of both detectors and evaders over time. Until then, the asymmetry holds: hiding a model's identity incurs a cost in meaning that most real-world authors cannot afford to lose, so stylometric attribution remains a viable line of defence in the AI text arms race.

Author Contributions: Conceptualization, K.K.; methodology, K.K.; software, K.K.; validation, K.K.; investigation, K.K.; resources, K.K.; data curation, K.K.; writing—original draft preparation, K.K.; writing—review and editing, K.K. and T.A.; visualisation, K.K.; supervision, T.A.; project administration, T.A.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: The presented research is supported by the National Science Fund of Bulgaria, with Contract No. KII-06-H85/16.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LLMs	Large Language Models
SVM	Support Vector Machine
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
RTT	Round-Trip Translation
CoPE	Contrastive Perturbation Enhancement
MT	Machine Translation
TF-IDF	Term Frequency–Inverse Document Frequency
BLEU	Bilingual Evaluation Understudy
CI(s)	Confidence Interval(s)
RQ	Research Question
BERT	Bidirectional Encoder Representations from Transformers
PAWS	Paraphrase Adversaries from Word Scrambling
AI	Artificial Intelligence
API	Application Programming Interface

Appendix A

Table A1. Wilson 95% confidence intervals for all workflow × model configurations.

Workflow	Model	SVM Accuracy (%)	Wilson 95% CI (%)
Baseline	Qwen 2.5	98.00	93.0–99.4
	Gemma 3	100.00	96.3–100.0
RTT (googletrans) single-hop chains	Qwen 2.5	99.00	94.6–99.8
	Gemma 3	76.00	66.8–83.3
RTT (4o-mini) single-hop chains	Qwen 2.5	98.00	93.0–99.4
	Gemma 3	86.00	77.9–91.5
RTT (googletrans) multi-hop chains	Qwen 2.5	99.00	94.6–99.8
	Gemma 3	66.00	56.3–74.5
RTT (4o-mini) multi-hop chains	Qwen 2.5	96.00	90.2–98.4
	Gemma 3	78.00	68.9–85.0
Grammarly ‘fix’	Qwen 2.5	96.00	90.2–98.4
	Gemma 3	100.00	96.3–100.0

Table A1. Cont.

Workflow	Model	SVM Accuracy (%)	Wilson 95% CI (%)
Grammarly ‘rephrase’	Qwen 2.5	98.00	93.0–99.4
	Gemma 3	84.00	75.6–89.9
Parrot	Qwen 2.5	97.00	91.5–99.0
	Gemma 3	95.00	88.8–97.8
Pegasus-PAWS	Qwen 2.5	99.00	94.6–99.8
	Gemma 3	56.00	46.2–65.3
Undetectable.ai ‘humanize’ option	Qwen 2.5	100.00	88.6–100.0
	Gemma 3	73.33	55.6–85.8

Table A2. Full McNemar test results with continuity correction.

Workflow	Model	<i>b</i>	<i>c</i>	χ^2	<i>p</i> (Corrected)	Exact <i>p</i>
RTT (googletrans) single-hop chains	Qwen 2.5	0	1	0.00	1.000	1.000
	Gemma 3	24	0	22.04	<0.001	<0.001
RTT (4o-mini) single-hop chains	Qwen 2.5	1	1	0.50	0.480	1.000
	Gemma 3	14	0	12.07	<0.001	<0.001
RTT (googletrans) multi-hop chains	Qwen 2.5	0	1	0.00	1.000	1.000
	Gemma 3	34	0	32.03	<0.001	<0.001
RTT (4o-mini) multi-hop chains	Qwen 2.5	3	1	0.25	0.617	0.625
	Gemma 3	22	0	20.05	<0.001	<0.001
Grammarly ‘fix’	Qwen 2.5	2	0	0.50	0.480	0.500
	Gemma 3	0	0	—	—	—
Grammarly ‘rephrase’	Qwen 2.5	1	1	0.50	0.480	1.000
	Gemma 3	16	0	14.06	<0.001	<0.001
Parrot	Qwen 2.5	2	1	0.00	1.000	1.000
	Gemma 3	5	0	3.20	0.074	0.062
Pegasus-PAWS	Qwen 2.5	0	1	0.00	1.000	1.000
	Gemma 3	44	0	42.02	<0.001	<0.001
Undetectable.ai ‘humanize’ option	Qwen 2.5	0	0	—	—	—
	Gemma 3	8	0	6.12	0.013	0.008

b: baseline correct, workflow incorrect; *c*: baseline incorrect, workflow correct. χ^2 values include the continuity correction; the two-sided exact binomial *p*-value is also given, as it is more reliable for small discordant counts. Dashes: no discordant pairs (test undefined).

References

1. Savoy, J. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*; Springer: Cham, Switzerland, 2020. [CrossRef]
2. Sharma, A.; Nandan, A.; Ralhan, R. An investigation of supervised learning methods for authorship attribution in short Hinglish texts using char & word N-grams. *arXiv* **2018**, arXiv:1812.10281. [CrossRef]
3. Riis, D.E.; Rasmussen, E.D.; Tampakis, P. Identifying Features of Constructive Journalism in News Articles: An Explainable ML Approach. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*; Meo, R., Silvestri, F., Eds.; ECML PKDD 2023. Communications in Computer and Information Science; Springer: Cham, Switzerland, 2025; Volume 2134. [CrossRef]
4. Kopanov, K.; Atanasova, T. A comparative pattern analysis of Qwen 2.5 and Gemma 3 text generation. *WSEAS Trans. Inf. Sci. Appl.* **2025**, *22*, 604–615. [CrossRef]
5. Ruder, S.; Ghaffari, P.; Breslin, J.G. Character-Level and Multi-Channel Convolutional Neural Networks for Large-Scale Authorship Attribution. *arXiv* **2016**, arXiv:1609.06686. [CrossRef]
6. Jafariakinabad, F.; Tarnpradab, S.; Hua, K.A. Syntactic Recurrent Neural Network for Authorship Attribution. *arXiv* **2019**, arXiv:1902.09723. [CrossRef]
7. AlZahrani, F.M.; Al-Yahya, M. A transformer-based approach to authorship attribution in classical Arabic texts. *Appl. Sci.* **2023**, *13*, 7255. [CrossRef]
8. Bauersfeld, L.; Romero, A.; Muglikar, M.; Scaramuzza, D. Cracking double-blind review: Authorship attribution with deep learning. *PLoS ONE* **2023**, *18*, e0287611. [CrossRef] [PubMed]

9. Brennan, M.; Afroz, S.; Greenstadt, R. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.* **2012**, *15*, 12. [CrossRef]
10. Mondal, S.K.; Zhang, H.; Kabir, H.M.D.; Ni, K.; Dai, H.-N. Machine translation and its evaluation: A study. *Artif. Intell. Rev.* **2023**, *56*, 10137–10226. [CrossRef]
11. Zhuo, T.Y.; Xu, Q.; He, X.; Cohn, T. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 319–337. Available online: <https://aclanthology.org/2023.findings-acl.22> (accessed on 1 May 2025).
12. Ahmed, W.; Javed, A.R.; Jalil, Z.; Iqbal, F. Authorship analysis with machine learning. In *Encyclopedia of Machine Learning and Data Science*; Phung, D., Webb, G.I., Sammut, C., Eds.; Springer: New York, NY, USA, 2022. [CrossRef]
13. Nguyen-Quoc, T.; Nguyen-Cong, H.; Le-Thanh, H. Contrastive Perturbation Enhancement for LLM-Based Machine Translation. In *Proceedings of the Information and Communication Technology*; Buntine, W., Fjeld, M., Tran, T., Tran, M.T., Huynh Thi Thanh, B., Miyoshi, T., Eds.; SOICT 2024. Communications in Computer and Information Science; Springer: Singapore, 2025; Volume 2351. [CrossRef]
14. Rei, R.; Guerreiro, N.M.; Pombal, J.; van Stigt, D.; Treviso, M.; Coheur, L.; de Souza, J.G.C.; Martins, A.F.T. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. *arXiv* **2023**. [CrossRef]
15. Guerreiro, N.M.; Rei, R.; van Stigt, D.; Coheur, L.; Colombo, P.; Martins, A.F.T. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 979–995. [CrossRef]
16. Pombal, J.; Yoon, D.; Fernandes, P.; Wu, I.; Kim, S.; Rei, R.; Neubig, G.; Martins, A.F.T. M-Prometheus: A suite of open multilingual LLM judges. *arXiv* **2025**. [CrossRef]
17. Prithivida. “Prithivida/Parrot_Paraphraser_on_T5”, Hugging Face. Available online: https://huggingface.co/prithivida/parrot_paraphraser_on_T5 (accessed on 1 May 2025).
18. Domenicrossati. “Domenicrossati/Pegasus-Xsum-Finetuned-Paws”, Hugging Face. Available online: <https://huggingface.co/domenicrossati/pegasus-xsum-finetuned-paws> (accessed on 1 May 2025).
19. Kopanov, K.; Atanasova, T. Evaluating semantic preservation in multilingual round-trip translation. In *Proceedings of the 39th Annual European Simulation and Modelling Conference (ESM’2025)*, Ghent, Belgium, 22–24 October 2025; EUROSIS-ETI: Ostend, Belgium, 2025; pp. 87–94.
20. Han, S. Googletrans (Version 4.0.2). GitHub. 2024. Available online: <https://github.com/ssut/py-googletrans> (accessed on 1 May 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.