# STAMP: Stylometric Text Anonymization with Memory-guided Policy Optimization

Zhan Shi
*Santa Clara University*
*Santa Clara, USA*
*Email: ashi2@scu.edu*

Yefeng Yuan
*Santa Clara University*
*Santa Clara, USA*
*Email: yyuan4@scu.edu*

Liang Cheng
*eBay Inc.*
*San Jose, USA*
*Email: liacheng@ebay.com*

Yuhong Liu
*Santa Clara University*
*Santa Clara, USA*
*Email: yhliu@scu.edu*

***Abstract*—Textual data used in modern ML systems frequently contains sensitive personal information. While conventional anonymization removes explicit identifiers, texts remain vulnerable to authorship inference attacks that exploit implicit stylometric fingerprints. Existing LLM-based rewriting methods apply rigid, one-size-fits-all obfuscation ignore stylometric outliers and incur high deployment cost-yielding suboptimal privacy utility trade-offs. We propose STAMP (Stylometric Text Anonymization with Memory-guided Policy Optimization), a reinforcement learning framework that trains a lightweight local rewriter to adaptively balance privacy and utility at the instance level. A teacher LLM bootstraps high-quality rewrites; a lightweight model is then fine-tuned via policy optimization guided by a Style Manifold Memory (SMM), which tracks population-level style distributions, detects high-risk outliers, and adjusts reward weighting accordingly applying aggressive obfuscation to highly identifiable texts while preserving semantic utility for low-risk ones. Experiments show STAMP substantially reduces authorship re-identification while consistently outperforming baselines on downstream task performance.**

## 1. Introduction

The widespread deployment of LLMs in sensitive domains including healthcare, finance, and social media amplifies the risk of exposing personal information embedded in text. A key threat is *stylometric deanonymization*: authorship classifiers exploit writing-style fingerprints to re-identify authors even after surface-level redaction [1].

Existing approaches have two critical gaps. First, they treat all texts uniformly, ignoring that *stylometric outliers* which are in sparse, distinctive style regions are far more vulnerable to re-identification and require more aggressive obfuscation; a one-size-fits-all policy unnecessarily distorts low-risk texts and under-protects high-risk ones. Second, LLM-based anonymizers rely on large models at inference time, limiting scalability [2].

STAMP addresses both gaps via a **Style Manifold Memory (SMM)**, which models population-level style geometry through prototype clusters, estimates instance-level privacy risk from geometric isolation, and retrieves target styles for direction-aware rewriting. A two-stage teacher–student pipeline first produces a high-quality bootstrapping corpus, then refines a lightweight local model via risk-adaptive RL yielding stronger privacy–utility trade-offs at substantially lower deployment cost.

## 2. Methodology

**Problem formulation.** Let $\mathcal{A}$ be a candidate author set. Given document $x$ from author $a \in \mathcal{A}$, we seek a policy $\pi_\theta$ that produces rewrite $y$ obscuring authorship while preserving semantic content. An attacker uses $f_{\text{auth}} : \mathcal{Y} \to \Delta^{|\mathcal{A}|}$ to infer authorship from released texts.

**Privacy and utility objectives.** Rather than merely inducing misclassification, we maximize the uncertainty of $f_{\text{auth}}$ via entropy, combined with geometry-aware style guidance and identifier suppression:

$$R_{\text{ent}}(y) = -\sum_{a' \in \mathcal{A}} P(a' \mid y) \log P(a' \mid y),$$
$$R_{\text{priv}}(x, y) = R_{\text{ent}}(y) + R_{\text{aux}}(x, y).$$

Semantic fidelity is captured by $R_{\text{util}}(x, y)$. The full objective is:

$$\max_\theta \; \mathbb{E}_{x,\, y \sim \pi_\theta}\Big[\lambda_{\text{util}}\, R_{\text{util}}(x, y) + \lambda_{\text{priv}}(x)\, R_{\text{priv}}(x, y)\Big],$$

where the instance-dependent weight $\lambda_{\text{priv}}(x) = w_0 + \alpha\, w(x)$, with $w(x) \in [0, 1]$ derived from the SMM outlier score, increases privacy pressure for high-risk outliers.

**Style Manifold Memory.** To capture population-level stylometric structure, we introduce the Style Manifold Memory (SMM), a structured memory over the global style distribution. SMM fulfills three roles: (1) modeling global style geometry via prototypes, (2) estimating instance-level privacy risk, and (3) retrieving target styles for direction-aware anonymization.

Authorship classifiers achieve higher accuracy on texts in sparse style regions [3]. We treat geometric isolation in a prototype-based style space as a tractable proxy for re-identification risk. The SMM maintains a prototype set $\mathcal{M} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ over style embeddings. For each prototype, we compute its mean pairwise distance to all others; an outlier threshold is then set as $\tau = \text{mean}(\bar{d}_k) + \lambda \cdot \text{std}(\bar{d}_k)$. An input $x$ is classified as a stylometric outlier if its mean

divergence $\bar{d}_x = \frac{1}{K}\sum_k d_{\cos}(z_x, \boldsymbol{\mu}_k)$ exceeds $\tau$. The normalized risk score

$$w(x) = \mathrm{clip}\Big(\tfrac{\bar{d}_x - d_{\min}}{d_{\max} - d_{\min}}, 0, 1\Big)$$

drives adaptive reward weighting. The SMM also retrieves a target prototype $\boldsymbol{\mu}(x)$: for outliers, a prototype from a denser region; for inliers, a nearby but stylistically distinct one.

**Teacher–student bootstrapping.** Directly applying RL to a lightweight LLM is unstable. We first construct a high-quality SFT corpus: the SMM provides a target prototype $\boldsymbol{\mu}(x)$ conditioned on $w(x)$, which is fed alongside $x$ to a frozen teacher LLM to generate candidates. We retain only candidates satisfying $\mathrm{BERTScore}(x, y) \geq \delta_{\mathrm{sem}}$ and select

$$y^* = \arg\max_y \Big[d_{\cos}(z_x, z_y) - \gamma\, d_{\cos}(z_y, \boldsymbol{\mu}(x))\Big],$$

i.e., the rewrite that maximizes stylistic departure from the source while aligning to the target prototype. A Llama-3.2-3B student is then fine-tuned on these $(x, y^*)$ pairs.

**Reward design.** Starting from the SFT checkpoint, GRPO [4] optimizes a composite reward with LoRA adapters. The privacy reward has three components [5]:

- *Entropy*: $r_{\mathrm{ent}}(y) = H(f_{\mathrm{auth}}(y))/\log n_c$, maximizing author attribution uncertainty.
- *Style*: $r_{\mathrm{style}}(x, y) = d_{\cos}(z_x, z_y) - \mathbf{1}[x \in \mathcal{O}] \cdot d_{\cos}(z_y, \boldsymbol{\mu}(x))$, encouraging stylistic departure and, for outliers, prototype alignment.
- *Entity suppression*: $r_{\mathrm{entity}}(x, y) = -|\mathrm{NER}(x) \cap \mathrm{NER}(y)|/|\mathrm{NER}(x)|$, penalizing retained named entities.

Utility is measured by BERTScore: $r_{\mathrm{sem}}(x, y) = \max(0, (\mathrm{BS}_{F1}(x, y) - f)/(1 - f))$ with floor $f = 0.80$. The final reward is

$$r_{\mathrm{total}} = \lambda_{\mathrm{priv}}(x) \underbrace{[r_{\mathrm{ent}} + r_{\mathrm{style}} + r_{\mathrm{entity}}]}_{r_{\mathrm{priv}}} + \lambda_{\mathrm{util}}\, r_{\mathrm{sem}}.$$

**Memory-guided inference.** After RL training, the SMM is reconstructed over the finalized prototype pool to recalibrate the global style manifold and threshold $\tau$. At inference, we compute $\bar{d}_x$ for an unseen text $x$ against $\mathcal{M}$. For outliers ($\bar{d}_x > \tau$), we retrieve a reference prototype sampled from the top-$K$ most visited prototypes, pulling the text toward denser, lower-risk regions of the manifold. For inliers ($\bar{d}_x \leq \tau$), we select a prototype that maximizes stylistic distance from $z_x$, encouraging sufficient variation while staying within the densely populated style region.

**Experimental setup.** We evaluate on YELP, TWITTER, and IMDB (author-ID and gender inference as privacy tasks; sentiment as utility), plus SYNTHPAI [6] for attribute leakage. Baselines include PRESIDIO [7], STYLEMIX, DIPPER [8], DP-MLM [9], and TAROT [2].

STAMP achieves best or second-best privacy across all datasets while leading on utility. On SYNTHPAI attribute leakage, age and education inference drop to 0.0943 and 0.0776 respectively; gender leakage remains higher across all methods, indicating gender-linked stylistic markers are harder to suppress.

TABLE 1. PRIVACY (ATTACKER MACRO-$F_1$, ↓) AND UTILITY (SENTIMENT MACRO-$F_1$, ↑). BEST IN **BOLD**, SECOND UNDERLINED.

| | **Privacy (↓)** | | | **Utility (↑)** | | |
|---|---|---|---|---|---|---|
| **Method** | **Yelp** | **Twitter** | **IMDb** | **Yelp** | **Twitter** | **IMDb** |
| Original | 0.892 | 0.914 | 0.867 | 0.924 | 0.885 | 0.912 |
| Presidio | 0.804 | 0.893 | 0.824 | 0.857 | 0.823 | 0.834 |
| StyleMix | 0.584 | 0.497 | 0.603 | 0.779 | 0.813 | 0.802 |
| DIPPER | 0.674 | 0.527 | 0.681 | 0.723 | 0.696 | 0.761 |
| DP-MLM | 0.664 | 0.603 | 0.470 | 0.723 | 0.750 | 0.758 |
| TAROT-PPO | 0.515 | 0.641 | 0.889 | 0.805 | 0.791 | 0.775 |
| TAROT-DPO | 0.307 | 0.550 | **0.347** | 0.772 | 0.764 | 0.571 |
| **Ours** | **0.292** | **0.311** | 0.352 | **0.809** | **0.814** | **0.819** |

## 3. Conclusion

STAMP is a risk-adaptive RL framework for privacy-preserving text rewriting. By combining teacher–student bootstrapping with SMM-guided policy optimization, it differentiates obfuscation pressure between stylometric outliers and inliers, achieving strong privacy protection without sacrificing utility. Limitations include reliance on a non-adaptive attacker model and the use of geometric isolation as a proxy for formal privacy risk.

## References


[1] T. Deußer, L. Sparrenberg, A. Berger, M. Hahnbück, C. Bauckhage, and R. Sifa, "A survey on current trends and recent advances in text anonymization," in *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2025, pp. 1–9.

[2] G. Loiseau, D. Sileo, D. Riquet, M. Meyer, and M. Tommasi, "Tarot: Task-oriented authorship obfuscation using policy optimization methods," in *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, 2025, pp. 14–31.

[3] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast, "Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection," in *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, 2018, pp. 1–25.

[4] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[5] Z. Shi, Y. Yuan, L. Cheng, and Y. Liu, "Reinforcement learning-guided large language model fine-tuning for privacy-preserving text rewriting," in *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing*, 2025, pp. 1–7.

[6] H. Yukhymenko, R. Staab, M. Vero, and M. Vechev, "A synthetic dataset for personal attribute inference," *Advances in Neural Information Processing Systems*, vol. 37, pp. 120 735–120 779, 2024.

[7] Microsoft, "Presidio: Data protection and de-identification sdk," 2019. [Online]. Available: https://github.com/microsoft/presidio

[8] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.

[9] S. Meisenbacher, M. Chevli, J. Vladika, and F. Matthes, "Dp-mlm: Differentially private text rewriting using masked language models," 2024. [Online]. Available: https://arxiv.org/abs/2407.00637

# STAMP: Stylometric Text Anonymization with Memory-guided Policy Optimization

Zhan Shi[1], Yefeng Yuan[1], Liang Cheng[2], Yuhong Liu[1]

[1]Santa Clara University, [2]eBay

## Abstract

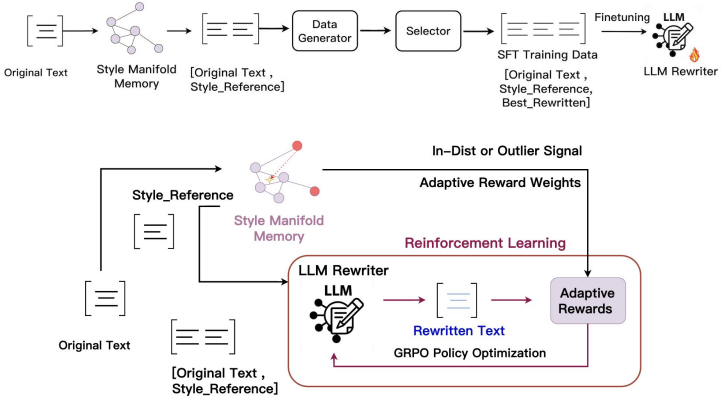
Text data often leaks sensitive information through implicit stylometric fingerprints, making it vulnerable to authorship inference attacks. Existing text anonymization rewriting methods apply uniform obfuscation, failing to account for stylometric outliers and leading to suboptimal privacy–utility trade-offs.

We propose **STAMP**, a reinforcement learning framework that leverages a **Style Manifold Memory (SMM)** to model population-level style distributions and adaptively adjust rewriting strength at the instance level. Experiments show that STAMP significantly reduces re-identification risk while preserving downstream utility.

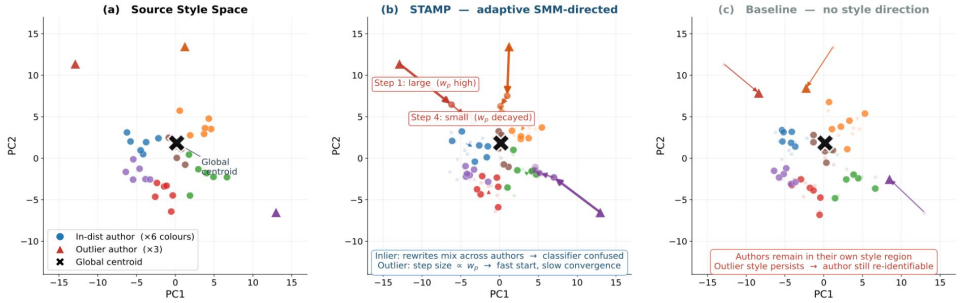
**Attack Model** The attacker is modeled as a stylometric classifier that infers authorship from rewritten text, evaluated under three increasingly strong threat models: (1) a baseline closed-set BERT attacker trained on rewrites, (2) a StyleEmb transfer attacker trained on frozen style embeddings from original texts and tested on rewrites, and (3) an adaptive BERT attacker fine-tuned on a subset of rewrites.

STAMP achieves better authorship F1 across all settings and shows the smallest degradation under adaptation, demonstrating strong robustness.

## Methodology




## Motivation




## Evaluation

| Method | Yelp | | Twitter | | IMDb | |
|---|---|---|---|---|---|---|
| | Privacy $F_1$↓ | Utility↑ | Privacy $F_1$↓ | Utility↑ | Privacy $F_1$↓ | Utility↑ |
| Original | 0.8915 | 0.9240 | 0.9142 | 0.8850 | 0.8672 | 0.9120 |
| Presidio | 0.8041 | 0.8571 | 0.8925 | 0.8227 | 0.8237 | 0.8341 |
| StyleMix | 0.5837 | 0.7792 | 0.4970 | 0.8130 | 0.6025 | 0.8020 |
| DIPPER | 0.6737 | 0.7227 | 0.5274 | 0.6962 | 0.6812 | 0.7614 |
| DP-MLM | 0.6640 | 0.7231 | 0.6027 | 0.7500 | 0.4695 | 0.7583 |
| TAROT-PPO | 0.5149 | 0.8051 | 0.6412 | 0.7912 | 0.8889 | 0.7754 |
| TAROT-DPO | 0.3074 | 0.7715 | 0.5500 | 0.7635 | **0.3474** | 0.5714 |
| **Ours** | **0.2917** | **0.8085** | **0.3107** | **0.8137** | 0.3519 | **0.8192** |

Table 1: Main results. Privacy is attacker macro-$F_1$: Author-ID for Yelp/IMDb and Gender for Twitter (lower is better). Utility is downstream sentiment macro-$F_1$ (higher is better). **Bold** and underlined values denote best and second-best among generative methods (excluding Original and Presidio).