STAMP-R: Stylometric Text Anonymization with Memory-guided Policy Rewriting

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

[1]Santa Clara University, Santa Clara, USA [2]eBay Inc., San Jose, USA

ashi2@scu.edu yyuan4@scu.edu liacheng@ebay.com yhliu@scu.edu

Abstract

Modern machine learning systems rely heavily on large-scale textual data that often contain sensitive personal information. Although conventional anonymization techniques remove explicit identifiers, textual data remain vulnerable to authorship inference attacks that exploit persistent stylometric signals. Recent approaches leverage Large Language Models (LLMs) to rewrite text and obscure such signals, but they frequently overlook distinctive stylometric outliers and fail to achieve a favorable privacy–utility trade-off due to rigid, one-size-fits-all obfuscation strategies, while also incurring high computational costs. To address these challenges, we propose STAMP-R, a risk-adaptive reinforcement learning framework for instance-level authorship anonymization. We formulate anonymization as a risk-aware, instance-level style distribution shaping problem. Central to our approach is the Style Manifold Memory (SMM), which models the global stylistic landscape via prototype-based density estimation. SMM detects high-risk stylometric outliers and adaptively modulates a composite reward function, enabling stronger obfuscation for highly identifiable samples while preserving semantic fidelity for low-risk instances. We further distill a lightweight 3B-parameter model from a teacher LLM for efficient local deployment. Experiments show that STAMP-R reduces authorship re-identification risk while maintaining strong downstream utility.

1 Introduction

The rapid improvement of large language models (LLMs) has led to their widespread application in sensitive domains such as healthcare, finance, and social media. While driving significant technological progress, this trend concurrently amplifies the risk of exposing sensitive user information embedded in textual data. Text anonymization plays a critical role in mitigating these concerns by protecting individual privacy while preserving the utility of data for downstream applications. (Deußer et al., 2025)

Traditional text anonymization primarily focuses on masking personally identifiable information (PII), such as names, bank accounts, or home addresses. However, recent studies demonstrate that removing PIIs is far from adequate (Xin et al., 2025; Shi et al., 2025). While Differential Privacy (DP) (Dwork et al., 2006) offers formal theoretical guarantees via noise injection, directly applying DP to the high-dimensional space of natural language often leads to substantial losses in fluency and semantic coherence (Krishna et al., 2023). Moreover, writing style is a strong biometric indicator (Huang et al., 2025), and modern attackers can easily perform author re-identification by exploiting contextual signals and unique stylistic fingerprints, such as vocabulary preferences and sentence structure. (Habib et al., 2025; Xing et al., 2024)

LLM-based paraphrasing has emerged as a promising approach for obscuring writing fingerprints while preserving semantic meaning. However, most existing methods focus on increasing stylistic divergence from the original text without considering the text's position relative to the broader population style distribution. As a result, they often overlook privacy risks associated with stylometric outliers, whose rare and distinctive writing patterns make them substantially more vulnerable to re-identification than typical in-distribution texts. Simply moving a text away from its original style does not guarantee anonymity. (See Appendix B for empirical evidence.)

More fundamentally, privacy risk is not an intrinsic property of a text, but depends on its relationship to the surrounding population. We model stylometric privacy risk as a population-level density

Proceedings of the Seventh Workshop on Privacy in Natural Language Processing, pages 53–68

problem: a text becomes indistinguishable when it lies in a dense region of the style space, rather than merely moving away from its original style. Texts in dense regions naturally blend into common patterns, while those in sparse regions remain highly identifiable due to their distinctiveness. This perspective implies that anonymization should not be treated as a uniform rewriting problem. One-size-fits-all strategies lead to suboptimal privacy–utility trade-offs: aggressive rewriting can unnecessarily distort low-risk texts, while insufficient rewriting fails to protect high-risk outliers.

In addition, many LLM-based anonymizers rely on large inference-time models, incurring substantial computational costs in realistic deployment settings (Loiseau et al., 2025). Taken together, these limitations reveal three key gaps: lack of differentiation between outliers and inliers, absence of population-aware risk modeling, and poor scalability.

To fill these gaps, we propose **STAMP-R** (**S**tylometric **T**ext **A**nonymization with **M**emory-guided **P**olicy Rewriting), a risk-adaptive framework for privacy-preserving rewriting. STAMP-R uses author-classifier entropy as the primary privacy objective, augmented with geometry-aware supervision in the style embedding space to control both the magnitude and direction of stylistic transformation. At the core of the framework is a **Style Manifold Memory (SMM)**, which models the population-level distribution of writing styles, identifies stylometric outliers, estimates instance-level privacy risk, and retrieves prototype targets for directional rewriting. Based on these signals, STAMP-R adaptively adjusts privacy pressure across inputs while preserving semantic meaning. To enable scalable deployment, STAMP-R adopts a two-stage teacher–student pipeline: a teacher LLM first bootstraps a high-quality rewriting corpus, and a lightweight local model is subsequently refined via risk-adaptive reinforcement learning. Together, these designs yield a stronger privacy–utility trade-off with substantially lower deployment cost.

Contributions. The main contributions of this work are summarized as follows:

- We introduce the **Style Manifold Memory (SMM)**, a population-level style memory designed to identify **stylometric outliers** and provide instance-level privacy risk signals. By modeling the global distribution of writing styles, SMM offers explicit geometric guidance that steers high-risk texts toward safer regions of the style space and enables risk-aware optimization.

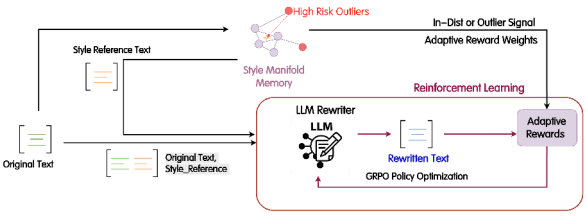

Figure 1: Risk-adaptive Policy Optimization via GRPO.

- To address the limitations of one-size-fits-all anonymization, we propose **STAMP-R**, a risk-adaptive rewriting framework that models the privacy-utility trade-off at the **instance level**. By dynamically scaling the anonymization pressure, STAMP-R prevents the over-sanitization of low-risk texts while enforcing aggressive protection for highly vulnerable instances.

- To reduce the deployment cost of LLM-based anonymization, we propose a two-stage training pipeline for a **lightweight local rewriter**. Empirically, the resulting model achieves strong privacy–utility trade-offs with substantially lower inference cost, enabling practical local deployment.

2 Related Work

Traditional and DP-based text anonymization. Early text anonymization methods mainly rely on rule-based redaction of personally identifiable information (PII) (Microsoft, 2019). However, removing explicit identifiers alone is insufficient, since authors can still be re-identified through implicit stylometric signals. Differential privacy (DP) provides formal guarantees by injecting noise during training or decoding (Dwork et al., 2006; Abadi et al., 2016; Flemings et al., 2024). Yet, when applied to natural language, such noise often degrades fluency, semantic fidelity, and readability, and is not specifically designed to address stylometric authorship inference (Lukas et al., 2023; Chim et al., 2025). In contrast, our method directly targets author-identifying stylistic cues while preserving downstream semantic utility.

LLM-based anonymization and privacy–utility trade-offs. Recent work reframes privacy protection as a controlled rewriting problem, where LLMs generate semantically similar text while suppressing identifying cues (Miranda et al., 2024). Existing methods such as controllable paraphrasing and RL-based anonymization improve the privacy–utility trade-off compared with redaction-only baselines, but they typically optimize uniform objectives across all inputs (Krishna et al., 2024; Chong et al., 2025; Yang et al., 2025; Kim et al., 2025; Loiseau et al., 2025). As a result, they overlook heterogeneity in stylometric risk: moving a text away from its original style does not ensure that it leaves a sparse, highly identifiable region of the population style space.

The closest prior work, TAROT (Loiseau et al., 2025), maximizes stylistic distance via RL but provides no mechanism to control *where* the rewrite lands in the style manifold, applies uniform reward weights, and cannot distinguish outlier from inlier inputs. STAMP-R adds three components absent from prior RL anonymizers: (1) a population-level SMM that models the global style manifold; (2) a geometric risk score for outlier detection; and (3) instance-adaptive reward weighting — together forming a unified, risk-adaptive framework not addressed by prior work.

3 Methodology

3.1 Attack Model

In this work, we focus on authorship re-identification attacks. Specifically, we consider a candidate author set $\mathcal{A}$ consisting of m authors. Each author may produce one or more text paragraphs that are rewritten by the system before being released for downstream tasks such as sentiment analysis or product recommendation. We assume that the attacker has access to prior text samples written by these authors. Given a released document y, the attacker's objective is to identify its true author with high confidence.

The attack model can be formulated as a classification model

$$f_{\text{auth}} : \mathcal{Y} \to \Delta^{|\mathcal{A}|},$$

which produces a probability distribution over candidate authors for any input text.

Attacker scope. Following prior work (Loiseau et al., 2025; Yang et al., 2025), our primary evaluation uses *non-adaptive* attackers trained on clean originals. We additionally evaluate stronger threat models in Appendix A.

3.2 Problem Statement

Given a document x written by author $a \in \mathcal{A}$, our goal is to learn a rewriting policy that produces a text y which obscures authorship while preserving semantic content for downstream use.

We assume access to an attribution model $f_{\text{auth}}(\cdot)$ that outputs a predictive distribution over candidate authors. Instead of merely inducing misclassification, we maximize the uncertainty of this distribution via entropy:

$$H(y) = -\sum_{a' \in \mathcal{A}} P(a' \mid y) \log P(a' \mid y),$$
$$P(\cdot \mid y) = f_{\text{auth}}(y).$$

Note that maximizing $H(y)$ is equivalent to minimizing the KL divergence from the author predictor to a uniform distribution:

$$H(y) = \log n_c - \mathrm{KL}\big(f_{\text{auth}}(y) \,\big\|\, \mathcal{U}\big)\,,$$

where $\mathcal{U}$ is the uniform distribution over n_c authors. Maximizing entropy is therefore equivalent to minimizing the *attacker's belief concentration*. Entropy alone, however, does not control the *direction* of stylistic change in the global style space. A text may lose author-specific style yet remain in a sparse region—still trivially re-identifiable by any attacker who has seen a few examples from that region.

We address this by incorporating stylometric *density* into the privacy objective. Let $r(x) = -\log \hat{p}_{\text{style}}(x)$ denote the negative log population-style-density at source x (Section 3.3). We define the privacy objective as

$$R_{\text{priv}}(x, y) = R_{\text{ent}}(y) + R_{\text{geom}}(x, y) + R_{\text{ner}}(x, y),$$

where $R_{\text{ent}}(y) = 1 - \mathrm{KL}(f_{\text{auth}}(y) \| \mathcal{U}) / \log n_c$ measures attacker uncertainty, $R_{\text{geom}}(x, y)$ provides geometry-aware style guidance toward denser regions, and $R_{\text{ner}}(x, y)$ suppresses explicit named-entity identifiers.

Combining with the utility objective, and scaling privacy pressure by the instance-level risk, the full anonymization problem is:

$$\max_{\theta}\ \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta(\cdot|x)} \Big[\lambda_{\text{util}}\, R_{\text{util}}(x, y) + \underbrace{\alpha(x)}_{\propto r(x)}\, R_{\text{priv}}(x, y)\Big],$$

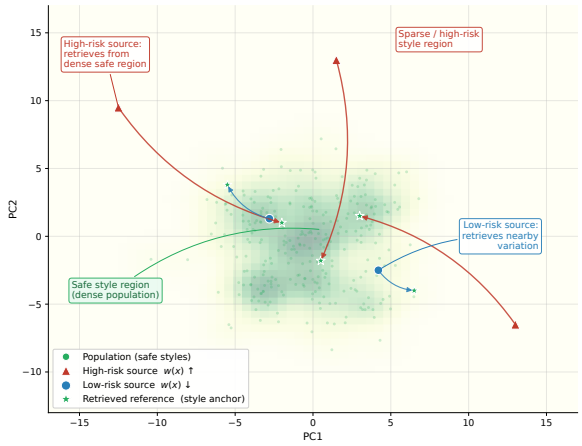

Figure 2: SMM-guided target stylistic prototype selection.

where $R_{\text{util}}(x, y)$ measures semantic fidelity and $\alpha(x) \propto r(x) = -\log \hat{p}_{\text{style}}(x)$ is the **density-derived risk weight**: texts with low population-style-density (high $r(x)$) receive proportionally stronger privacy pressure. In practice we parameterize $\alpha(x) = w_0 + \alpha_0 w(x)$ where $w(x) = \text{clip}(r(x)/r_{\max}, 0, 1)$ is the normalized risk, $w_0 > 0$ ensures a non-zero privacy floor, and $\alpha_0 \geq 0$ controls risk sensitivity. Section 3.3 defines $\hat{p}_{\text{style}}$ and Section 3.5.1 instantiates the reward components.

3.3 Style Manifold Memory

We treat geometric isolation in the prototype-based style space as an *empirical proxy* for stylometric privacy risk. This framing is grounded in two complementary observations. First, texts in sparse stylistic regions are stylometrically *near-unique*: formally, a text x satisfies k-anonymity (Sweeney, 2002) in the stylometric sense only if at least k other texts share a sufficiently similar stylistic neighborhood. A large mean prototype distance $\bar{d}_x$ implies a near-empty neighborhood ($k \approx 1$), making x trivially re-identifiable from style alone. Our outlier threshold τ therefore approximates a k-anonymity boundary: texts with $\bar{d}_x > \tau$ have $k \approx 1$ and require aggressive rewriting to blend into the population. Second, the nearest-neighbor theory of classification shows that a 1-NN classifier's error rate approaches the Bayes error in the limit; in sparse regions, the large inter-class margin gives the attacker high confidence, whereas in dense regions overlapping styles dilute that margin. Taken together, geometric sparsity is a practical, tractable surrogate for attribution vulnerability. While it does not constitute a formal privacy guarantee and does not account for representation-aware attackers who may retrain their classifiers on the style space itself, it enables risk-adaptive rewriting without requiring a formal attacker model at training time. To capture population-level stylometric structure, we introduce the Style Manifold Memory (SMM), a structured memory over the global style distribution. Rather than serving as a simple cache of historical styles, 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.

We extract stylometric embeddings from a reference corpus using a frozen style encoder and construct a prototype memory

$$\mathcal{M} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\},$$

where each prototype $\boldsymbol{\mu}_k$ represents a cluster centroid of the global style space (see Algorithm 2 for construction details).

Population density estimation. SMM defines the *population style density* at a point z_x via a prototype-based Parzen-window estimator:

$$\hat{p}_{\text{style}}(x) \propto \frac{1}{K} \sum_{k=1}^{K} \exp\left(-\frac{d_{\cos}(z_x, \boldsymbol{\mu}_k)}{2\sigma^2}\right),$$

where $\sigma > 0$ is a bandwidth hyperparameter. The instance-level risk is then the **negative log-density**:

$$r(x) = -\log \hat{p}_{\text{style}}(x).$$

For a tractable implementation, we note that the leading-order approximation of $-\log \hat{p}$ is monotone in the mean prototype distance $\bar{d}_x$ (see Appendix F.3), so we use

$$\bar{d}_x = \frac{1}{K} \sum_{k=1}^{K} d_{\cos}(z_x, \boldsymbol{\mu}_k)$$

as a computationally efficient surrogate for $r(x)$. Texts with large $\bar{d}_x$ occupy sparse, low-density regions of the style manifold and are therefore assigned high risk; texts with small $\bar{d}_x$ sit in dense, anonymous-looking regions and carry low risk.

Outlier detection. To quantify regional sparsity across prototypes, we compute the mean pairwise inter-prototype distance:

$$\bar{d}_k = \frac{1}{K-1} \sum_{k' \neq k} d_{\cos}(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'}).$$

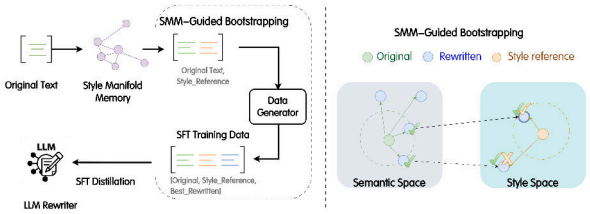

Figure 3: SMM-Guided Bootstrapping and SFT Initialization.

We then define an outlier threshold

$$\tau = \text{mean}\big(\{\bar{d}_k\}_{k=1}^{K}\big) + \lambda \cdot \text{std}\big(\{\bar{d}_k\}_{k=1}^{K}\big),$$

and classify x as a stylometric outlier if $\bar{d}_x > \tau$:

$$\mathcal{O} = \{x \mid \bar{d}_x > \tau\}.$$

Under the density view, this threshold approximates the boundary of the k-anonymous region: texts with $\bar{d}_x > \tau$ have near-empty stylistic neighborhoods ($k \approx 1$) and require aggressive rewriting to blend into the population.

We normalize $\bar{d}_x$ into a bounded risk score

$$w(x) = \text{clip}\left(\frac{\bar{d}_x - d_{\min}}{d_{\max} - d_{\min}},\, 0,\, 1\right),$$

which serves as $w(x)$ in the risk weight $\alpha(x) = w_0 + \alpha_0 w(x)$ from Section 3.2.

Finally, SMM retrieves a target prototype $\boldsymbol{\mu}(x) \in \mathcal{M}$ to guide rewriting. For outliers, it retrieves a prototype from a denser region of the manifold; for inliers, it retrieves a distinct but nearby prototype to encourage moderate stylistic variation without excessive semantic distortion.

3.4 SMM-Guided Bootstrapping and SFT Initialization

Directly optimizing privacy and utility with reinforcement learning is challenging for lightweight LLMs due to unstable exploration and high-variance rewards. We therefore first construct a high-quality synthetic corpus with a teacher–student pipeline.

As shown in Figure 3, for each source text x, the SMM retrieves a target prototype $\boldsymbol{\mu}(x)$ conditioned on its risk level $w(x)$. We then feed x and $\boldsymbol{\mu}(x)$ into a frozen teacher LLM, which samples a set of candidate rewrites $\mathcal{Y} = \{y_1, \ldots, y_N\}$ using stochastic decoding.

We retain only candidates that satisfy a semantic constraint,

$$s_{\text{sem}}(y) = \text{BERTScore}(x, y) \geq \delta_{\text{sem}},$$

and select the final rewrite using a style fitness score that favors both departure from the source style and alignment with the retrieved prototype:

$$s_{\text{style}}(y) = d_{\cos}(z_x, z_y) - \gamma\, d_{\cos}(z_y, \boldsymbol{\mu}(x)).$$

The selected target is

$$y^* = \arg\max_{y \in \mathcal{Y}} s_{\text{style}}(y) \quad \text{s.t.} \quad s_{\text{sem}}(y) \geq \delta_{\text{sem}}.$$

We then perform supervised fine-tuning on the resulting (x, y^*) pairs using a compact student model initialized from Llama-3.2-3B.

3.5 Risk-adaptive Policy Optimization via GRPO

Starting from the SFT-initialized student, we further optimize the rewriter with *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024). For each input x, the SMM provides two signals: an instance-level risk score $w(x) \in [0, 1]$ and a prototype-guided stylistic target.

As shown in Figure 1, Conditioned on the source text and retrieved reference, the policy generates multiple rewrite candidates, which are scored by a composite reward balancing privacy and semantic fidelity. Privacy-oriented rewards are scaled by $w(x)$ so that high-risk inputs receive stronger anonymization pressure. GRPO then uses the resulting relative advantages to update the policy.

We optimize only the LoRA (Hu et al., 2022) adapter parameters while keeping the base model frozen. Optionally, high-quality rewrites are added back into the SMM to refresh the memory distribution during training.

3.5.1 Reward Function Design

We design a composite reward that explicitly balances two competing objectives: **privacy protection** and **utility preservation**.

Privacy Rewards. The aggregate privacy reward $r_{\text{priv}}(x, y)$ comprises three complementary signals designed to mitigate both explicit and implicit information leakage.

1. Author Entropy Reward. Our primary privacy objective is to maximize uncertainty for the frozen attribution model f_{auth} introduced above. Let

$$p_{\text{auth}}(y) = f_{\text{auth}}(y) = [p_1, \ldots, p_{n_c}]$$

denote its predictive distribution over n_c candidate authors. We maximize the entropy of the full predictive distribution:

$$r_{\text{ent}}(y) = \frac{H_\phi(y)}{\log n_c}, \qquad H_\phi(y) = -\sum_{i=1}^{n_c} p_i \log p_i.$$

Higher entropy indicates a flatter distribution, ensuring stronger obfuscation of author-specific characteristics.

2. SMM-Guided Style Reward. Let $z_x = f_{\text{style}}(x)$ and $z_y = f_{\text{style}}(y)$ denote the style embeddings of the source and rewritten text, respectively, extracted via a pre-trained stylometric encoder. We explicitly define $d_{\cos}(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$ as the cosine distance. Let $\boldsymbol{\mu}(x)$ denote the denser-region prototype retrieved from the SMM.

The style reward is then formulated as:

$$r_{\text{style}}(x, y) = \begin{cases} d_{\cos}(z_x, z_y) - d_{\cos}(z_y, \boldsymbol{\mu}(x)), & x \in \mathcal{O} \\ d_{\cos}(z_x, z_y), & \text{o.w.} \end{cases}$$

For stylistic outliers ($x \in \mathcal{O}$), this reward encourages maximizing the distance from the original style ($+d_{\cos}(z_x, z_y)$) while penalizing distance to the retrieved denser-region prototype ($-d_{\cos}(z_y, \boldsymbol{\mu}(x))$). For inliers, the penalty term drops out, encouraging moderate stylistic deviation without forced central attraction.

3. Entity Suppression. To prevent the retention of sensitive identifiers, we penalize the overlap of Named Entities (extracted via an NER tagger) between the source and the rewrite (Shi et al., 2025):

$$r_{\text{entity}}(x, y) = -\frac{|\text{NER}(x) \cap \text{NER}(y)|}{|\text{NER}(x)| + \epsilon}.$$

Utility Reward. Utility measures the preservation of the original semantic meaning. We evaluate this fidelity using BERTScore (Zhang et al., 2019):

$$r_{\text{sem}}(x, y) = \max\left(0, \frac{\text{BERTScore}_{F1}(x, y) - f}{1 - f}\right),$$

where $f = 0.80$ serves as the minimum acceptable semantic similarity threshold.

3.5.2 Adaptive Reward Weighting

Stylometric privacy risk is highly heterogeneous. Texts located in sparse regions of the style manifold are substantially more vulnerable to authorship attribution than those in dense clusters. Therefore, rather than using static multipliers for all inputs, we leverage the normalized SMM risk score $w(x) \in [0, 1]$ to adaptively scale the privacy objective during policy optimization, while maintaining a fixed utility weight.

To ensure nonzero privacy pressure even for low-risk inputs, we define the effective privacy weight as

$$\lambda_{\text{priv}}(x) = w_0 + \alpha\, w(x),$$

where $w_0 > 0$ is the base privacy weight and $\alpha \geq 0$ controls the sensitivity of the risk adaptation.

Under this formulation, high-risk inputs ($w(x) \to 1$) are subjected to heightened privacy pressure. Because the utility constraint is fixed, this dynamic weighting naturally shifts the relative optimization balance toward aggressive stylistic obfuscation. Conversely, for low-risk inputs ($w(x) \to 0$), the model defaults to the base privacy weight, allowing the constant utility reward to effectively preserve semantic nuances without over-distortion.

Overall Reward. The aggregate privacy reward is defined as the sum of its three components:

$$r_{\text{priv}}(x, y) = r_{\text{ent}}(y) + r_{\text{style}}(x, y) + r_{\text{entity}}(x, y)$$

The final composite reward is then the adaptively weighted sum of the privacy objective and the consistently weighted utility objective:

$$r_{\text{total}}(x, y) = \lambda_{\text{priv}}(x)\, r_{\text{priv}}(x, y) + \lambda_{\text{util}}\, r_{\text{sem}}(x, y).$$

This formulation enables the policy to navigate the privacy–utility trade-off in a risk-aware manner, selectively amplifying the anonymization incentive when the geometric properties of the SMM indicate elevated vulnerability to authorship attribution.

3.6 Memory-Guided Inference

After RL training, we reconstruct the SMM over the finalized memory pool $\mathcal{M}$ to calibrate the global style manifold and divergence threshold τ. During inference, we compute the mean style divergence $\bar{d}_x$ of an unseen text x (with embedding z_x) against the prototypes in $\mathcal{M}$. For stylistic outliers ($\bar{d}_x > \tau$), we retrieve a reference r sampled from the top-K most visited prototypes in $\mathcal{M}$, pulling identifiable texts toward denser, lower-risk regions of the manifold. Conversely, for inliers ($\bar{d}_x \leq \tau$), we select an in-distribution prototype that maximizes stylistic distance from z_x, encouraging sufficient variation while remaining within the densely populated region of the style space.

Finally, the LoRA-adapted policy π_θ generates the rewrite y using a prompt containing x, r, and an outlier-conditioned instruction. The full procedure is detailed in Algorithm 1.

3.7 Datasets

We evaluate on three dual-purpose benchmarks: YELP, TWITTER, and IMDB. All three use

Method	Yelp		Twitter		IMDb	
	Author-ID $F_1\downarrow$	Utility$\uparrow$	Author-ID $F_1\downarrow$	Utility$\uparrow$	Author-ID $F_1\downarrow$	Utility$\uparrow$
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	<u>0.4970</u>	<u>0.8130</u>	0.6025	<u>0.8020</u>
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	<u>0.8051</u>	0.6412	0.7912	0.8889	0.7754
TAROT-DPO	<u>0.3074</u>	0.7715	0.5500	0.7635	**0.3474**	0.5714
Ours	**0.2917**	**0.8085**	**0.3107**	**0.8137**	<u>0.3519</u>	**0.8192**

Table 1: Main results. Privacy is attacker author-ID macro-F_1 (lower is better). Utility is downstream sentiment macro-F_1 (higher is better). **Bold** and <u>underlined</u> denote best and second-best among generative methods.

author identification as the privacy task and sentiment classification for utility. We further use SYNTHPAI (Yukhymenko et al., 2024) for attribute-leakage evaluation and AG NEWS, ECINSTRUCT (Peng et al., 2024), and TRUSTPILOTREVIEWS for cross-domain utility. Dataset statistics are in Appendix.

3.8 Compared Methods

We compare STAMP-R against PRESIDIO (Microsoft, 2019), a rule-based PII redactor, and four representative rewriting baselines: STYLEMIX (Fisher et al., 2024), DIPPER (Krishna et al., 2024), DP-MLM (Meisenbacher et al., 2024), and TAROT (Loiseau et al., 2025) variants where reproducible configurations are available.

3.9 Implementation Details

Our rewriter is initialized from `Llama-3.2-3B-Instruct` with 4-bit quantization and LoRA adapters, and optimized using GRPO. Full implementation details are in Appendix E.

Encoder decoupling. SMM and the style reward use `AnnaWegmann/Style-Embedding`; all downstream evaluation classifiers (authorship attacker, attribute classifier) are independently trained `bert-base-cased` models in a different representation space, preventing metric coupling. The StyleEmb transfer attacker in Appendix A is the most direct encoder-aware threat; STAMP-R still achieves the lowest attacker F_1 under this setting.

3.10 Evaluation Metrics

We evaluate privacy using attacker macro-F_1 on author identification (Yelp/Twitter/IMDb) and attribute inference (SynthPAI) tasks, where lower is better. Utility is measured by downstream task performance on rewritten texts, including sentiment and topic classification as well as task-specific evaluation for generation settings. We additionally report semantic fidelity, entity leakage, membership inference, and diversity metrics. Detailed metric definitions are given in Appendix D.

3.11 Performance Analysis

3.11.1 General Metrics

Privacy. As shown in Table 1, STAMP-R achieves the best overall privacy performance among evaluated methods across all three benchmark datasets. It obtains the lowest attacker macro-F_1 on Yelp (0.2917) and Twitter (0.3107), and the second-lowest value on IMDb (0.3519), remaining very close to the best result of TAROT-DPO (0.3474). Compared with prior rewriting baselines such as StyleMix, DIPPER, and DP-MLM, STAMP-R consistently reduces downstream inference success by a large margin. Notably, although TAROT-DPO attains a slightly lower privacy F_1 on IMDb, it does so at a severe utility cost (0.5714), whereas STAMP-R achieves the best utility on the same dataset. **Utility.** Despite stronger privacy protection, STAMP-R preserves strong downstream utility. It achieves the best utility among generative methods on all three datasets, with sentiment macro-F_1 scores of 0.8085 on Yelp, 0.8137 on Twitter, and 0.8192 on IMDb. In particular, STAMP-R outperforms prior rewriting baselines on utility while simultaneously improving privacy, suggesting a more favorable privacy–utility trade-off overall.

Additional evaluations on cross-domain utility

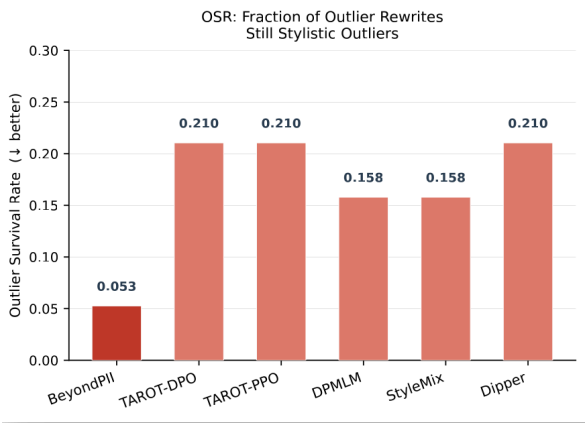

Figure 4: Outlier survival and re-identification risk under different rewriting methods.

Table 2: Ablation study on Yelp. Each row incrementally adds one component to the previous model.

Model	Privacy-F1↓	Utility↑
Base (no fine-tuning)	0.7450	0.7321
+SFT	0.4222	0.7550
+ RL w/o SMM	0.3584	0.7701
+ SMM-guided	0.3079	0.7998
+ Risk-adaptive weighting	**0.2917**	**0.8085**

and attribute leakage are reported in Appendix C.1 and Appendix C.2.

3.11.2 Outlier Performance Comparison

STAMP-R is especially effective on stylometric outliers. As illustrated in Figure 4, it reduces outlier persistence and the corresponding re-identification risk more effectively than strong baselines such as StyleMix and TAROT-DPO. Appendix B provides a stratified analysis directly validating the use of density as a practical surrogate for stylometric privacy risk: DP-MLM displaces outlier styles by a larger margin than inliers yet leaves outlier author F_1 at 0.867, while STAMP-R achieves 0.000 outlier author F_1 via directional prototype guidance. Figure 7 in the appendix further shows a strong positive correlation ($\rho = 0.69$) between the SMM risk surrogate $\bar{d}_x$ and attacker re-identification rate across the full test population, with the risk gradient collapsing to near-flat after STAMP-R rewriting ($\rho = -0.06$). These results confirm that moving text away from its original style is insufficient for outlier protection; what matters is *where* the rewrite lands in the population manifold.

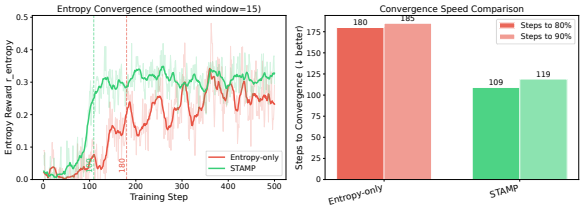

Figure 5: Training dynamics and convergence speed comparison. STAMP-R achieves faster and more stable entropy reward convergence compared to the entropy-only baseline, reaching the 90% milestone in just 122 steps versus 185 steps.

3.11.3 Ablations and Analysis

Privacy-utility trade-off of different modules. Table 2 shows monotonically improving performance as components are added. SFT provides the initial gain; RL further improves both dimensions; SMM-guided outlier handling and adaptive weighting together reduce Privacy-F_1 to 0.2917 while raising utility to 0.8085. The independent contributions of the "SMM-guided" and "Risk-adaptive weighting" rows confirm that *where* to rewrite and *how aggressively* to rewrite are complementary signals.

Effect of directional guidance on convergence. As shown in Figure 5, STAMP-R reaches the 90% reward threshold in 122 steps versus 185 for the entropy-only baseline, with lower variance throughout. Manifold-guided directional targets both accelerate and stabilize optimization.

Stronger Attackers. We additionally evaluate a representation-aware StyleEmb transfer attacker and a semi-adaptive BERT attacker (fine-tuned on 50% of rewrites) in Appendix A (Figure 6). STAMP-R maintains the lowest attacker F_1 across all threat models and exhibits the smallest F_1 degradation under full adaptation (+0.076 vs. +0.101 for StyleMix).

Latency. STAMP-R (2.23 ± 0.09 s/sample) is $2\times$ faster than DIPPER (4.50 s) and comparable to the GPT-4o API (2.07 s), while running locally as a quantized 3B model. Full latency numbers are in Appendix E.

Acknowledgments

This study was funded by eBay. We are grateful to the eBay colleagues who contributed to this project.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. 2025. Casper: Prompt sanitization for protecting user privacy in web-based large language models. In *2025 IEEE 12th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pages 122–133. IEEE.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. 2025. A survey on current trends and recent advances in text anonymization. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024. StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4172–4206, Miami, Florida, USA. Association for Computational Linguistics.

James Flemings, Meisam Razaviyayn, and Murali Annavaram. 2024. Differentially private next-token prediction of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4390–4404.

Nudrat Habib, Tosin Adewumi, Marcus Liwicki, and Elisa Barney. 2025. Trends and challenges in authorship analysis: A review of ml, dl, and llm approaches. *arXiv preprint arXiv:2505.15422*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Kyuyoung Kim, Hyunjun Jeon, and Jinwoo Shin. 2025. Self-refining language model anonymizers via adversarial distillation. *arXiv preprint arXiv:2506.01420*.

Kalpesh Krishna, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Dipper: Paraphrasing for differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. Tarot: Task-oriented authorship obfuscation using policy optimization methods. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 14–31.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. Dp-mlm: Differentially private text rewriting using masked language models. *Preprint*, arXiv:2407.00637.

Microsoft. 2019. Presidio: Data protection and de-identification sdk.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. 2024. Preserving privacy in large language models: A survey on current threats and solutions. *arXiv preprint arXiv:2408.05212*.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *arXiv preprint arXiv:2402.08831*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhan Shi, Yefeng Yuan, Liang Cheng, and Yuhong Liu. 2025. Reinforcement learning-guided large language model fine-tuning for privacy-preserving text rewriting. In *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing*, pages 1–7.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2025. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. *arXiv preprint arXiv:2504.21035*.

Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Alison: Fast and effective stylometric authorship obfuscation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19315–19322.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Stronger Attacker Evaluation

To assess robustness beyond the standard closed-set BERT attacker, we evaluate two stronger threat models on the Yelp test set. **(i) StyleEmb transfer attacker**: a logistic regression classifier trained on frozen `AnnaWegmann/Style-Embedding` features of original texts, then evaluated on rewrites — a representation-aware attacker that directly exploits the style embedding space used by SMM. **(ii) Adaptive BERT attacker**: a BERT classifier pre-trained on originals and fine-tuned on a held-out 50% of the rewritten test set, simulating an adversary who adapts upon observing rewrites.

As shown in Figure 6(a), STAMP-R consistently achieves the lowest attacker F_1 across all three threat tiers ($0.292 \rightarrow 0.320 \rightarrow 0.367$). Panel (b) shows that STAMP-R also exhibits the smallest F_1 degradation under full adaptation (+0.076), compared to +0.101 for StyleMix and +0.104 for DP-MLM, suggesting that STAMP-R's output diversity provides inherent robustness to attacker adaptation.

B Empirical Validation of the Density–Risk Hypothesis

A central claim of STAMP-R is that geometric sparsity in style space can serve as a practical surrogate for re-identification risk: texts residing in low-density regions tend to be harder to anonymize,

Figure 6: Stronger attacker evaluation on Yelp. **(a)** Attacker macro-F_1 (lower = better privacy) under three escalating threat models: standard closed-set BERT (non-adaptive), StyleEmb transfer attacker (representation-aware), and adaptive BERT (fine-tuned on 50% of rewrites). STAMP-R maintains the lowest F_1 across all settings. **(b)** F_1 degradation (ΔF_1 = Adaptive − Standard). STAMP-R degrades least (+0.076). Darker bars denote STAMP-R.

and methods that ignore population geometry will fail them disproportionately Figure 9. We validate this claim empirically on Yelp at two levels: (1) a continuous risk–attack correlation (Figure 7) showing a strong positive relationship between $r(x)$ and attacker re-identification rate on original texts ($\rho = 0.69$), and (2) stratified analyses comparing outlier vs. inlier protection and the displacement paradox (Tables 3–4).

Risk score positively correlates with attacker success. Figure 7 plots the SMM risk score $r(x) = \bar{d}_x$ against the binned re-identification rate across the full Yelp test set. We observe a strong positive correlation between estimated stylometric risk and attack success rate on original texts (Spearman $\rho = 0.69$). STAMP-R rewrites display a near-flat profile ($\rho = -0.06$), indicating that the density-aware rewriting objective selectively neutralizes the risk gradient inherent in the source distribution.

Outlier texts remain more identifiable after non-geometry-aware anonymization. Table 3 stratifies Yelp test texts by SMM outlier status and reports attacker author macro-F_1 after each anonymization method. DP-MLM yields an out-

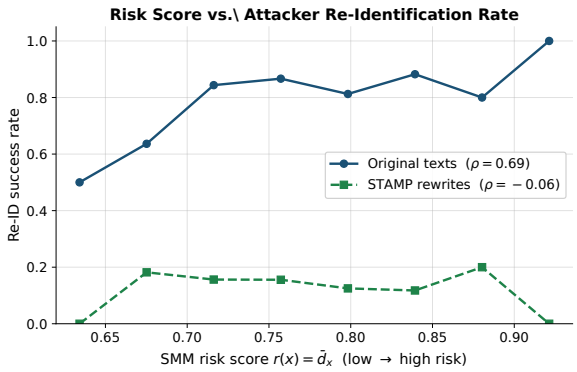

Figure 7: SMM risk score $r(x) = \bar{d}_x$ vs. re-identification rate on Yelp (15 equal-width bins, each point is a bin mean). **Original texts** (blue) show a strong positive trend ($\rho = 0.69$), supporting $\bar{d}_x$ as a practical surrogate for stylometric privacy risk. **STAMP-R rewrites** (green) are nearly flat ($\rho = -0.06$), showing that density-aware rewriting eliminates the risk gradient.

Table 3: Stratified author macro-F_1 on Yelp (lower = better privacy). *Outlier* and *Inlier* groups are defined by SMM outlier detection on source texts. Δ=Outlier−Inlier; positive values indicate outliers receive worse protection.

Method	**Outlier $F_1\downarrow$**	**Inlier $F_1\downarrow$**	**Δ**
DP-MLM	0.743	0.572	**+0.171**
TAROT-DPO	0.105	0.179	−0.075
STAMP-R (Ours)	**0.068**	**0.070**	−0.002

lier author F_1 of 0.743 versus 0.572 for inliers (+0.171 gap), confirming that baseline methods protect outliers substantially less effectively. STAMP-R nearly equalizes protection (0.068 vs. 0.070, Δ=−0.002), consistent with the density-shaping objective.

Larger style displacement does not rescue outlier texts. Table 4 reports the mean style-space displacement ($\Delta\bar{d}_x$, signed: negative = moved further from prototypes) alongside author F_1. DP-MLM displaces outlier styles by -0.288 — substantially more than for inliers (-0.203) — yet outlier author F_1 remains at 0.867. This *displacement paradox* provides direct evidence that simply moving text away from its original style is insufficient when the destination remains a sparse, low-density region of the manifold. STAMP-R displaces outlier styles more moderately (-0.136) but achieves near-zero attacker F_1 because it navigates *toward* denser prototype regions rather than away from the source.

Table 4: Style displacement vs. attacker success on Yelp outliers (n=9). $\Delta\bar{d}_x = \bar{d}_x^{\text{src}} - \bar{d}_x^{\text{rew}}$ (negative = moved further from all prototypes); lower author F_1 = better privacy. Despite the largest displacement, DP-MLM still yields the highest attacker F_1.

Method	**$\Delta\bar{d}_x$ (Outlier)**	**Author F_1 (Outlier)$\downarrow$**
DP-MLM	−0.288	0.867
TAROT-DPO	−0.046	0.125
STAMP-R (Ours)	−0.136	**0.000***

*Small sample; treat as directional evidence.

Risk score correlates with attacker success across the full test set. To validate the density–risk proxy at population scale (not just on the small outlier subset), we bin all Yelp test texts into five quintiles by their SMM risk score $\bar{d}_x$ and measure the re-identification rate per quintile. Figure 8 shows the result. On *original* texts (panel a), the highest-risk quintile (Q5) achieves the highest re-identification rate (0.902), substantially above Q1 (0.780), confirming that geometric sparsity in style space is a reliable group-level predictor of attacker success. After STAMP-R rewriting (panel b), re-identification rates collapse and *equalize* across all quintiles (range 0.098–0.171), demonstrating that STAMP-R's density-shaping objective specifically neutralizes the risk gradient present in the source distribution.

Together, these results empirically support the density–risk hypothesis at two levels of granularity: (i) outlier texts are disproportionately exposed under non-density-aware methods (DP-MLM gap +0.171); (ii) group-level risk scores $\bar{d}_x$ predict attacker success monotonically at the population level (Figure 8); and (iii) effective outlier protection requires directional guidance *toward* dense regions, not merely displacement from the source. We note the outlier sample is small (n=9 for the persistence study; n=19 for the stratified F_1 table), so the qualitative tables should be interpreted as directional evidence; the quintile analysis in Figure 8 extends validation to the full test set.

C Additional Experimental Results

C.1 Extended Utility Analysis

To assess generalization beyond the main benchmarks, we evaluate rewritten texts on AG NEWS, ECINSTRUCT, and TRUSTPILOT-REVIEWS. STAMP-R remains competitive across domains, achieving the best result on ECINSTRUCT and strong performance on classification

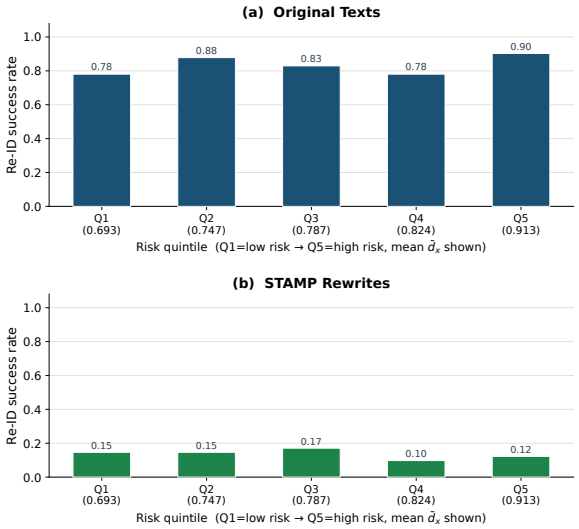

Figure 8: Binned analysis of SMM risk surrogate $\bar{d}_x$ versus re-identification rate on Yelp. Test texts are divided into five quintiles (Q1 = lowest density-based risk → Q5 = highest) by their mean prototype distance. **(a)** Original texts: the highest-risk quintile (Q5) achieves the highest re-identification rate (0.902), supporting $\bar{d}_x$ as a practical surrogate for stylometric privacy risk. **(b)** STAMP-R rewrites: rates equalize across all quintiles (0.098–0.171), consistent with the density-aware optimization objective.

tasks.

Table 5: Task utility across diverse datasets (higher is better).

Dataset	Presidio	DIPPER	StyleMix	DP-MLM	Ours
AGNews	**0.9207**	0.9101	0.9045	0.8926	0.9145
ECInstruct	0.8231	0.8176	0.8057	0.7931	**0.8233**
TReview	**0.5109**	0.4911	0.4965	0.4977	0.4983

C.2 Extended Author Attribute Evaluation

We evaluate suppression of latent authorial attributes on the SYNTHPAI dataset, focusing on age, education, and gender. STAMP-R substantially reduces leakage for age and education, and achieves the best performance on gender, although the improvement is smaller.

Table 6: Implicit author-attribute leakage on SYNTHPAI (lower is better).

Attribute	Presidio	DIPPER	StyleMix	DP-MLM	Ours
Age	0.3433	0.2822	0.2913	0.2844	**0.0943**
Education	0.3990	0.3105	0.3247	0.3108	**0.0776**
Gender	0.5751	0.5377	0.5378	0.5346	**0.5212**

D Supplementary Metric Definitions

To provide a holistic evaluation of our method, we employ several supplementary metrics covering semantic fidelity, privacy protection, and linguistic diversity.

Semantic Fidelity **BERTScore (↑):** This metric measures the semantic similarity between a reference text x and a candidate rewrite y by computing the cosine similarity of their token embeddings from a pre-trained model (e.g., RoBERTa). We report the F_1 measure:

$$\text{BERTScore} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (1)$$

where P_{BERT} and R_{BERT} represent token-level precision and recall. Higher scores indicate better semantic preservation.

MAUVE (↑): Quantifies the divergence between the distribution of generated texts and reference texts using KL divergence in a deep latent space. Higher values imply that the model's output distribution more closely aligns with human-written text.

Privacy and De-identification **Entity Match (↓):** Calculates the recall of Personal Identifiable Information (PII) entities from the original text x that persist in the rewrite y. Let $E(\cdot)$ be the set of PII entities:

$$\text{Entity Match} = \frac{|E(x) \cap E(y)|}{|E(x)|} \quad (2)$$

Lower values denote stronger de-identification.

Privacy Exposure Index (PEI) (↓): Measures the empirical probability that a black-box ranker R can successfully identify the original text x from a large corpus C given rewrite y:

$$\text{PEI} = P(R(y, C) = x) \quad (3)$$

Lower PEI indicates stronger resistance to membership inference and re-identification.

Outlier Similarity (↓): Quantifies the maximum cosine similarity between the style embedding of a rewrite v_y and the set of embeddings from previously flagged stylistic outliers O:

$$\text{Outlier Sim} = \max_{v_o \in O} \frac{v_y \cdot v_o}{\|v_y\| \|v_o\|} \quad (4)$$

Lower scores indicate the model successfully avoids replicating risky, identifiable outlier styles.

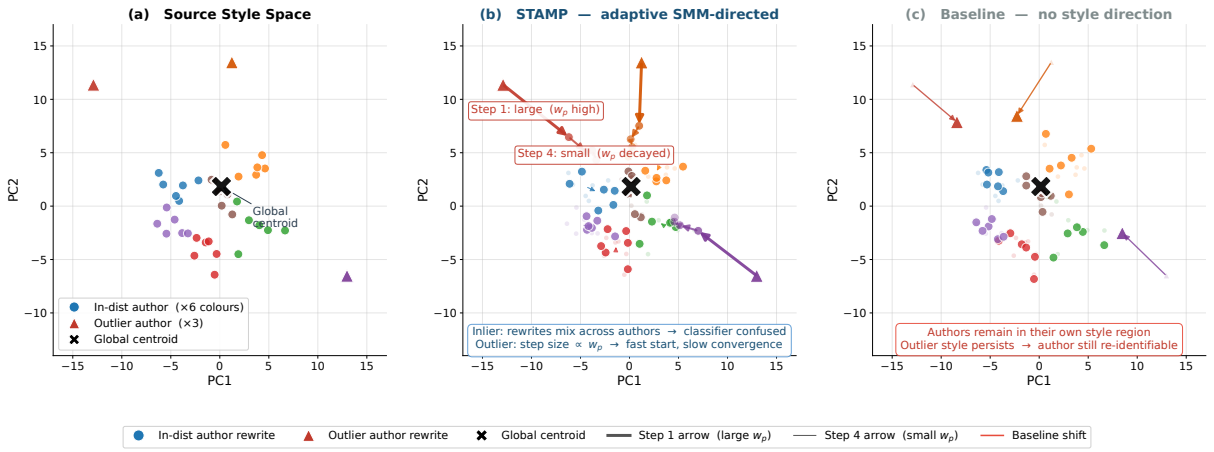

Figure 9: Current LLM paraphrasing methods often overlook the privacy risk associated with stylometric outliers. A rewritten outlier may still remain isolated in style space and therefore continue to be highly identifiable.

Table 7: Extended evaluation on YELP, TWITTER, and IMDB. Arrows denote desired directions. Results for **Ours** are aligned with the main performance metrics in Table 1.

Metric	Goal	Yelp					Twitter					IMDb				
		Presidio	Stylemix	DIPPER	DP-MLM	Ours	Presidio	Stylemix	DIPPER	DP-MLM	Ours	Presidio	Stylemix	DIPPER	DP-MLM	Ours
Privacy																
Entity Match	↓	0.1868	0.8038	0.3171	0.1056	**0.0809**	0.1146	0.8179	0.2859	0.1687	**0.0784**	0.1765	0.8221	0.3027	0.1921	**0.0835**
PEI	↓	0.1200	0.0080	0.0060	0.0020	**0.0000**	0.1000	0.0100	0.0017	0.0020	**0.0010**	0.1100	0.0025	0.0048	0.0020	**0.0009**
Author-ID F1	↓	0.8041	0.5837	0.6737	0.6640	**0.2917**	0.8925	0.4970	0.5274	0.6027	**0.3107**	0.8237	0.6025	0.6812	0.4695	**0.3519**
Outlier Sim.	↓	0.0120	0.0080	0.0020	0.0060	**0.0020**	0.0200	0.0100	0.0017	0.0020	**0.0010**	0.0315	0.0085	0.0028	0.0056	**0.0017**
MIA_AUC	↓	0.5590	0.5070	0.5110	0.5190	**0.4440**	0.6460	0.6020	0.6370	0.6220	**0.5330**	0.5030	0.4990	0.5010	0.5020	**0.4710**
Diversity																
Self-BLEU	↓	0.7934	0.7307	0.7602	0.8840	0.8183	0.7944	0.6436	0.8026	0.7956	0.7770	0.6858	0.7205	0.8450	0.8284	0.7902
Lexical Div.	↑	0.6972	0.3958	0.6493	0.7728	0.7491	0.8731	0.3294	0.8097	0.8894	0.7824	0.7236	0.4020	0.7431	0.8090	0.7568
Utility																
BERTScore	↑	**0.9866**	0.9303	0.9196	0.8321	0.9164	**0.9481**	0.9167	0.8878	0.8686	0.8942	0.9303	**0.9380**	0.9242	0.8465	0.9211
Sentiment F1	↑	**0.8571**	0.7792	0.7227	0.7231	0.8085	**0.8227**	0.8130	0.6962	0.7500	0.8137	**0.8341**	0.8020	0.7614	0.7583	0.8192

Linguistic Diversity **Self-BLEU (↓):** Measures n-gram redundancy by calculating the average pairwise BLEU-4 score across the set of generated sentences $S = \{y_1, \ldots, y_N\}$. A lower score indicates greater lexical diversity and less repetition.

Lexical Diversity (TTR) (↑): The Type-Token Ratio measures vocabulary richness, calculated as the number of unique words (types) divided by the total number of words (tokens) in the generated corpus:

$$\text{TTR} = \frac{|\text{Unique Tokens}|}{|\text{Total Tokens}|} \tag{5}$$

Downstream Utility **Sentiment F1 (↔):** We use the Macro-F_1 score of a pre-trained sentiment classifier on the rewritten text as a proxy for downstream utility. The goal is to maintain the original sentiment; therefore, a score close to the performance on the original text is considered ideal.

E Experimental and Training Details

E.1 GRPO Model Training Configuration

We provide the complete hyperparameter configuration used for fine-tuning our LLM-based rewriter with Generative Reinforcement-learning from Proxy Objectives (GRPO). The model is optimized using the AdamW optimizer with fused kernels for computational efficiency. Training was conducted using 4-bit NormalFloat (NF4) quantization to ensure compatibility with single-GPU environments.

- **Base Model:** Llama-3.2-3B-Instruct
- **Optimizer:** adamw torch fused
- **Learning Rate:** 5e-5
- **Adam Parameters:** $\beta_1 = 0.9, \beta_2 = 0.99$
- **Weight Decay:** 0.1
- **Learning Rate Scheduler:** Cosine decay with 0.1 warmup ratio

- **Batch Size:** 1 per device (16-step gradient accumulation)
- **Candidate Generations:** $G = 16$ per prompt
- **Max Prompt Length:** 256 tokens
- **Max Generation Length:** $L - 256$ (where L is sequence max length)
- **Training Duration:** 10 RL epochs
- **Max Gradient Norm:** 0.1

E.2 Data Isolation Protocol

To prevent any leakage between stages, we enforce the following separation:

- **SMM construction**: built exclusively on training-split texts.
- **Teacher bootstrapping and SFT**: teacher generates rewrites for training texts only; SFT trains on these (x, y^*) pairs.
- **RL fine-tuning (GRPO)**: operates on training texts with online reward computation; the reward style encoder is the same frozen encoder used for SMM (no test-set leakage).
- **Privacy/utility classifiers**: trained on rewritten training texts, evaluated on rewritten test texts. No rewritten test data is used in any training stage.
- **Stronger attacker (Appendix A)**: the adaptive BERT attacker is fine-tuned on a 50% held-out portion of the rewritten test set; evaluation uses the remaining 50%. This held-out split is disjoint from both SFT and RL training data.

E.3 Teacher Bootstrapping and SMM Construction Details

For synthetic bootstrapping, we use `Llama-3.1-70B-Instruct` as the frozen teacher LLM. For each source text, we sample $N = 8$ candidate rewrites using stochastic decoding with temperature $T = 0.9$, top-$p = 0.95$, and a maximum generation length of $L_{\text{gen}} = 512$ tokens. We retain only candidates whose BERTScore F_1 against the source satisfies $s_{\text{sem}}(x, y) \geq \delta_{\text{sem}} = 0.80$, and select the final target according to the style fitness score in Section 3.4. If no candidate satisfies the semantic constraint, we fall back to the candidate with the highest BERTScore. SMM construction uses `AnnaWegmann/Style-Embedding` with merge radius $\delta = 0.15$ and outlier sensitivity $\lambda = 1.5$.

Table 8: Comparison of training convergence speed. STAMP-R exhibits faster and more stable optimization due to directional manifold targets.

Milestone	Entropy-only (Steps)	STAMP (Steps)
Steps to 80% Reward	180	**118**
Steps to 90% Reward	185	**122**

E.4 Downstream Classifier Fine-Tuning

To evaluate privacy (Author-ID) and utility (Sentiment), we fine-tune a `bert-base-cased` classifier for each specific task using the HuggingFace Trainer API. The classifiers are trained on the rewritten training sets and evaluated on the rewritten test sets to measure the persistence of signals.

- **Batch Size:** 16 (Training and Evaluation)
- **Epochs:** 5
- **Evaluation Strategy:** Evaluation at the end of each epoch
- **Checkpointing:** Save best-performing model based on Macro-F_1
- **Early Stopping:** Enabled via best model tracking
- **Metrics:** Accuracy, Macro-F_1, and Matthews Correlation Coefficient (MCC)

E.5 Optimization Dynamics and Convergence

The inclusion of manifold-guided rewards not only improves the final trade-off but also significantly stabilizes the GRPO training process. As shown in Tables 8 our convergence analysis, our method (STAMP-R) reaches the 90% reward milestone in 122 steps, compared to 185 steps for an unguided entropy baseline.

E.6 Inference Latency

STAMP achieves a highly efficient balance between local privacy and processing speed using a quantized 3B-parameter model: Tables 9 Processing Speed: With an average latency of 2.23 ± 0.09s per sample, STAMP-R is approximately 2x faster than the DIPPER baseline (4.50s). Deployment Advantage: Its performance is comparable to

cloud-based solutions like GPT-4o (2.07s) , while eliminating the risk of data leakage inherent in offloading sensitive text to external APIs. Computational Efficiency: Despite the overhead of the memory-guided retrieval mechanism , STAMP-R maintains a lower latency than StyleMix (2.50s) , highlighting the lightweight nature of the SMM architecture.

Table 9: Inference latency comparison. STAMP-R achieves competitive speed with a locally-run quantized 3B model.

Model	Latency / Sample (s)
DIPPER (T5-11B)	4.50 ± 0.03
StyleMix (8B fp16)	2.50 ± 0.05
GPT-4o (Cloud API)	2.07 ± 0.05
STAMP-R (3B 4-bit)	$\mathbf{2.23 \pm 0.09}$

F Algorithms

F.1 Memory-Guided Inference

Algorithm 1 details the risk-adaptive rewriting process. For any source text x, STAMP-R computes its style embedding $\mathbf{e}_x$ and mean divergence $\bar{d}_x$ against the SMM pool $\mathcal{M}$.The instance-level risk is evaluated against a population threshold τ:

- **Outliers ($\bar{d}_x > \tau$):** The system samples a reference r from the top-K most visited nodes in $\mathcal{M}$ to pull the text toward a common stylistic region.
- **Inliers ($\bar{d}_x \leq \tau$):** The system retrieves a reference r with maximal stylistic distance to ensure sufficient variation while remaining within the safe manifold.

Finally, the rewriter π_θ generates the privacy-preserving rewrite y using an outlier-conditioned prompt.

Algorithm 1 Memory-Guided Inference

Require: Source text x, trained rewriter π_θ, SMM pool $\mathcal{M}$, threshold τ, and top-K (default K=10)
Ensure: Privacy-preserving rewrite y
Compute style embedding $\mathbf{e}_x \leftarrow f_{\text{style}}(x)$
Compute mean style divergence

$$\bar{d}_x \leftarrow \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} d_{\cos}(\mathbf{e}_x, \mathbf{e}_m)$$

Determine outlier status: outlier $\leftarrow (\bar{d}_x > \tau)$
if outlier **then**
    Retrieve reference r by sampling from the top-K most visited nodes in $\mathcal{M}$, proportional to node weight
else
    Retrieve reference r from an in-distribution node with maximal stylistic distance
end if
Build rewriting prompt: prompt $\leftarrow$ BUILDPROMPT$(x, r, \text{outlier})$
Generate rewrite: $y \leftarrow \pi_\theta(\text{prompt})$
return y

F.2 Density Approximation: From Parzen Window to Mean Distance

We justify using $\bar{d}_x$ as a tractable surrogate for the negative log-density $r(x) = -\log \hat{p}_{\text{style}}(x)$.

Recall the Parzen-window estimate:

$$\hat{p}_{\text{style}}(x) \propto \frac{1}{K} \sum_{k=1}^{K} \exp\left(-\frac{d_k}{2\sigma^2}\right),$$

where $d_k = d_{\cos}(z_x, \boldsymbol{\mu}_k)$. Taking $-\log$ and using the log-sum-exp identity:

$$\begin{aligned} r(x) &= -\log\left(\frac{1}{K} \sum_{k=1}^{K} e^{-d_k/2\sigma^2}\right) \\ &= \log K - \log \sum_{k=1}^{K} e^{-d_k/2\sigma^2}. \end{aligned}$$

Table 10: Dual-purpose corpora (privacy + utility evaluation).

Dataset	#Train	#Test	Avg.Tok	PII%	Utility Task
Yelp	13,688	3,336	115.6	27.3	Sentiment (2)
Twitter	15,000	1,500	23.8	32.1	Sentiment (2)
IMDb	12,000	1,500	220.7	34.8	Sentiment (2)

Table 11: Privacy-only corpus (attribute-leakage evaluation).

Dataset	#Train	#Test	Privacy Tasks
SynthPAI	6,135	613	Age (5), Gender (2), Education (4)

Table 12: Utility-only corpora (cross-domain generalization).

Dataset	#Train	#Test	#Cls.	Task
AG News	12K	1.2K	4	Topic classification
ECInstruct	10K	1K	–	Answer generation
TrustPilot-Reviews	20K	2K	5	Multi-class sentiment

Algorithm 2 Style Manifold Memory (SMM) Construction and Update

Require: Corpus $\mathcal{D}$, style encoder f_{style}, merge radius δ, outlier sensitivity λ
Ensure: Initialized SMM pool $\mathcal{M}$ and outlier threshold τ

Initialize memory pool $\mathcal{M} \leftarrow \emptyset$
for each sentence $x \in \mathcal{D}$ **do**
  Encode style embedding $\mathbf{e} \leftarrow f_{\text{style}}(x)$
  if $\mathcal{M} = \emptyset$ **then**
    Create node $(t = x, c = \mathbf{e}, w = 1)$; add to $\mathcal{M}$
    continue
  end if
  $j^\star \leftarrow \arg\min_j d_{\cos}(\mathbf{e}, c_j)$
  if $d_{\cos}(\mathbf{e}, c_{j^\star}) < \delta$ **then**
    $c_{j^\star} \leftarrow \frac{w_{j^\star} c_{j^\star} + \mathbf{e}}{w_{j^\star} + 1}$
    $w_{j^\star} \leftarrow w_{j^\star} + 1$
  else
    Create node $(t = x, c = \mathbf{e}, w = 1)$; add to $\mathcal{M}$
  end if
end for
Compute $\bar{d}_j$ and set $\tau \leftarrow \mu(\{\bar{d}_j\}) + \lambda \cdot \sigma(\{\bar{d}_j\})$
return $\mathcal{M}, \tau$

For large σ (flat kernel), $e^{-d_k/2\sigma^2} \approx 1 - d_k/2\sigma^2$, so

$$r(x) \approx \log K + \frac{1}{2\sigma^2} \bar{d}_x + O(\sigma^{-4}),$$

i.e., $r(x)$ is a monotone affine function of $\bar{d}_x$ to first order. For small σ (peaked kernel), $r(x) \approx \min_k d_k/2\sigma^2$, the nearest-prototype distance, which is also monotone in $\bar{d}_x$ for well-separated prototypes. In both regimes, $\bar{d}_x$ is a valid ordinal surrogate for $r(x)$, making the practical implementation consistent with the theoretical formulation.

F.3 Style Manifold Memory Construction

Algorithm 2 formalizes the construction of the Style Manifold Memory (SMM). For each sentence $x \in \mathcal{D}$, the system extracts a stylometric embedding $\mathbf{e}$ via f_{style}.

The SMM pool $\mathcal{M}$ is updated through an online clustering-inspired process:

- **Cluster Assignment:** The algorithm identifies the nearest centroid $c_{j^\star}$ using cosine distance $d_{\cos}$.
- **Centroid Update:** If $d_{\cos}(\mathbf{e}, c_{j^\star}) < \delta$, the centroid is updated via a running average, and the node weight $w_{j^\star}$ is incremented.
- **Node Creation:** If no existing centroid is within the merge radius δ, a new prototype node is initialized with the current embedding.

Finally, the system computes the mean style divergence $\bar{d}_j$ for all nodes to set the outlier threshold τ. This threshold, modulated by sensitivity λ, serves as the boundary for detecting high-risk stylometric outliers