# I Am No One: Style-Aware Paraphrasing for Text Anonymization

Anonymous ACL submission

## Abstract

Online content, despite being posted under pseudonyms, presents significant privacy risks as it often contains subtle stylistic cues that can be exploited to identify authors. Various studies have highlighted the importance of adding noise to textual data for anonymization, particularly through differential privacy; however, such methods often degrade the quality and utility of the original text. In this work, we propose an alternative approach to text anonymization that leverages the ability of pretrained large language models to capture and modify subtle stylistic attributes present in user generated text. Our method constructs an author's stylistic profile from minimal text samples and rewrites it using targeted paraphrasing to obscure identifiable style markers while preserving the original content. This strategic style manipulation allows us to significantly reduce the effectiveness of Authorship attribution attacks. On a real-world Google review dataset, our approach achieves a 50% reduction in authorship attribution success rates while maintaining content quality. We conduct extensive experiments across multiple datasets and rigorously evaluate our approach to assess its effectiveness in balancing the privacy-utility trade off.

## 1 Introduction

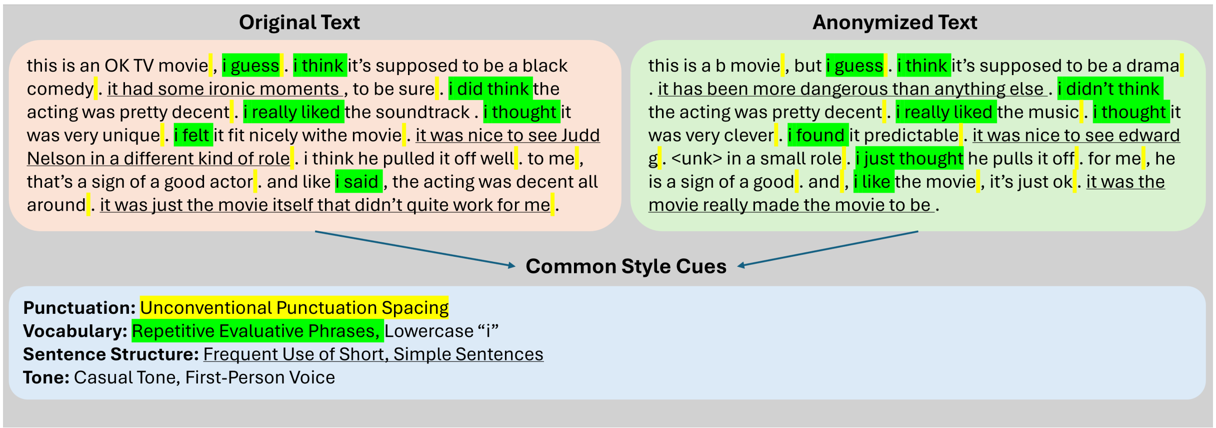
The widespread sharing of online content has raised significant concerns about user privacy. Subtle stylistic patterns embedded within user-generated data can be leveraged to identify and trace the original author, even in cases where users attempt to remain anonymous. This practice, known as "authorship attribution", poses a serious threat to individuals' privacy, as it can expose sensitive personal information and undermine their ability to control the dissemination of their digital footprint.

Traditional anonymization techniques, such as removing explicit personal identifiers, are proven to be inadequate in the face of modern machine learning models capable of extracting nuanced stylistic cues (Lison et al., 2021). This underscores a critical gap: current anonymization techniques largely overlook stylistic fingerprints as privacy risks.

Recently, differential privacy (DP) based methods such as DP-VAE (Weggenmann et al., 2022), DP-Prompt (Utpala et al., 2023) and DP-MLM (Meisenbacher et al., 2024) have attracted growing attention for textual anonymization, typically introducing calibrated noise or paraphrasing under privacy budgets. DP provides a mathematically grounded privacy guarantee, but the choice of privacy budget ($\varepsilon$) can drastically affect model utility: High-$\varepsilon$ settings may yield negligible noise and thus poor privacy, whereas low-$\varepsilon$ can degrade text utility to the point of unreadability.

Figure 1 shows that the anonymized text generated by DP-VAE retains identifiable stylistic patterns despite formal privacy guarantees. These limitations stem from the inherent design of DP-based anonymization methods, which typically introduce perturbations either to latent representations or through alterations of individual output tokens. Latent-level privacy mechanisms, often designed to bound representation sensitivity, frequently fail to disrupt higher-order syntactic structures or discourse patterns in the text. Conversely, while token-level mechanisms effectively obscure local patterns (e.g., word choice), their cumulative noise injection often degrade text utility to the point of unreadability. These findings challenge the assumption that DP's mathematical guarantees suffice for anonymity, as privacy-driven noise harms utility without fully erasing authorship cues.

Recent work has begun questioning whether rigorous DP is actually essential to guard against authorship attribution (Meisenbacher and Matthes, 2024). In practice, many real-world scenarios might benefit from less disruptive rewriting strategies, which are more focused on neutralizing author-specific style attributes rather than complete reliance on random perturbations. This could poten-



Original Text

this is an OK TV movie , i guess . i think it's supposed to be a black comedy . it had some ironic moments , to be sure . i did think the acting was pretty decent . i really liked the soundtrack . i thought it was very unique . i felt it fit nicely withe movie . it was nice to see Judd Nelson in a different kind of role . i think he pulled it off well . to me , that's a sign of a good actor . and like i said , the acting was decent all around . it was just the movie itself that didn't quite work for me .

Anonymized Text

this is a b movie , but i guess . i think it's supposed to be a drama . it has been more dangerous than anything else . i didn't think the acting was pretty decent . i really liked the music . i thought it was very clever . i found it predictable . it was nice to see edward g . <unk> in a small role . i just thought he pulls it off . for me , he is a sign of a good . and , i like the movie , it's just ok . it was the movie really made the movie to be .

Common Style Cues

**Punctuation:** Unconventional Punctuation Spacing
**Vocabulary:** Repetitive Evaluative Phrases, Lowercase "i"
**Sentence Structure:** Frequent Use of Short, Simple Sentences
**Tone:** Casual Tone, First-Person Voice

Figure 1: User generated text (*left*) and its anonymized version (*right*) by DP-VAE. Despite applying differential privacy, many stylistic cues remain, potentially enabling a classifier to re-identify the author.

tially help strike a better balance between privacy and content utility. Hence, we posit the following research question:

*How can we leverage stylistic transformations to anonymize user-generated text while preserving content utility, without resorting to noise injection?*

To address this question, we propose a style-aware, prompt-driven anonymization technique that uses an LLM to extract an author's key stylistic traits, and generate anonymized text. Our method first extracts core features such as text length and structure, punctuation patterns, vocabulary choices, and tone, which collectively form a concise style profile for each author. These dimensions are grounded in classical stylometry research (Holmes, 1994; Stamatatos, 2009; Usha and Thampi, 2017). This profile then guides the generation of anonymized text that preserves content utility while effectively masking author identity.

Our experiments on two real-world datasets spanning short-form and long-form text scenarios confirm that style-focused rewriting can effectively reduce authorship attribution accuracy while maintaining near-original meaning and, even better, readability. By avoiding explicit noise injection, our method focuses solely on removal of an author's stylistic markers, yielding coherent, anonymized text for a variety of lengths and genres.

Our work makes the following contributions to the study of text anonymization:

1. We introduce a prompt-driven, style-aware text anonymization framework that modifies author-specific stylistic features to protect identity while maintaining the utility. Our approach is model-agnostic, making it compatible with a wide range of language models and adaptable to diverse use cases.

2. We demonstrate that constructing a distinctive style profile for each author requires only a minimal number of samples. Our analysis reveals, for the first time, the relative importance of stylistic features like text length, tone, punctuation patterns, and vocabulary choices in the anonymization process, providing an explainable foundation for optimizing style-aware rewriting techniques.

3. We establish, through empirical evaluation on two user-generated datasets, that our approach reduces authorship-attribution accuracy by 50–70 % relative to strong baselines, while achieving the highest utility scores and preserving near-original readability.

## 2 Related Work

Text anonymization aims to obscure not only explicit identifiers but also subtle stylistic fingerprints such as syntax, vocabulary, and discourse patterns that can be leveraged for authorship attribution (Sundararajan and Woodard, 2018). Earlier studies on text anonymization frequently employed sequence labeling techniques (Lison et al., 2021), primarily focusing on removing explicit identifiers. However, these methods often do not adequately address higher-order linguistic structures contributing to an author's unique style. Recent research has increasingly explored approaches using either differential privacy (DP) or paraphrasing-based methods to more comprehensively anonymize textual data while balancing privacy and utility.

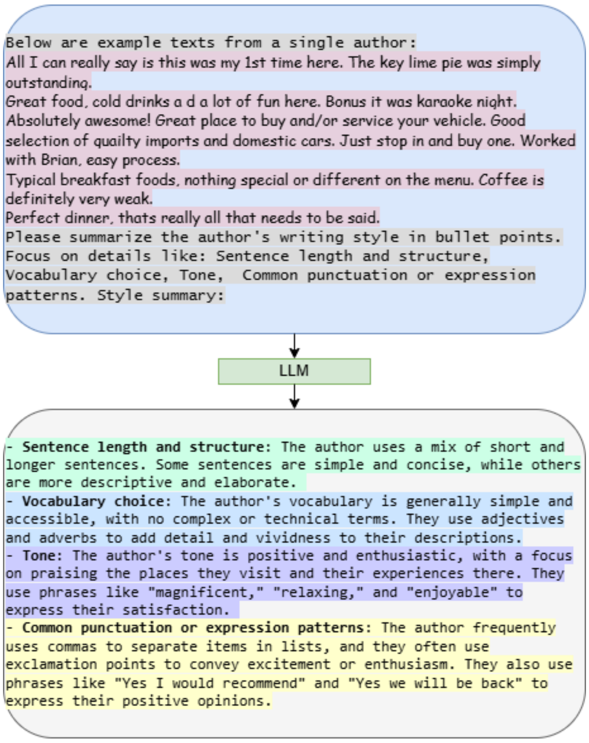
### 2.1 DP-based Methods

ER-AE (Bo et al., 2021) was among the first works in DP-based anonymization methods. It employs a Seq2Seq autoencoder architecture, perturbing latent embeddings through a two-set exponential mechanism. While effective in generating interpretable anonymized texts, ER-AE experiences limitations on longer texts due to stringent privacy budgets. Other DP-based methods such as DP-Paraphrase (Mattern et al., 2022), DP-MLM (Meisenbacher et al., 2024), and DP-Prompt (Utpala et al., 2023) have been developed to introduce calibrated noise at the token level by adjusting token probabilities during generation. For instance, DP-Prompt employs logit clipping and temperature scaling to achieve differential privacy guarantees; however, these modifications often negatively impact readability and semantic coherence at tighter privacy bounds (lower $\varepsilon$ values).

To mitigate some of these drawbacks, latent-space approaches such as DP-VAE (Weggenmann et al., 2022) perturb embeddings within an autoencoder framework, preserving fluency better than token-level perturbations. Despite these advantages, latent-space approaches may struggle with semantic coherence, particularly in short-form texts. Recently, (Meisenbacher and Matthes, 2024) proposed a Quasi-DP variant of DP-Prompt, removing logit clipping and retaining only temperature-based sampling. However, this approach achieves better performance only at weak privacy guarantees (higher $\varepsilon$ values), highlighting potential trade-offs between privacy bounds and data utility.

### 2.2 Non-DP Paraphrasing Methods

Early non-DP methods, such as Adversarial Stylometry (Brennan et al., 2012), relied on manual rewriting and editing, requiring significant effort from users to consciously alter their stylistic signatures. Recent automated approaches, such as JAMDEC (Fisher et al., 2024b), extract key content tokens from texts and utilize constrained diverse beam decoding with pre-trained language models (e.g., GPT2-XL) to generate paraphrases while maintaining semantic integrity. Similarly, STYLEREMIX (Fisher et al., 2024a) uses fine-grained LoRA adapters to systematically transform text over 7 specific style axes, such as formality and sentence length; however, this method necessitates additional pre-training on carefully curated style-specific corpora.



Below are example texts from a single author:
All I can really say is this was my 1st time here. The key lime pie was simply outstanding.
Great food, cold drinks a d a lot of fun here. Bonus it was karaoke night.
Absolutely awesome! Great place to buy and/or service your vehicle. Good selection of quailty imports and domestic cars. Just stop in and buy one. Worked with Brian, easy process.
Typical breakfast foods, nothing special or different on the menu. Coffee is definitely very weak.
Perfect dinner, thats really all that needs to be said.
Please summarize the author's writing style in bullet points. Focus on details like: Sentence length and structure, Vocabulary choice, Tone,  Common punctuation or expression patterns. Style summary:

- **Sentence length and structure:** The author uses a mix of short and longer sentences. Some sentences are simple and concise, while others are more descriptive and elaborate.
- **Vocabulary choice:** The author's vocabulary is generally simple and accessible, with no complex or technical terms. They use adjectives and adverbs to add detail and vividness to their descriptions.
- **Tone:** The author's tone is positive and enthusiastic, with a focus on praising the places they visit and their experiences there. They use phrases like "magnificent," "relaxing," and "enjoyable" to express their satisfaction.
- **Common punctuation or expression patterns:** The author frequently uses commas to separate items in lists, and they often use exclamation points to convey excitement or enthusiasm. They also use phrases like "Yes I would recommend" and "Yes we will be back" to express their positive opinions.

Figure 2: *Style Profiling*. Given a set of example texts from a single author, the large language model is prompted to generate a concise textual summary of the author's writing style.

(Meisenbacher and Matthes, 2024) also introduced a non-DP variant of DP-Prompt that discards explicit differential privacy mechanisms (logit clipping, temperature scaling) in favor of adjusting the top-$k$ sampling parameter alone.

Our proposed method aligns with recent work in leveraging large language models (LLMs) for text anonymization (Utpala et al., 2023; Meisenbacher and Matthes, 2024), but it significantly differs by utilizing explicit, prompt-driven stylistic profiles. Unlike previous methods relying on random noise injection or general paraphrasing, our style-aware paraphrasing explicitly targets stylistic markers, providing fine-grained control to balance anonymization and semantic content preservation.

## 3 Method

Given a set of authors $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, where each author $A_i$ has a corpus of texts $\mathcal{D}_i$, our objective is to produce an anonymized version of $\mathcal{D}_i$ such that an attacker is unlikely or less likely to correctly identify the ground-truth author $A_i$. We define the attacker model as an authorship attribution classifier $\mathcal{F}(\cdot)$, which, given a text $x \in \mathcal{D}_i$, returns either the predicted author label or a probability distribution over authors. Let $\hat{x}$ denote the anonymized version of $x$; our goal is to ensure:

$$\mathcal{F}(\hat{x}) \neq i \quad \text{(or low probability on } i\text{)}, \qquad (1)$$

i.e., the classifier should either misclassify $\hat{x}$ or assign a low likelihood to the true author $i$, while preserving the core semantic content of $x$.

Unlike existing DP-based approaches that anonymize by injecting noise into $x$, we propose to achieve anonymization by identifying and altering the stylistic markers in $x$ that distinguish the original author from others. Our pipeline (Algorithm 1) proceeds in two stages: (i) *Style Profiling*, and (ii) *Rewriting*, as described below.

### 3.1 Style Profiling

We define a function $\textsc{StyleSummarize}(\cdot)$ that uses a LLM to produce a short textual descriptor of an author's stylistic cues. Let $\mathcal{D}_i^{\text{train}} \in \mathcal{D}_i$ be the training subset for author $A_i$. We randomly sample $K$ texts from $\mathcal{D}_i^{\text{train}}$, forming a mini-batch $\mathcal{S}_i = \{x_1, x_2, \ldots, x_K\}$.

$$s_i \;=\; \textsc{StyleSummarize}\big(\mathcal{S}_i\big), \qquad (2)$$

where $s_i$ is a short textual summary of the author's writing style (Figure 2). Since Authorship attribution typically relies on consistent lexical and syntactic patterns (Sundararajan and Woodard, 2018), capturing these stylometric clues can help us systematically transform them during rewriting. In practice, $\textsc{StyleSummarize}$ is implemented by prompting a pretrained LLM:

$$s_i \sim p_\theta\big(\text{“Style summary”} \mid \mathcal{S}_i\big), \qquad (3)$$

with some temperature or top-$k$ sampling (Utpala et al., 2023).

### 3.2 Zero-Shot Rewriting

Given a text $x$ and style profile $s_i$, our goal is to produce an anonymized $\hat{x}$ that preserves meaning but reduces the author-specific style. We prompt the LLM with an instruction that references $s_i$ and requests rewriting of $x$ without those cues:

$$\hat{x} \;=\; \textsc{Rewrite}\big(x, s_i\big). \qquad (4)$$

Here, $\textsc{Rewrite}$ is implemented as a single forward pass of the LLM:

$$\hat{x} \sim p_\theta\big(\text{“Rewrite: no style”} \mid x, s_i\big). \qquad (5)$$

We assume no knowledge of the adversary's classifier (i.e., zero-shot).

**Algorithm 1** Style-aware Paraphrasing for Text Anonymization

**Input:** Language model (LM), Author corpora $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$, Style summarization function $\textsc{StyleSummarize}$, Rewriting prompt function $\textsc{Rewrite}$
**Output:** Anonymized corpus $\{\hat{\mathcal{D}}_1, \ldots, \hat{\mathcal{D}}_N\}$

**Style Summarization Prompt:** “*Below are example texts from a single author: {$\mathcal{S}_i$}. Please summarize the author's writing style in bullet points. Focus on details like: Sentence length and structure, Vocabulary choice, Tone, Common punctuation or expression patterns. Style summary:*”

**Rewriting Prompt:** “*Here is the author's style profile: {$s_i$}. Rewrite the following text so that it does **not** reflect these style cues, but retains the original meaning: {$x$}* Answer with the rewritten text only.”

```
1:  for each author $A_i$ in $\{1, \ldots, N\}$ do
2:      $\mathcal{S}_i \leftarrow$ random sample from $\mathcal{D}_i^{\text{train}}$
3:      $s_i \leftarrow \textsc{StyleSummarize}(\mathcal{S}_i)$      ▷ Style Summarization Prompt
4:  end for
5:  for each author $A_i$ in $\{1, \ldots, N\}$ do
6:      for each text $x$ in $\mathcal{D}_i$ do
7:          $p \leftarrow \textsc{Rewrite}(x, s_i)$      ▷ Rewriting Prompt
8:          $\hat{x} \leftarrow \text{LM}(p)$      ▷ One forward pass for rewritten text
9:          $\hat{\mathcal{D}}_i \leftarrow \hat{\mathcal{D}}_i \cup \{\hat{x}\}$
10:     end for
11: end for
12: return $\{\hat{\mathcal{D}}_1, \ldots, \hat{\mathcal{D}}_N\}$
```

### 3.3 Anonymization Framework

The anonymization pipeline is explained in Algorithm 1. Lines 1–4 describe the style-profiling step for each author, while lines 5–11 illustrate how we rewrite each text in a single pass, with an optional retry if the output is empty.

We highlight key advantages of our approach. First, the method is model-agnostic, requiring no training or fine-tuning to integrate with other LLMs. Second, unlike black-box anonymization techniques, it provides human-readable style profiles, which ensures transparency by identifying targeted linguistic patterns for removal. Third, the system generalizes effectively from as few as 5 samples per author, while handling texts with varying length and genres without domain adaptation, addressing scenarios common in real-world applications. The style-aware rewriting step preserves semantic content while suppressing style markers, as validated by both privacy and utility evaluations.

# 4 Experimental Setup

## 4.1 Baselines

Our baselines are originated from the study proposed by (Meisenbacher and Matthes, 2024), who extended the DP-Prompt method of (Utpala et al., 2023) and create three variants: (i) *DP*, a strict implementation of DP-Prompt with varying $\epsilon$ values; (ii) *Quasi-DP*, a relaxed version omitting logit clipping but retaining temperature-based sampling tied to $\epsilon$; and (iii) *Non-DP*, a paraphrasing variant removing DP constraints entirely. In our experiments, we compare against all three, with Non-DP serving as a particularly relevant baseline since it paraphrases the input without noise injection and also yields better results compared to the DP variants in most cases (Table 3). We report results for three representative parameter values per method: DP ($\epsilon = 25, 100, 250$), Quasi-DP ($\epsilon = 25, 100, 250$), and Non-DP ($k = 50, 10, 3$). These values capture the *low*, *moderate*, and *high* privacy levels. Due to space limitation, we exclude intermediate settings (e.g., $\epsilon = 50, 150$ or $k = 25, 5$) presented in the original study, as their results closely follow the expected trend between adjacent values. Additionally, we do not report results for other approaches discussed in 2.1, as their evaluation frameworks often diverge significantly from ours, particularly in terms of datasets. The selected baselines are not only methodologically aligned with our approach (e.g., LLM-based frameworks) but also share comparable evaluation protocols, enabling a fair and meaningful comparison.

## 4.2 Dataset

Similar to (Meisenbacher and Matthes, 2024), we adopt the Author10 subset from the Blog Authorship Corpus (Schler et al., 2006). In addition to this, we consider a smaller, short-form text scenario using a subset of Google Reviews (Li et al., 2022). The original Google Reviews dataset contains over 666 million reviews from more than 113 million users across the United States. In this paper, we focus on the state of Illinois, one of the states with the largest number of reviews, and then select the top 9 most frequent reviewers. This yields a set of approximately 3,959 reviews, each typically 1–2 sentences in length. We refer to this dataset as Illinois9 in our experiments. A summary of both datasets, including document length distributions, is provided in Table 1.

| Dataset | Authors | Docs | D/A | W/S | S/D |
|---|---|---|---|---|---|
| **Author10** | 10 | 15,070 | 1,507 | 14.3<br>(±12.8) | 4.74<br>(±4.56) |
| **Illinois9** | 9 | 3,959 | 439.89 | 9.19<br>(±4.82) | 2.24<br>(±1.89) |

Table 1: Dataset statistics for Author10 (long-form blogs) and Illinois9 (short reviews). *D/A* = documents per author, *W/S* = words per sentence, *S/D* = sentences per document. Values in (.) indicate standard deviations.

## 4.3 Evaluation Metrics

We follow the overall evaluation protocol from (Meisenbacher and Matthes, 2024) using both utility and privacy metrics to assess our anonymization strategy. However, to better capture the divergence from the informational content of the original text, we also report the distance between the original and paraphrased text using a weighted KL divergence, which penalizes the disappearance of rare, informative tokens by weighting the KL divergence with their IDF.

*Utility metrics.* We evaluate model utility by measuring the semantic similarity between the original and anonymized versions of the text. As in (Meisenbacher and Matthes, 2024), we embed each text with three pre-trained models ALL-MINILM-L6-V2, ALL-MPNET-BASE-V2, and GTE-SMALL (Li et al., 2023) and report the averaged *cosine similarity (CS)* across the models. In addition, we report the *perplexity (PPL)* (Weggenmann et al., 2022) for both versions of texts, where the perplexity is computed with GPT-2 (Radford et al., 2019). A higher *CS* and lower *PPL* typically signify better content preservation and readability. For methods delivering comparable privacy scores (Table 5), we additionally report a *weighted KL divergence* that is sensitive to the loss of rare yet informative tokens. Let $P$ and $Q$ denote the normalized term–frequency distributions of the original document and its anonymized counterpart, respectively; rare tokens are up-weighted by their inverse document frequency IDF:

$$D_{\text{W-KL}}(P \parallel Q) = \sum_{t \in V} \mathrm{IDF}(t)\, P(t)\, \log \frac{P(t)}{Q(t)},$$

|  | $CS(P, Q_i)$ | $D_{\mathrm{KL}}(P \| Q_i)$ | $D_{\mathrm{W\text{–}KL}}(P \| Q_i)$ |
|---|---|---|---|
| $Q_1$ | 0.67 | 7.31 | 7.31 |
| $Q_2$ | 0.67 | 7.31 | **36.55** |

Table 2: Toy example against $P$: "quokka quietly grazes" →{quokka, quietly, grazes}. $Q_1$: "quokka quietly munches" →{quokka, quietly, munches}. $Q_2$: "animal quietly grazes" →{animal, quietly, grazes}. All KL divergences are computed with smoothing $\epsilon = 10^{-10}$.

where $V$ is the vocabulary of the original corpus. For instance, as shown in Table 2, the weighted KL divergence can capture the semantic importance of missing "quokka". While both paraphrased sentences, i.e., $Q_1$ and $Q_2$ share the same cosine similarity with the original sentence, $P$, the weighted KL score sharply penalizes $Q_2$ for dropping a rare, informative word. As can be seen, lower values indicate that the anonymized text better retains the informative token distribution of the original text.

*Privacy metrics.* We report the *BLEU* (Papineni et al., 2002) score between the original and anonymized versions of the text. A lower BLEU may indicate greater lexical transformation, potentially beneficial for privacy. Second, we quantify privacy by measuring how effectively anonymized text conceals the identity of its original author. Specifically, we adopt an authorship attribution scenario (Mattern et al., 2022; Weggenmann et al., 2022), where an authorship classifier is trained on the original training set but tested on the anonymized test set. A lower F1 indicates stronger privacy, as it reflects the classifier's diminished ability to link anonymized text to the correct author. For the Author10 dataset, we trained a DeBERTa-v3 (He et al., 2021) model for authorship attribution using the same parameters as (Meisenbacher and Matthes, 2024). For the Illinois9 dataset, we employed BERT (Devlin, 2018), as it has demonstrated high performance, particularly on short-form review datasets such as IMDb, which closely resemble Illinois9 (Fabien et al., 2020). To jointly assess the balance between utility preservation and privacy protection, we additionally report the *relative gain* metric $\gamma$ (Mattern et al., 2022; Meisenbacher and Matthes, 2024):

$$\gamma := \frac{S_p}{S_o} - \frac{A_p}{A_o},$$

where $A$ and $S$ are authorship F1 scores and CS on original (subscript $o$) and anonymized (subscript $p$) data, respectively.

### 4.4 Parameter Variations and Setup

*LLM Models.* We adopt Llama-3.2-3B-Instruct (Touvron et al., 2023) as our primary model due to its instruction-tuned architecture, which makes it well suited for generating rich style profiles. Furthermore, its open-source nature aligns with our privacy-centric design goals, ensuring transparency and reproducibility in our framework. To demonstrate the model-agnostic nature of our framework, we also run our experiment with MiniCPM3-4B (Hu et al., 2024), a versatile and highly capable model that surpasses many larger models (e.g., GPT-3.5-Turbo and Phi-3.5-mini-Instruct) on various benchmarks demonstrating its generalization, reasoning, and instruction-following capabilities.

*Sample Size for Building Style Profiles.* We experimented with $K \in \{2, 5, 10, 20\}$ texts per author to prompt the LLM for a "style profile," as introduced in Section 3.1, where $K$ denotes the number of example texts used to capture an author's stylistic features. As shown in Appendix C, even $K = 5$ examples are sufficient to produce robust and distinctive profiles for both Author10 and Illinois9, so we adopt this as our default setting.

## 5 Results & Discussion

Table 3 summarizes the performance of three major approaches listed in Section 2.1 and 2.2, i.e., DP-Prompt with varying $\varepsilon$, Quasi-DP, and Non-DP against *Ours*.

### 5.1 Privacy Performance

As shown in Table 3, our method reduces authorship attribution accuracy by 50–70%, achieving an F1 score of 26.02 (vs. original 66.45) on the Author10 dataset and 20.76 (vs. original 76.78) on the short-form Illinois9 corpus. These results significantly outperform the Non-DP paraphrasing baseline and surpass DP-Prompt at nearly all privacy levels. For example, on Author10 at $k$=3, our method reduces F1 to 26.02, a 51% improvement over the Non-DP paraphrasing method (53.10), which retains identifiable authorship signals. The performance on Illinois9 in particular demonstrates robust anonymization even in short-text domains where stylistic cues are sparse, contrasting with Non-DP paraphrasing, which struggles to disrupt authorship patterns in such settings.

As far as DP-Prompt is concerned, the sole exception is Author10 at $\varepsilon$=25, where DP achieves

| | | Baseline | DP ($\varepsilon$) | | | Quasi-DP ($\varepsilon$) | | | Non-DP ($k$) | | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\varepsilon$ / $k$ | | 25 | 100 | 250 | 25 | 100 | 250 | 50 | 10 | 3 | Llama | MiniCPM |
| **Author10** | CS ↑ | 1 | 0.589 | 0.812 | 0.832 | 0.347 | 0.810 | 0.833 | 0.710 | 0.750 | 0.787 | 0.702 | 0.82 |
| | BLEU ↓ | 1 | 0.077 | 0.123 | 0.153 | 0.001 | 0.121 | 0.153 | 0.049 | 0.063 | 0.088 | 0.023 | 0.213 |
| | PPL ↓ | 41 | 8770 | 928 | 905 | 16926 | 982 | 925 | 816 | 1080 | 837 | 42.47 | 61.53 |
| | Author F1 (s) ↓ | 66.45 | 7.13 | 58.10 | 60.60 | 6.59 | 57.84 | 61.13 | 46.83 | 49.88 | 53.10 | 26.02 | 49.46 |
| | Gain ($\gamma$) | - | 0.482 | -0.062 | -0.080 | 0.248 | -0.060 | -0.087 | 0.005 | -0.001 | -0.012 | 0.31 | 0.076 |
| **Illinois9** | CS ↑ | 1 | 0.592 | 0.894 | 0.914 | 0.595 | 0.892 | 0.916 | 0.812 | 0.84 | 0.879 | 0.709 | 0.89 |
| | BLEU ↓ | 1 | 0.013 | 0.432 | 0.497 | 0.015 | 0.424 | 0.52 | 0.255 | 0.292 | 0.373 | 0.022 | 0.31 |
| | PPL ↓ | 98.83 | 220.66 | 89.15 | 96.99 | 222.16 | 89.63 | 94.93 | 82.19 | 75.89 | 84.72 | 53.36 | 43.05 |
| | Illinois F1 (s) ↓ | 76.78 | 23.42 | 61.73 | 64.86 | 21.84 | 59.90 | 65.35 | 49.32 | 51.22 | 54.94 | 20.76 | 47.90 |
| | Gain ($\gamma$) | - | 0.287 | 0.09 | 0.069 | 0.311 | 0.112 | 0.065 | 0.17 | 0.173 | 0.163 | 0.439 | 0.266 |

Table 3: Privacy and Utility performance on the Author10 and Illinois9 datasets. Metrics include Cosine Similarity (CS), BLEU, Perplexity (PPL), Authorship F1 (*static*), and Relative Gain ($\gamma$). Baseline refers to the original (unaltered) input text. For DP-based methods, smaller $\epsilon$ values correspond to tighter privacy guarantees.

stronger anonymity with F1 = 7.13 vs. ours (26.02). However, this comes at a significant cost: DP's perplexity at $\varepsilon$=25 is 8,770, more than 200× higher than ours (42.47) which suggests that such privacy guarantees have come at the cost of significantly reduced readability.

## 5.2 Utility Preservation

We report near-baseline perplexity for the LLaMA variant (42.47 and 53.36 vs. original 41 and 98.83), and even lower perplexity for MiniCPM on Illinois9 (Table 3). This outperforms DP methods at strong privacy levels (PPL >1,000) and Non-DP paraphrasing on Author10 (PPL 800–1,080). Quasi-DP performs even worse in some settings: on Author10 at $\varepsilon$=25, its perplexity exceeds 16,000, and remains well above baseline even at higher $\varepsilon$ levels. It is important to highlight that despite significant stylistic perturbation (BLEU = 0.023), the perplexity remains stable, indicating that our approach successfully targets *stylistic* rather than *semantic* features. This is further confirmed by the high cosine similarity (≈ 0.70) even as the authorship F1 drops.

## 5.3 Full *vs.* Single-Dimension Style Profiles

| Metric | Baseline | Full Profile | Length |
|---|---|---|---|
| CS ↑ | 1.000 | 0.709 | 0.713 |
| BLEU ↓ | 1.000 | 0.022 | 0.022 |
| PPL ↓ | 98.83 | 53.36 | 54.71 |
| Author F1 (s) ↓ | 76.78 | 20.76 | 19.32 |
| Gain ($\gamma$) | – | 0.439 | 0.461 |

Table 4: Ablation on style-profile types (Illinois9). See Appendix C for full results.

All earlier results rely on the default configuration of Algorithm 1. During StyleProfiling (lines 1–4) the LLM produces a *Full* profile based on four features: tone, vocabulary choice, length, and punctuation. To measure the impact of each cue, we repeat exactly the same procedure with four single-dimension profiles (*Tone*, *Length*, *Vocab*, and *Punc*). We then supply the chosen profile together with the source text in a separate rewriting prompt, asking the LLM to preserve the original meaning while removing the style markers named in the profile. As Table 4 shows, the *Full* profile delivers the best overall privacy–utility trade-off. However, the *Length*-only variant comes surprisingly close for Illinois9. It matches or slightly exceeds the *Full* profile in privacy gain (ΔF1) while preserving marginally higher cosine similarity. We attribute this to the short, structurally uniform nature of review texts, where sentence length is already a strong authorial fingerprint. On the more varied Author10 corpus the gap narrows: *Tone*, *Length*, and even *Vocab* come within 0.2–0.3 pt of the Full profile on both F1 and CS (see Appendix C). While no single cue dominates across datasets, the Full profile remains the most *stable* choice, delivering strong privacy–utility trade-offs.

## 5.4 Effect of the Base LLM: Llama-3.2-3B *vs.* MiniCPM3-4B

Table 3 also shows how the choice of LLM affects the performance of our method. Across both datasets, Llama consistently yields substantially lower Authorship-F1 and BLEU scores, indicating that it is more effective at neutralising stylistic cues and tends to perform more aggressive lexical rewrites. By contrast, MiniCPM produces higher CS scores, reflecting stronger surface-level content preservation, but also higher Authorship-F1, signalling that a greater share of stylistic signal remains detectable. With regard to perplexity, the two models display complementary behaviour. Llama demonstrates slightly higher PPL on short texts, while MiniCPM consistently maintains lower perplexity, but at the cost of weaker privacy. Therefore, choosing Llama leads to a balanced privacy-utility tradeoff as indicated by the higher relative gain. This advantage likely stems from Llama's stronger instruction-following capabilities.

### 5.5 Style-Guided, Semi-Guided, and Unguided (Paraphrase) Rewrites

*Rewrite variants.* Our default *style-guided* pipeline gives the LLM an explicit style profile and asks it to "rewrite the text so it no longer reflects these cues." We ablate two other variants: (i) *Semi-guided* uses the same prompt as above but *without* supplying a profile, and (ii) *Unguided / Paraphrase* does not mention style at all; the model is prompted to "rewrite the text."

Table 5 contrasts our standard *style-guided* rewrite with the other two variants (Llama-3.2-3B-Instruct). On Author10 the *semi-guided* setting achieves slightly better Author-F1 scores but at the cost of noticeably worse utility: *PPL* rises, *CS* falls, and *weighted KL divergence* increases, indicating greater distortion of informative tokens. Further analyses in Appendix A (GPT-4 preference judgements and named-entity retention) confirm that our *style-guided* rewrite maintains more informational content while achieving nearly the same privacy levels as the *semi-guided* variant. Taken together, these patterns also hint that the authorship classifier is not relying solely on stylistic cues but also draws on residual content when identifying the author. A fully *unguided/paraphrase* version does even worse, yielding the largest utility loss while leaving a sizable authorship signal, underscoring the value of explicit style guidance.

| | Style-Guided | Semi-Guided | Paraphrase |
|---|---|---|---|
| **Author10** | | | |
| CS ↑ | 0.702 | 0.694 | 0.752 |
| BLEU ↓ | 0.023 | 0.018 | 0.034 |
| PPL | 42.47 | 42.39 | 56.04 |
| Author F1 ↓ | 26.02 | 24.10 | 36.61 |
| Weighted KL ↓ (All) | 48.75 | 49.48 | 41.11 |
| Weighted KL ↓ (Subset) | 39.39 | 42.96 | 33.60 |
| **Illinois9** | | | |
| CS ↑ | 0.709 | 0.746 | 0.791 |
| BLEU ↓ | 0.022 | 0.027 | 0.039 |
| PPL | 53.36 | 52.85 | 61.46 |
| Author F1 ↓ | 20.76 | 22.64 | 29.37 |
| Weighted KL ↓ (All) | 58.33 | 57.16 | 46.87 |
| Weighted KL ↓ (Subset) | 52.10 | 49.97 | 41.59 |

Table 5: Comparison of Style-Guided vs. Semi-guided and Paraphrase rewrites on Author10 and Illinois9."Subset" shows performance on only the subset of samples where one method successfully prevented author re-identification while the other method did not.

### 5.6 Discussion

Our results confirm that explicit, multi-dimension style profiles are the key to a favourable privacy–utility balance. Style guided rewrites consistently suppress authorship signal by 50–70% while maintaining high CS and near baseline PPL scores, something neither DP noise nor token-level paraphrasing can match. Although the semi-guided variant trims Author-F1 slightly further on Author10, the extra privacy gain is achieved at a clear utility cost: the weighted KL divergence rises while CS falls. Further analysis provided in Appendix. A.1 and Appendix A.2 also confirms that the higher privacy levels of *semi-guided* variant is achieved largely by discarding meaning rather than by precise style removal.

The ablation in Table 4 adds nuance: on the short, uniform Illinois9 reviews, *Length*-only guidance nearly matches the *Full* profile, suggesting that minimal cues can suffice when structure is consistent. For longer, more varied posts, however, the full profile remains clearly superior.

Overall, our findings reinforce that structured, style-aware guidance offers a reliable way to navigate the privacy–utility tradeoff in anonymized text generation.

## 6 Conclusion

We present a novel text anonymization framework that employs explicit, prompt-driven stylistic profiling to effectively mask authors' identity while preserving semantic content. Our approach demonstrated robust performance across both long-form (blogs) and short-form (reviews) text datasets, significantly outperforming differential privacy-based and paraphrasing baselines. Comprehensive evaluations suggests that our style-guided anonymization better retains the original content and meaning compared to noise-based anonymization approaches. By rigorously analyzing the privacy–utility tradeoff, we highlighted the critical role of precise stylistic manipulation in anonymization effectiveness.

## 7 Limitations

While this study demonstrates promising outcomes, it also presents several avenues for future investigation. Notably, the proposed method depends on the capacity of LLMs to accurately extract and generalize stylistic features such as syntactic constructions (e.g. passive-voice preferences), idiosyncratic collocations, or rhetorical patterns (parallelism, anaphora). As illustrated in Appendix D (Table 9), certain authorial fingerprints can slip through our hand-picked cues.Therefore, anonymization performance may be shaped by the quality and characteristics of the specific LLM employed. This highlights the need for further research into model selection and refinement, as well as the development of automated or unsupervised style profiling techniques to reduce potential biases associated with manually selected features.

Privacy in our study is primarily assessed via the drop in F1 of an authorship classifier trained on the original corpus, which has been a standard approach in prior work. However, as shown in Section 5.5, the classifier may also leverage content signals, meaning lower F1 can reflect both style obfuscation and content loss. This underscores the need for more nuanced evaluation frameworks that disentangle these factors and provide clearer guidance for improving stylistic anonymization.

## 8 Ethical Consideration and Potential Risks

Our style-aware paraphrasing framework effectively conceals authorship, but it also carries the risk of adversarial misuse. Malicious actors could leverage this technique to disguise the provenance of disinformation, hate speech, or other harmful content, thereby evading forensic attribution and undermining content moderation efforts. Such abuse underscores the need for responsible deployment. For detailed results on demographic-attribute inference (gender and age), see Appendix B.

## References

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online. Association for Computational Linguistics.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3).

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024a. 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.

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024b. Jamdec: Unsupervised authorship obfuscation using constrained decoding over small language models. *Preprint*, arXiv:2402.08761.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

David I. Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Jiacheng Li, Jingbo Shang, and Julian McAuley. 2022. UCTopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6169, Dublin, Ireland. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130*.

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

Stephen Meisenbacher and Florian Matthes. 2024. Thinking outside of the differential privacy box: A case study in text privatization with language model prompting. *Preprint*, arXiv:2410.00751.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Athira Usha and Sabu M. Thampi. 2017. Authorship analysis of social media contents using tone and personality features. In *Security, Privacy, and Anonymity in Computation, Communication, and Storage (SpaCCS)*, pages 212–228. First Online: 07 December 2017.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 721–731, New York, NY, USA. Association for Computing Machinery.

## A Supplementary Analyses on Guided *vs.* Semi-guided Rewriting

To better understand the small privacy edge of the *Semi-guided Rewrite* over the *Style-guided* variant in Author10 dataset, we ran two complementary analyses: an LLM-based content judgement and a named-entity retention study, alongside the weighted KL results already reported in Table 5. Each analysis was carried out on (i) the full test set and (ii) the subset where the *semi-guided* method achieves lower Author-F1 (121 cases).

| | **Style-Guided** | **Semi-guided** |
|---|---|---|
| Full Corpus | 847 | 649 |
| Subset (n=121) | 105 | 16 |

Table 6: GPT-4 preferences for content retention between style-guided and semi-guided rewrites (Author10).

### A.1 LLM-Based Content Judgment

Recent NLP studies have utilized LLMs such as GPT-4 (OpenAI et al., 2024) for human-aligned evaluative tasks, demonstrating strong agreement with human judgments on content quality and informativeness (Liu et al., 2023) (Li et al., 2024).

Leveraging this capability, we employ GPT-4 to objectively compare our style-guided and semi-guided rewriting methods with respect to content retention. Such automated evaluations have been shown to correlate strongly with human annotation, providing a scalable and consistent measurement approach (Wang et al., 2023).

We conducted a comparative evaluation using GPT-4 to systematically assess content retention between the two anonymized versions. For each triplet $\langle x, \hat{x}_{\mathrm{style-guided}}, \hat{x}_{\mathrm{semi-guided}} \rangle$, GPT-4 was prompted to select the anonymized variant that better preserves the semantic content of the original.

On the full corpus, GPT-4 preferred the style-guided version in 847 cases compared to 649 for the *semi-guided Rewrite* (Table 6), highlighting that incorporating explicit stylistic profiles significantly improves semantic preservation. Even within the 122-text subset where the *semi-guided Rewrite* demonstrated better privacy than the style-guided version for all samples, the style-guided method was overwhelmingly favored 105 to 16. This indicates that the marginal privacy gains achieved by the semi-guided method were largely due to excessive removal or distortion of meaningful content, rather than targeted style anonymization. Our style-guided approach demonstrates superior performance in maintaining semantic integrity, effectively balancing the trade-off between privacy and utility.

### A.2 Named-Entity Retention

We evaluated named-entity retention, which is a critical indicator of semantic fidelity given the nature of our datasets. In blog posts (Author10), named entities often include personal references, locations, and events crucial for maintaining narrative coherence. We define the entity retention loss per document as:

$$\Delta_{\text{ents}}(m) = |\text{ents}(x)| - |\text{ents}(\hat{x}_m)|,$$

$$m \in \{\mathrm{style-guided}, \mathrm{semi-guided}\}.$$

Across the full test set ($N = 1{,}503$), the profile-guided method yielded an average loss of 1.95 entities per text, whereas the *semi-guided Rewrite* exhibited a higher average loss of 2.11 entities. Of the total 5,434 entities originally identified, the profile-guided approach preserved 2,507 entities (46.1%), significantly outperforming the *semi-guided Rewrite*, which retained only 2,267 entities (41.7%). These results confirm that the *semi-guided Rewrite* attains marginally better privacy metrics by at the cost of sacrificing named entities that are inherently crucial to the semantic and contextual value of the original text.

## B Age and Gender Inference

Prior work on text privatization, most notably (Meisenbacher et al., 2024) also measures privacy by how well an adversary can infer sensitive attributes such as *gender* and *age*. To keep our results comparable, we adopt their *static* evaluation setting on Author10. A DEBERTA-V3-BASE classifier is fine-tuned for three epochs on the original train split and then evaluated on the anonymized test split.[1] Because Author10 is less demographically diverse than Topic10 (the dataset used by (Meisenbacher et al., 2024)), we can form only *three* age bins (rather than five)

As shown in Table 7, we observe significant drops in attribute inference: Gender-F1 falls from 0.66 to 0.55 (17%), and age F1 from 0.73 to 0.43 (41%). These gains are achieved without any attribute-specific paraphrasing, suggesting that demographic signals are strongly encoded in writing style and are attenuated as a by-product of our style neutralization, echoing prior sociolinguistic findings (Johannsen et al., 2015).

| Method | Gender F1 ↓ | Age F1 ↓ |
|---|---|---|
| Baseline | 0.66 | 0.73 |
| Full | 0.56 | 0.44 |
| Len | 0.56 | 0.44 |
| Tone | 0.57 | 0.43 |
| Vocab | 0.55 | 0.44 |
| Punc | 0.55 | 0.43 |

Table 7: Adversarial inference of gender and age on anonymized Author10. Lower F1 implies stronger privacy.

## C Change in Style Profiles

Table 8 reports performance when the LLM is guided by either a full style profile or a single stylistic dimension (Length, Tone, Vocab, Punc). While the full profile performs best overall, the margin over single-feature profiles is surprisingly small; in several cases *Length* even matches the full profile on relative gain ($\gamma$) while preserving slightly higher cosine similarity. This suggests that the LLM can implicitly infer additional stylistic cues even when prompted with only one salient feature.

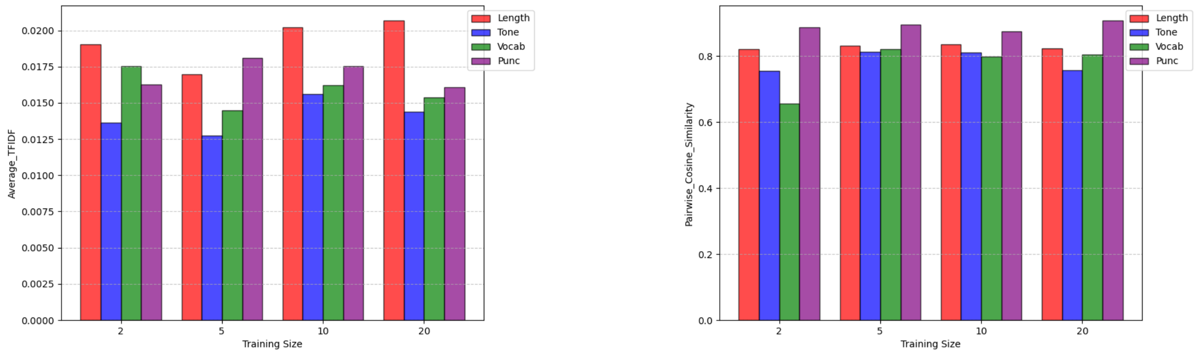
*Distinctiveness of individual profiles.* To understand why some single features still work well, we measure two corpus-level statistics for each feature-specific profile set, varying the number of training examples used to create the profile ($K \in \{2, 5, 10, 20\}$). We run this analysis on the shorter Illinois9 reviews, because with longer Author10 posts the $K{=}20$ prompt would exceed the LLM's context window, preventing reliable profile generation.

- *Average TF–IDF* $\overline{T}$ (Fig. 3a, left) estimates how much *unique lexical information* each profile carries. *Length* consistently yields the highest $\overline{T}$, followed by *Vocab*, whereas *Punc* scores lowest.
- *Average pairwise cosine similarity* $\overline{C}$ (Fig. 3b, right) gauges how *distinct* profiles are across authors. Lower values imply stronger author discrimination; again *Length* is most distinctive, while *Punc* clusters tightly.

Both $\overline{T}$ and $\overline{C}$ increase modestly from $K{=}2$ to $K{=}5$ and then level off, confirming that as few as five random examples are sufficient to capture a representative style profile. The strong performance of *Length* in the privacy metrics aligns with its combination of high lexical uniqueness and clear inter-author separability.

*Correlation with downstream performance.* Across the 16 feature–size combinations (four fea-

[1] Adaptive results are omitted because our focus is authorship obfuscation; the static scores are sufficient to illustrate the trend.



(a) Average TF-IDF

(b) Average Pairwise CS

Figure 3: Lexical uniqueness ($\overline{T}$, left) and inter-author distinctiveness ($\overline{C}$, right) of single-feature style profiles on Illinois9, across different profile sizes $K$.

| | **Illinois9** | | | | | | **Author10** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | Baseline | Full | Length | Tone | Vocab | Punc | Baseline | Full | Length | Tone | Vocab | Punc |
| CS ↑ | 1 | 0.709 | 0.713 | 0.711 | 0.712 | 0.714 | 1 | 0.702 | 0.702 | 0.703 | 0.704 | 0.703 |
| BLEU ↓ | 1 | 0.022 | 0.022 | 0.021 | 0.021 | 0.023 | 1 | 0.023 | 0.027 | 0.024 | 0.029 | 0.028 |
| PPL ↓ | 98.83 | 53.36 | 54.71 | 54.67 | 55.66 | 54.57 | 41 | 42.47 | 42.75 | 42.27 | 41.96 | 42.57 |
| Authorship F1(s) ↓ | 76.78 | 20.76 | 19.32 | 20.36 | 21 | 19.24 | 66.45 | 26.02 | 25.85 | 24.35 | 24.78 | 24.77 |
| Gain ($\gamma$) | - | 0.439 | 0.461 | 0.446 | 0.438 | 0.463 | - | 0.31 | 0.313 | 0.337 | 0.331 | 0.33 |

Table 8: Performance using full vs. single-feature profiles on Illinois9 and Author10. Lower is better for BLEU, PPL, and Authorship F1; higher is better for CS and $\gamma$.

tures, four values of $K$), a random-forest LOOCV indicates the same pattern suggested by Figure 3: the average TF–IDF score $\overline{T}$ is the strongest predictor of privacy gain $\Delta$F1 (feature importance = 0.67; $R^2 = 0.19$), whereas average pairwise cosine $\overline{C}$ is most predictive of utility loss in PPL (importance = 0.55; $R^2 = 0.48$). Although the explained variance is moderate, there is still a trend. Profiles that are both lexically rich and author-distinctive (most notably **Length**) produce larger privacy gains with an acceptable utility cost, helping single-feature guidance approach the performance of the full profile in many cases.

## D Qualitative Case Study

The central dilemma of text anonymization lies in removing stylistic fingerprints without erasing essential content. Low–$\varepsilon$ DP outputs clearly prioritise privacy, but the added noise obliterates the meaning of a sentence. Conversely, Non-DP and Quasi-DP varaints often seem to rephrase fluently but semantically drift from the original text. For example in Table 9, the original text is a back-handed compliment: it sounds positive while implicitly teasing the person for merely meeting a low bar. Temperature sampling flattens this nuance into a bland affirmation like *"Rosie is a great speaker."* Although grammatically correct, the rewrite erases the original sarcasm, thereby degrading information fidelity while still leaving enough anchors for re-identification.

Our *style-guided* rewrites deliberately keep a named entity such as "Rosie" when it is essential to the sentence's meaning, accepting a slight privacy cost to avoid distorting the author's intent. This leads to the identification of the author, echoing with our earlier findings in section 5.6 that the authorship classifier not just relies on style cues but also content cues. Erasing every salient token would prevent identification but at the price of losing the intended message. Across all other cases, our style-guided variant successfully replaces distinctive tokens or restructures the sentence, delivering strong privacy gains while consistently preserving the meaning. Additional examples are provided in Table 10 and Table 11.

| Method | $\epsilon/k$ | Anonymized Text | Identified? | Weighted KL |
|---|---|---|---|---|
| Original text | — | Wow. I'm glad Rosie talks in complete sentences. It'd be kind of annoying if she didn't. | Yes | — |
| **DP** | | | | |
| | 25 | Wow, it will be like "wow if they doesn't!" so you've probably heard what Rosie's heard. "Wow, and indeed Rosie spelled it correctly," this was a reaction the ghuggish female leader honed. | No | 42.06 |
| | 50 | Rosie is a complete speaker. | Yes | 62.84 |
| | 100 | Rosie is a complete speaker. | Yes | 62.84 |
| | 150 | Rosie is very good at talking in sentences. | Yes | 61.13 |
| | 250 | Rosie is a good speaker. | Yes | 67.93 |
| **Quasi-DP** | | | | |
| | 25 | Sore words | No | 71.88 |
| | 50 | Wow. I'm glad Rosie talks in complete sentences. It would be kind of annoying if she didn't. | Yes | 3.03 |
| | 100 | Rosie talks in complete sentences. | Yes | 49.35 |
| | 150 | Rosie is good at talking. | Yes | 67.93 |
| | 250 | Rosie is very good at speaking. | Yes | 67.97 |
| **Non-DP** | | | | |
| | 3 | I'm glad Rosie talks in complete sentences. | Yes | 25.59 |
| | 5 | Rosie is a great speaker. | Yes | 67.93 |
| | 10 | Rosie is good at words. | Yes | 67.93 |
| | 25 | Rosie talks in full sentences. | Yes | 54.44 |
| | 50 | I'm glad Rosie talks in terms of sentences. | Yes | 29.96 |
| **Ours** | | | | |
| | Style guided | Rosie's ability to speak in full sentences is refreshing; otherwise, her conversations might become tedious. | Yes | 44.40 |
| | Semi-guided | It's refreshing to hear someone communicate effectively and clearly without filler words or short phrases. | No | 54.00 |

Table 9: Rewritten texts (short form) from Author10 dataset. Note: For DP-based methods, smaller $\epsilon$ values correspond to tighter privacy guarantees.

## E Implementation and Hyperparameter Settings

All experiments were run on a single-GPU kernel in Amazon SageMaker Studio, using the default PyTorch `pytorch_p310` container (Python 3.10.8). We used NVIDIA T4 (16 GB) or A10G (24 GB) GPUs with CUDA driver/toolkit 12.1. The software stack comprised PyTorch 2.2.2, Transformers 4.51.3, Hugging Face Hub 0.31.2 and the OpenAI Python client 1.23.6. Additional packages included `sentence-transformers` 4.1.0, `evaluate` 0.4.3, `sentencepiece` 0.2.0, `bitsandbytes` 0.45.5, spaCy 3.7.4 (model *en_core_web_sm* 3.7), pandas 2.2.3, numpy 1.26.4, tqdm 4.67.1, seaborn 0.13.2 and scikit-learn 1.4.2.

### E.1 Large-Language-Model Inference

All LLM calls were inference only (no fine-tuning, LoRA or gradient checkpointing). We ran both MiniCPM (`openbmb/MiniCPM3-4B`) and LLaMA (`meta-llama/Llama-3.2-3B-Instr.`) in BF16 with automatic device mapping. For each model, *style profiling* used a temperature of 0.3, top-p of 0.9, up to 256 new tokens and a repetition penalty of 1.2. *Rewriting* used a temperature of 0.7, top-p of 0.7, up to 64 new tokens and the same repetition penalty. Unguided paraphrasing also ran in BF16 with temperature 1.0, top-p 0.9, up to 64 new tokens and repetition penalty 1.2. Judgments were obtained via the GPT-4 API (`gpt-4-0314`) with temperature set to 0 and a 2048-token context window.

### E.2 Style-Profiling Parameters

Each style profile was built from up to five author snippets (`max_examples=5`). We experimented with five prompt variants (`prompt_type` $\in$ {*full*, *length*, *vocab*, *tone*, *punc*}) and allowed up to three retry attempts for both profile generation and rewriting.

### E.3 Authorship and Attribute Classifiers

We trained a BERT-base(Devlin, 2018) sequence classifier on the Illinois9 data for three epochs (batch size 32; learning rate $2 \times 10^{-5}$; warm-up ratio 0.1; weight decay $10^{-2}$; max-norm 1.0).

| Method | $\epsilon/k$ | Anonymized Text | Identified? | Weighted KL |
|---|---|---|---|---|
| Original text | — | So much to see and do. Shops are upscale and offer such a great selection of products. | Yes | — |
| **DP** | | | | |
| | 25 | Visit. Also love going shopping by shopping street. | Yes | 62.76 |
| | 50 | Among the many shopping shops: the high-end resorts, all-inclusive resort or just the old-fashioned way. | No | 57.32 |
| | 100 | The shops are upscale and offer such a great selection of products. | No | 17.59 |
| | 150 | Great shopping and restaurants. | No | 58.65 |
| | 250 | The shopping is upscale and the restaurants are very good. | No | 52.12 |
| **Non-DP** | | | | |
| | 3 | There are so many things to see and do in the city. | No | 43.65 |
| | 5 | The shops offer a wide range of goods and services. | No | 45.71 |
| | 10 | The upscale shopping area is a great place to visit if you are in the mood for upscale shopping. | No | 49.44 |
| | 25 | Shops at Luxor are very well appointed and have beautiful architecture. The hotel is very good for family or business. | No | 52.63 |
| | 50 | Amazing shopping and unique shops. | Yes | 54.97 |
| **Quasi-DP** | | | | |
| | 25 | One can travel around city—there are so much to look at and do. | No | 44.04 |
| | 50 | They have numerous boutiques and restaurants, all in a nice hotel. | No | 59.76 |
| | 100 | The shopping area is spacious and well kept. | No | 60.80 |
| | 150 | The shops are upscale and offer such a great selection of products. | No | 17.59 |
| | 250 | The shops are upscale and offer such a great selection of products. | No | 17.59 |
| **Ours** | | | | |
| | Style guided | The area boasts an impressive array of high-end shopping options, featuring a diverse range of premium goods. | No | 59.71 |
| | Semi-guided | There is plenty to explore in this area, with high-end stores providing an extensive range of goods available for purchase. | No | 59.15 |

Table 10: Rewritten texts from Illinois9 dataset. Note: For DP-based methods, smaller $\epsilon$ values correspond to tighter privacy guarantees.

For Author10 (including gender and age), we fine-tuned DeBERTa-v3-base (He et al., 2021) using the identical optimizer settings, batch size, and training schedule as in (Meisenbacher and Matthes, 2024).

### E.4 Evaluation Metrics

Perplexity was computed with GPT-2-base (Radford et al., 2019) (`max_length=512`); cosine similarity by averaging sentence embeddings from *all-MiniLM-L6-v2*, *all-mpnet-base-v2* and *thenlper/gte-small* (batch size 32); BLEU (Papineni et al., 2002) via the Hugging Face Evaluate v0.4.3 `bleu` metric; weighted KL divergence over a BERT-base-uncased word-piece tokens with smoothing $\varepsilon = 10^{-10}$; and named-entity retention by comparing original versus anonymized entity counts extracted with spaCy's *en_core_web_sm*.

### E.5 Dataset Splits

The Illinois9 data were split into 3,167 training examples (80%) and 792 test examples (20%). For Author10, we adopted the original 90/10 split (13,562 training examples and 1,507 test examples) as provided by (Meisenbacher and Matthes, 2024).

### E.6 Compute Budget

Style profiling and rewriting with MiniCPM3–4B on NVIDIA T4 GPUs consumed approximately 2.8 GPU-hours for the Author10 dataset and 1.5 GPU-hours for Illinois9. Using LLaMA-3.2-3B on NVIDIA A10G GPUs required about 1.0 GPU-hour on Author10 and 0.3 GPU-hours on Illinois9.

| Method | $\epsilon/k$ | Anonymized Text | Identified? | Weighted KL |
|---|---|---|---|---|
| **Original text** | — | I read an interesting article on human stupidity. It is a pretty much long article, but it is kind of interesting. Read 'The Basic Laws of Stupidity'. It has been a couple of weeks since I posted some laws. So consider this a continuation from where I left off. So urlLink Friday the Thirteenth. And tomorrow is urlLink Saturday the Fourteenth. urlLink Valentine's day. Wonderful. I am looking out for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. BTW, does anyone else use an RSS reader? | Yes | — |
| **DP** | | | | |
| | 25 | Tomorrow is 3:30 am Sunday so Sundays are 2-day Saturday evening… and 3 hours Sunday the 31 February for Friday afternoon for Sundatnl. The New Year's of 2004. I can write blog content on my birthday but haven't posted it and must go to college. This page can… | No | 44.67 |
| | 50 | Yes, the website is on 'the web'. Hopefully it becomes 'blog to the gnters'—'The World's Next Top News site'. If you need directions—either look at our FAQ page. | No | 51.84 |
| | 100 | I am looking for new designs for this blog and urlLink QSS. | No | 48.62 |
| | 150 | I am looking out for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. | Yes | 41.17 |
| | 250 | I am looking for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. | Yes | 41.51 |
| **Quasi-DP** | | | | |
| | 25 | Just two days after we announced something! | Yes | 56.88 |
| | 50 | They have another deadline for posting one or more of your articles. I am ready… | Yes | 55.02 |
| | 100 | I am looking out for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. | No | 41.17 |
| | 150 | I am looking out for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. | No | 41.17 |
| | 250 | I am looking out for new designs for this blog and urlLink QSS. Anyone having any good color schemes, send them to me. | No | 41.17 |
| **Non-DP** | | | | |
| | 3 | I am looking for new designs for the website and urlLink QSS. | No | 49.50 |
| | 5 | 'I'm going to the dentist', 'I am going to be out on a date with a friend.' | Yes | 51.94 |
| | 10 | I am looking for new designs. | No | 52.85 |
| | 25 | I am looking for new designs for this blog and urlLink QSS. | Yes | 48.62 |
| | 50 | On the blog side. Today was urlLink Friday. Tomorrow was urlLink Saturday. Sunday was urlLink Valentine's day. Fantastic. | Yes | 47.50 |
| **Ours** | | | | |
| | Style guided | I recently came across an intriguing piece about human behavior patterns. The lengthy article explores "The Basic Laws of Stupidity," which I found particularly thought-provoking. This post serves as a follow-up to my previous discussion on the topic, now two weeks ago. For those interested in continuing the conversation, you can… | No | 43.54 |
| | Semi-guided | I recently came across an intriguing piece about human behavior and its limitations. The topic spans multiple pages, yet holds my attention due to its relevance. For those interested in exploring similar ideas, I recommend checking out 'The Basic Laws of Stupidity.' This concept has been discussed previously, with a follow-up installment scheduled soon. | Yes | 42.59 |

Table 11: Rewritten texts (long form) from Author10 dataset. Note: For DP-based methods, smaller $\epsilon$ values correspond to tighter privacy guarantees.