Supplementary Material – I Am No One: Style-Aware Paraphrasing for Text Anonymization

Anonymous submission to Interspeech 2026

This supplementary document provides additional material to accompany the main manuscript, including full implementation details, extended utility metrics, robustness checks, and qualitative examples across multiple datasets.

1. Extended Utility Metrics

1.1. Weighted KL Divergence Formulation

Let P and Q denote the normalized term–frequency distributions of the original and anonymized document, respectively. We define the weighted KL divergence as:

$$D_{\text{W-KL}}(P \parallel Q) = \sum_{t \in V} \text{IDF}(t)\, P(t)\, \log \frac{P(t)}{Q(t)}, \qquad (1)$$

where V is the vocabulary of the original corpus and $\text{IDF}(t)$ is the inverse document frequency of token t.

As illustrated in Table 1, we consider a toy example where P corresponds to "quokka quietly grazes", Q_1 to "quokka quietly munches", and Q_2 to "animal quietly grazes", with all KL divergences computed using smoothing $\varepsilon = 10^{-10}$. Weighted KL penalizes anonymizations that drop rare, informative words (e.g., *quokka*) more strongly than generic replacements, even when cosine similarity remains unchanged. Lower values of $D_{\text{W-KL}}$ indicate that the anonymized text better retains the distribution of informative tokens.

Table 1: *Toy Example Comparing KL and Weighted KL Divergence*

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

1.2. LLM-as-Judge for Content Retention

Following prior work using LLMs as evaluators, we use GPT-4 as a *pairwise preference judge* for content retention. For each triplet $\langle x, \hat{x}_{\text{style-guided}}, \hat{x}_{\text{semi-guided}} \rangle$, GPT-4 is prompted to select which anonymized version better preserves the original meaning of x. Table 2 reports the number of instances for which each variant is preferred. GPT-4 prefers the style-guided variant on the full AUTHOR10 test set (847 vs. 649) and even more strongly on the n=121 subset where the semi-guided mode attains lower Author-F1 (105 vs. 16). This suggests that the semi-guided mode's marginal privacy gains are largely driven by more aggressive content distortion.

Table 2: *GPT-4 content-retention preferences on* AUTHOR10.

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

2. Extended Profile Analysis

2.1. Full Results: Single-Dimension Style Profiles

Table 3 reports the complete results for single-dimension style profiles across both datasets, expanding upon the summary provided in the main manuscript.

2.2. Profile Distinctiveness and Saturation Analysis

To understand why some single features still work well, we measure two corpus-level statistics for each feature-specific profile set across different sample sizes ($K \in \{2, 5, 10, 20\}$) on ILLINOIS9.

As shown in Figure 1, *Average TF-IDF* ($\overline{T}$) estimates how much unique lexical information each profile carries. *Length* consistently yields the highest $\overline{T}$. *Average pairwise cosine similarity* ($\overline{C}$) gauges how distinct profiles are across authors (lower values imply stronger discrimination). Again, *Length* is most distinctive. Both metrics 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.

3. Robustness Checks

3.1. Evaluation Against a White-Box Adversary

To probe the limits of our style obfuscation technique, we simulate a *worst-case, method-aware* adversary on the ILLINOIS9 corpus. The attacker is fully aware of our anonymization pipeline and applies the *same* style-profiling and rewriting procedure to the original training texts, producing an attacker-specific training set. A BERT authorship classifier is then fine-tuned on this attacker view.

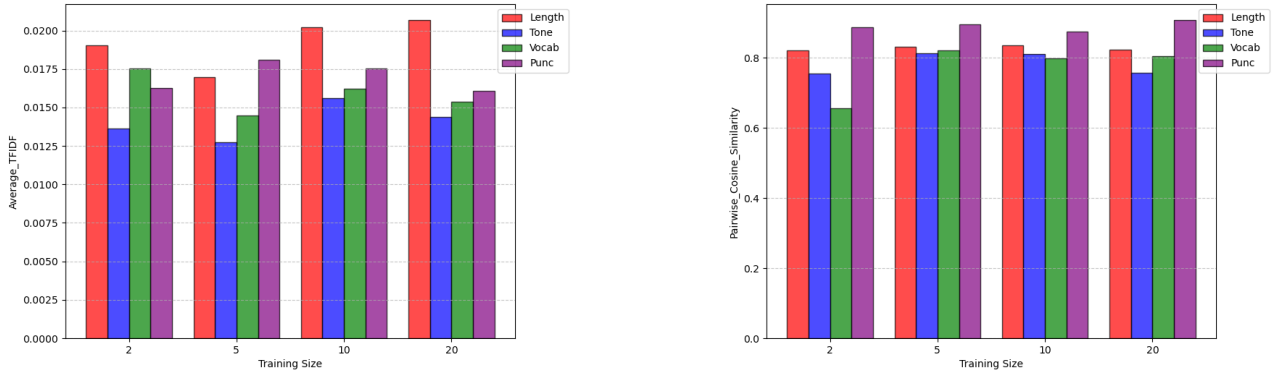
On the original reviews, the classifier achieves macro $F_1 = 77$. In the white-box setting, performance drops to $F_1 = 35$, a 55% reduction despite full method awareness and matched rewriting of the training data. This indicates that our framework substantially weakens stylometric cues even against a strong adversary.

3.2. Robustness Across Additional Domains

To test out-of-domain generalization, we apply our method to two public review datasets: Yelp and IMDB. Table 4 shows that

Table 3: *Performance of Full and Single-Feature Style Profiles on* ILLINOIS9 *and* AUTHOR10

	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 ↓	76.78	20.76	19.32	20.36	21.00	19.24	66.45	26.02	25.85	24.35	24.78	24.77
Gain (γ)	–	0.439	0.461	0.446	0.438	0.463	–	0.310	0.313	0.337	0.331	0.330



(a) Average TF-IDF

(b) Average Pairwise CS

Figure 1: *Lexical uniqueness ($\overline{T}$, left) and inter-author distinctiveness ($\overline{C}$, right) of single-feature style profiles on* ILLINOIS9*, across different profile sizes K.*

our method maintains strong privacy gains with minimal utility loss: authorship F1 drops by more than 85% on both datasets.

Table 4: *Utility and Privacy on Additional Benchmarks*

Dataset	**CS ↑**	**BLEU ↓**	**PPL**		**F1**		γ
			Orig	**Anon**	**Orig**	**Anon**	
Yelp	0.812	0.030	37.87	40.61	95.6	6.90	0.739
IMDB	0.719	0.019	45.56	36.51	98.9	14.0	0.577

3.3. Comparison with STYLEREMIX

Table 5 reports our re-run of STYLEREMIX [1] alongside our best style-guided variant. STYLEREMIX excels at literal content preservation (CS 0.97) but leaves most authorship signal intact (F1 71.5). Our approach trades a moderate amount of semantic similarity for a substantial reduction in attribution accuracy (F1 20.8), confirming its advantage for short, stylistically sparse texts.

4. 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.

Table 5: *Comparison with* STYLEREMIX *on* ILLINOIS9

Method	**CS ↑**	**BLEU ↓**	**PPL ↓**	**F1 ↓**	γ
Original	1.000	1.000	98.83	76.78	–
STYLEREMIX	0.972	0.696	37.37	71.50	0.041
Ours (LLAMA-3.2-3B)	0.709	0.022	53.36	20.76	0.439

4.1. 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 prior work.

4.2. 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-Instruct`) 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.

4.3. 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.

4.4. Authorship and Attribute Classifiers

We trained a BERT-base [2] sequence classifier on the ILLINOIS9 data for three epochs (batch size 32; learning rate 2×10^{-5}; warm-up ratio 0.1; weight decay 10^{-2}; max-norm 1.0). For AUTHOR10, we fine-tuned DeBERTa-v3-base [3] using the identical optimizer settings, batch size, and training schedule as in our DP baselines.

4.5. Compute Budget

With MINICPM3-4B on NVIDIA T4 GPUs, the pipeline 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.

4.6. Prompt Templates

All LLM calls use the same system message: `You are a helpful assistant that responds as instructed.` Figure 2 summarizes the user prompt templates used for (i) style profiling and (ii) style-guided rewriting.

5. Additional Qualitative Examples

This section provides additional qualitative examples that complement the case study in the main paper. Table 6 reports additional short-review examples from ILLINOIS9, Table 8 reports long-form examples from AUTHOR10, and Table 7 reports short-form examples from AUTHOR10. Each block presents the source text alongside anonymized outputs from DP, Quasi-DP, Non-DP, and our style-guided framework.

Style-Profiling Prompt (Full Profile)

```
Below are example texts from a single author.

[AUTHOR_TEXT_1]
...
[AUTHOR_TEXT_K]

Task: Summarize the author's writing style in bullet points.
Focus on:
- Sentence length and structure
- Vocabulary choice
- Tone
- Common punctuation patterns

Style summary:
```

Style-Profiling Prompt (Single-Dimension)

```
Analyze the texts below and summarize only the author's <FOCUS_AREA>.
Provide a concise bullet-point summary (do not quote specific sentences).

Texts:
[AUTHOR_TEXT_1]
...
[AUTHOR_TEXT_K]

Summary (<FOCUS_AREA>):
```

Style-Guided Rewriting Prompt

```
Author style profile:
[STYLE_PROFILE]

Rewrite the text below so that it does NOT reflect the style cues above,
while preserving the original meaning.

Text:
[ORIGINAL_TEXT]

Output (rewrite only; no commentary):
```

Figure 2: *Prompt templates used. We show the full multi-dimensional style-profiling prompt, the single-dimension profiling prompt used in ablations, and the style-guided rewriting prompt that conditions on the extracted profile.*

Table 6: *Example Rewrites from* ILLINOIS9 *(Google Reviews)*

Method	ϵ/k	Anonymized Text	Identified?	W-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
	100	The shops are upscale and offer such a great selection of products.	No	17.59
	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
	10	The upscale shopping area is a great place to visit if you are in the mood for upscale shopping.	No	49.44
	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
	100	The shopping area is spacious and well kept.	No	60.80
	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 7: *Example Short-Form Rewrites from* AUTHOR10

Method	ϵ/k	Anonymized text	Identified?	W-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
	100	Rosie is a complete speaker.	Yes	62.84
	250	Rosie is a good speaker.	Yes	67.93
Quasi-DP				
	25	Sore words	No	71.88
	100	Rosie talks in complete sentences.	Yes	49.35
	250	Rosie is very good at speaking.	Yes	67.97
Non-DP				
	3	I'm glad Rosie talks in complete sentences.	Yes	25.59
	10	Rosie is good at words.	Yes	67.93
	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 8: *Example Long-Form Rewrites from* AUTHOR10

Method	ϵ/k	Anonymized Text	Identified?	W-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
	100	I am looking for new designs for this blog and urlLink QSS.	No	48.62
	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
	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
	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
	10	I am looking for new designs.	No	52.85
	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

6. References

[1] J. Fisher, S. Hallinan, X. Lu, M. L. Gordon, Z. Harchaoui, and Y. Choi, "StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 4172–4206. [Online]. Available: https://aclanthology.org/2024.emnlp-main.241/

[2] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," 2021.