

# StyleRemix: Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements

Jillian Fisher\*<sup>♡</sup> Skyler Hallinan\*<sup>♡</sup> Ximing Lu<sup>♡♣</sup> Mitchell Gordon<sup>♡</sup>  
 Zaid Harchaoui<sup>♡</sup> Yejin Choi<sup>♡♣</sup>  
<sup>♡</sup>University of Washington <sup>♣</sup>Allen Institute for AI  
 {jrfish, hallisky}@uw.edu

## Abstract

Authorship obfuscation, rewriting a text to intentionally obscure the identity of the author, is an important but challenging task. Current methods using large language models (LLMs) lack interpretability and controllability, often ignoring author-specific stylistic features, resulting in less robust performance overall.

To address this, we develop STYLEREMIX, an adaptive and interpretable obfuscation method that perturbs specific, fine-grained *style elements* of the original input text. STYLEREMIX uses pre-trained Low Rank Adaptation (LoRA) modules to rewrite an input specifically along various stylistic axes (e.g., formality and length) while maintaining low computational cost. STYLEREMIX outperforms state-of-the-art baselines and much larger LLMs in a variety of domains as assessed by both automatic and human evaluation.

Additionally, we release AUTHORMIX, a large set of 30K high-quality, long-form texts from a diverse set of 14 authors and 4 domains, and DISC, a parallel corpus of 1,500 texts spanning seven style axes in 16 unique directions<sup>1</sup>.

## 1 Introduction

Authorship obfuscation, the act of rewriting text to conceal the author, is an important method for preserving the privacy of authors in sensitive contexts, e.g., anonymous discussion forums, double-blind reviews, and health services. However, it is inherently complex, requiring a substantial change in writing style to obscure the author’s identity while also preserving the original content and fluency.

Historically, authorship obfuscation methods have manipulated aspects of an author’s style to

\*Co-first authors

<sup>1</sup>We release 1) our code at <https://github.com/jfisher52/StyleRemix> 2) a demo of STYLEREMIX at <https://huggingface.co/spaces/hallisky/StyleRemix> and 3) the datasets (AUTHORMIX and DISC) and trained models in a [HuggingFace collection](#)

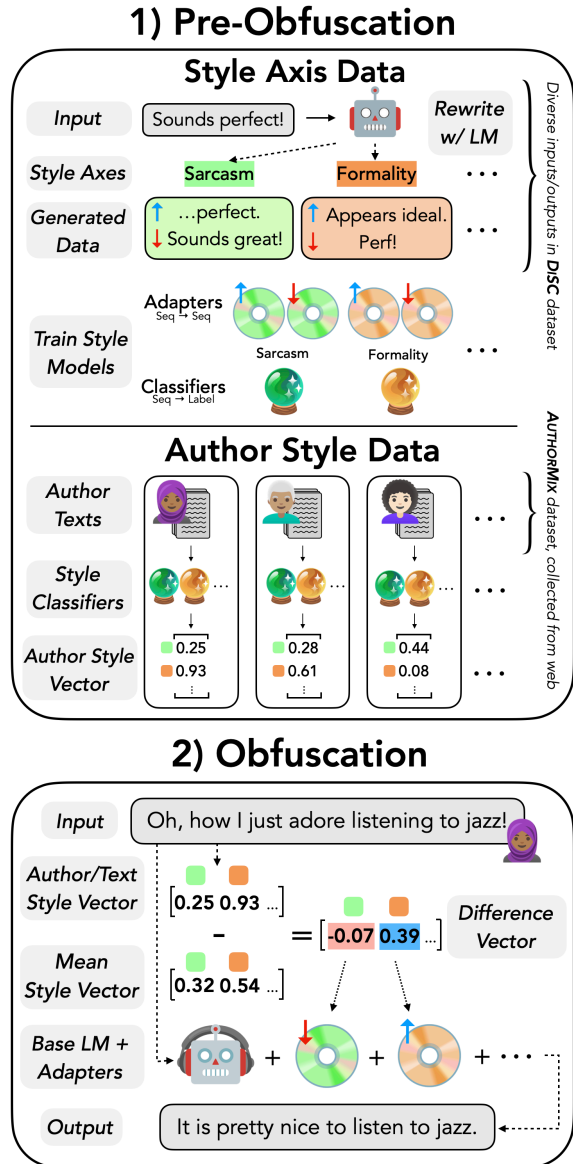



Figure 1: Overview of STYLEREMIX. In pre-obfuscation, distinct *style elements* are distilled from an LM into individual training sets, which are used to train specialized LoRA adapters. During obfuscation, the user can automatically or manually select the style adapter(s) which, when combined with the base LM, will best steer generations away from the original style.

obfuscate the original text (Karadzhov et al., 2017; Shetty et al., 2017; Bevendorff et al., 2019). These techniques typically use style aspects that are easy to automatically evaluate, such as text length, capitalization frequency, and punctuation, to alter the original text. However, these rule-based methods are often too rigid and lead to a degradation of fluency and grammaticality (Fisher et al., 2024).

Recent work demonstrates strong obfuscation performance using LLMs (Mahmood et al., 2019; Haroon et al., 2021; Weggenmann et al., 2022; Fisher et al., 2024), but the common challenge among these is a relative lack of interpretability and controllability on the obfuscation; these approaches do not incorporate any author-specific stylometric characteristics of the original author, leading to more generalized and ineffective obfuscations. For example, a method that relies solely on increasing language model fluency might effectively obfuscate more informal writing, but not formal writing.

To address this gap, we introduce  STYLEMIX, an interpretable, inference-time, author-specific obfuscation method that combines the fluency and steerability of LLMs with author-specific style information. STYLEMIX first detects unique stylistic elements of the author, either through automatic processes or manually. It then uses this information during obfuscation by integrating style-specific adapters with a base language model (an LLM) to guide the generated text away from the author’s original style.

STYLEMIX avoids high computational costs by utilizing pre-trained Low Rank Adaptation modules (LoRA; Hu et al., 2021), which we train to rewrite inputs towards specific directions on various stylistic axes (e.g., more/less length, more/less formality, higher/lower grade level). Drawing inspiration from the process of creating a *remix*, where musical elements of a song, such as tempo, key, and instrumentation are adjusted to form an entirely new track, in this work we seek to identify and manipulate different elements of an authorship style, and propose a simple yet effective approach to steer different components of the text with LoRA adapters. Our results show that STYLEMIX outperforms state-of-the-art authorship obfuscation methods and instruction-based models of similar and larger sizes. Additionally, our method has the added benefit of explainability and is customizable to any unique authorship style.

We make the following contributions:

- (I) We introduce STYLEMIX, an interpretable, inference-time algorithm designed for authorship obfuscation. This method offers the personalization and flexibility required for effectiveness across various styles and text types.
- (II) We release two datasets:
  - (1) AUTHORMIX, a comprehensive authorship dataset with over 30K paragraphs spanning four diverse domains (presidential speeches, novels, scholarly articles, and blogs) and 14 author styles, encompassing many more domains and styles than any previous work to our knowledge.
  - (2) DISTILLED STYLE COMPONENTS DATASET (DISC), a high-quality, validated, parallel dataset over 7 style axes. It features  $n = 1,500$  texts rewritten towards 16 distinct style directions for a total corpus size of 24K.

## 2 Methods

STYLEMIX is an obfuscation method that leverages *style elements* to adaptively rewrite texts. Specifically, it incorporates information about the style of the original author to guide the obfuscation process. Figure 1 illustrates this new approach, which consists of two phases.

The pre-obfuscation phase, conducted only once regardless of the number of authors, involves creating a diverse training set for each style axis we aim to modify (e.g., length variations, formality levels, grade level adjustments, etc.). These style-specific datasets are then used to train Low-Rank Adaptation (LoRA) adapters, which are low-parameter modules that can be seamlessly integrated with a larger base model to guide text generation along specific style axes.

In the obfuscation phase, users can choose the style axes that most effectively disguise the original author’s style, either automatically or manually. The selected pre-trained LoRA adapters are then used to steer the obfuscated text generation.

### 2.1 Stage 1: Pre-Obfuscation

**Style Axes** When selecting the style axis, our goal is to identify “author invariants”, which are text properties that are unique to a specific author. The widely accepted author invariants in the field of stylometry (the study of authorship style) include text length and the use of function words<sup>2</sup>

<sup>2</sup>Function words are words that express grammatical relationships among other words (if, up, would, etc.).

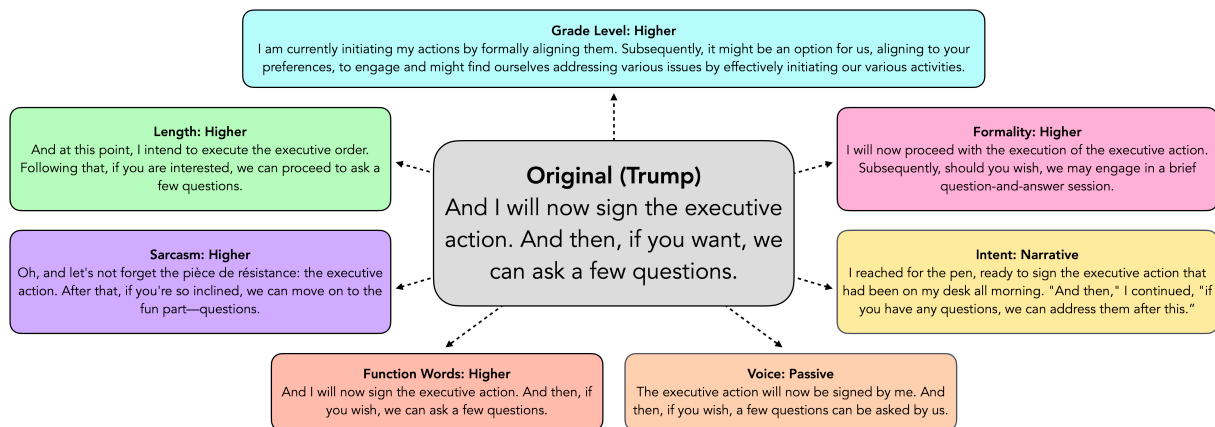


Figure 2: We compare generations from rewriting a text from AUTHORMIX-Speech using each of the style axis adapters *individually*. This demonstrates the distinct transformation capabilities of each adapter, highlighting variations in tone, formality, and other linguistic features. We choose the direction of the style axes based on the automatic style selection method described in Section 2.2.

(Peng and Hengartner, 2002). Additionally, we incorporate "grade level," which primarily measures discrete features like the number of syllables and sentence and word lengths. Since this measure can vary slightly, we averaged three similar metrics: the Flesch-Kincaid (FK; Flesch, 1948), Linsear Write (L; O'Hayre), and the Gunning Fog Index (GF; Gunning, 1952) metrics. For the exact formulas, see Appendix C.1.

Beyond formula-based properties, we also explore more abstract style axes such as the use of sarcasm, formality, voice (passive or active), and writing type (persuasive, descriptive, narrative, and expository). Due to the lack of existing formulas, we train model-based classifiers to measure these properties. More details on the training of these models can be found in Appendix C.1.

In total, we identify seven style axes, each with two directions ("higher" or "lower"), except writing style, which has four options. This results in 16 distinct style elements. We acknowledge that this is not an exhaustive list of all author invariants, but we observed noticeable differentiation among the authors in our experimentation using these metrics. For more details, see Appendix B.5.

**Adapter Training Data** With the above style axes chosen, we create DiSC, a 16-style-element parallel dataset which distills each style element from a large LLM. To standardize the style adapter and minimize content dependencies, we create a single base training set and used instruction prompting with a LLM to generate rewrites along the chosen style axes. The base dataset comprises a diverse range of domains to encompass different writing

types. Specifically, we randomly sample 500 paragraphs from sources including Wikipedia, books and plays (Kryściński et al., 2021), and diary-style blogs (Schler et al., 2006a). Each paragraph is cleaned and standardized, resulting in paragraphs of 2-5 sentences each. Using GPT-4 Turbo (OpenAI, 2023), we then generate new versions of these paragraphs along different style axes and directions ("higher" or "lower") using detailed instruction prompt tuning (see Appendix C.2). This results in 16 parallel datasets written in different style axis and directions.

Next, we evaluate the generated paragraphs to ensure that they accurately reflect the intended style axis and direction. Table 1 presents the evaluation results, both automatic and human, for the style training datasets created. The results demonstrate that our datasets effectively capture the desired styles. See Appendix C.2 for more details.

**Train LoRA Adapters** Next, our goal is to train the models to generate text along the chosen style axes. To minimize computational cost (Strubell et al., 2019), we bypass model fine-tuning and instead employ *Low Rank Adaption* (LoRA; Hu et al., 2021) adapters for each of the style axes. By freezing the base model and tuning only a small portion of injected features, LoRA guarantees lightweight training (Rebuffi et al., 2017; Houlsby et al., 2019) while also incurring *no additional inference latency*, ensuring both efficient training and deployment. We use Llama-3 8B (AI@Meta, 2024) as our base model, and train LoRA adapters on top of them for each direction on the style axes. See Appendix C.3 for more train-

Style Axis ( <i>metric</i> )	Orig.	More	Less
<b>Length</b> ( <i>words/sent</i> )	18.87	<b>23.04</b>	<u>18.24</u>
<b>Function Words</b> ( <i># func. words</i> )	40.08	<b>55.19</b>	<u>21.47</u>
<b>Grade Level</b> ( <i>avg. FK, L, GF</i> )	9.45	<b>11.08</b>	<u>6.72</u>
<b>Formality</b> ( <i>model score</i> )	0.68	<b>0.97</b>	<u>0.43</u>
<b>Accuracy</b> ( <i>human eval</i> )			
<b>Sarcasm</b>	97.7		
<b>Voice</b>	93.7		
<b>Writing Intent</b> ( <i>4 classes</i> )	77.7		

Table 1: Evaluation of the parallel style training datasets. Automatic evaluation (top) is shown for the original score, as well as the score for the dataset that had instruction to increase (More) or decrease (Less) the given style axis. The **highest value** is bolded and the lowest value is underlined. Other style axes required human evaluation (below). For this we randomly combine 10% of the high and low datasets (or all four types for Writing Type) and ask three NLP experts to label whether the style axis was high or low; average accuracy is shown.

ing details.

## 2.2 Stage 2: Obfuscation

**Style Axes and Weights Selection** During the obfuscation phase, a text or set of texts is presented for obfuscation. If a user has a clear idea of which style axes to adjust, they can input their desired styles and the corresponding weights of the adapters to control the strength of the generation. However, since this information is often unavailable, we develop a straightforward yet effective method for selecting which style axes to modify and the magnitude of the weights of these adapters.

For the given  $m$  authors in some genre (e.g. speech, novel), we first create an author vector  $\mathbf{x}_i \in \mathbb{R}^7$  for each author, which is composed of the automatic evaluation of the seven style axes. After normalizing with respect to all  $m$  authors, we calculate the “difference” vector between each author and the average, defined as  $\bar{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j$ . Using the absolute values in this difference vector  $|\bar{\mathbf{x}}_i|$ , users can select the top  $k$  style axes where the specific author deviates most from the average.

Next, the user needs to specify the weight for each chosen style adapter to merge with the base model. This procedure could be manual, but we also provide a heuristic to determine the weights automatically. Building on prior work, we find that LoRA adapters perform well with values in the range  $[-1.5, 1.5]$  (Huang et al., 2023). Next, we use the number of standard deviations an author vector deviates from the average to map each style axis to

a set of predetermined weights  $w_i$ . Specifically,

$$w_i \begin{cases} 0.7 & \text{std}(\bar{x}_i) \leq 1 \\ 0.9 & 1 < \text{std}(\bar{x}_i) \leq 2 \\ 1.2 & 2 < \text{std}(\bar{x}_i) \leq 3 \\ 1.5 & \text{std}(\bar{x}_i) > 3 \end{cases}$$

For detailed implementation, see Appendix C.5.

**Generation Techniques** During generation, we use the adapters corresponding to the selected style axes to rewrite the given text, steering these prominent styles toward the average. In addition, we experiment with multiple methods for combining these LoRA adapters.

- **Sequential:** We pass in the text through a sequence of adapters iteratively; the output from one adapter serves as the input for the next. This method provides additional interpretability by revealing how the text becomes obfuscated at different stages after altering specific style axes. However, it increases computation time, as it requires a forward pass for each chosen style axis.
- **Adapter Merging (AM):** We merge the weights of all the adapters before combining them with the base model by *concatenating* their weights (Yadav et al., 2023a; Yu et al., 2024). See Appendix C.4 for more details.
- **LoraHub<sup>+</sup>:** LoraHub is a framework designed to assemble multiple LoRA adapters with the goal of maximizing performance on specific tasks (Huang et al., 2023). It adjusts the weights of the given adapters to optimize the specified objective through gradient-free optimization. We extend this with LoraHub<sup>+</sup>, which defines a new objective function  $L$  designed to optimize for obfuscation by summing up the automatic evaluations of the selected style axes across a small set of test examples. We also add the fluency score to encourage more fluent text:

$$L = \sum_{v_i \in \text{selected axes}} \begin{cases} v_i & v_i \leq \frac{1}{m} \sum_{j=1}^m x_i \\ 1 - v_i & v_i > \frac{1}{m} \sum_{j=1}^m x_i \end{cases} + \alpha \cdot s_f$$

where  $v_i$  represents the automatic evaluation for a selected style axis on the subset of test examples,  $s_f$  represents fluency score, and  $\alpha$  denotes the discount factor. LoraHub<sup>+</sup> is used *in conjunction* with adapter merging.

Merging multiple adapters with concatenation is computationally efficient. Specifically, we find that



Model	Llama-2-Chat		Llama-3-Inst		Gemma-Inst	Paraphrase	MT	Stylo	JD	STYLEREMIX		
Size	7B	13B	8B	70B	7B					Seq.	AM	AM + LoraHub <sup>+</sup>
AUTHORMIX– Speech												
<b>Drop Rate</b>	18.2	24.0	17.6	16.8	23.1	24.1	10.3	15.1	29.2	34.9	41.2	31.4
<b>Grammar</b>	67.8	67.1	67.1	70.2	67.8	71.2	54.9	37.8	56.7	61.7	66.5	63.9
<b>Content</b>	83.8	80.8	80.8	80.2	78.6	83.9	89.1	89.5	56.4	71.3	77.3	73.9
<b>Overall</b>	10.3	13.0	9.5	9.5	12.3	14.4	5.1	5.1	9.4	<u>15.3</u>	<b>21.2</b>	14.8
AUTHORMIX– Novels												
<b>Drop Rate</b>	12.2	13.7	9.2	11.3	13.3	10.8	7.0	13.5	24.9	19.3	28.6	35.6
<b>Grammar</b>	71.8	73.8	73.1	75.4	70.0	68.3	46.3	36.8	61.2	72.6	68.1	63.5
<b>Content</b>	82.9	80.7	83.1	81.5	81.9	81.3	85.2	88.1	58.6	83.7	76.1	72.9
<b>Overall</b>	7.3	8.2	5.6	6.9	7.6	6.0	2.8	4.4	8.9	11.8	<u>14.8</u>	<b>16.5</b>
AUTHORMIX– Scholar												
<b>Drop Rate</b>	0.8	1.5	1.6	2.5	0.0	0.8	1.5	4.6	6.1	1.8	9.2	11.5
<b>Grammar</b>	64.3	64.9	64.1	66.6	65.3	69.1	54.5	31.0	62.3	65.8	48.6	44.7
<b>Content</b>	91.7	89.7	88.9	84.0	88.9	91.3	92.8	85.8	60.6	78.0	75.3	68.8
<b>Overall</b>	0.5	0.9	0.9	1.4	0.0	0.5	0.8	1.2	2.3	0.9	<u>3.4</u>	<b>3.5</b>
AUTHORMIX– Blog												
<b>Drop Rate</b>	17.7	21.3	21.8	18.9	27.5	22.2	9.4	12.1	56.4	34.4	41.0	42.0
<b>Grammar</b>	68.4	69.1	71.3	74.0	69.0	69.8	41.9	29.1	60.6	66.7	64.9	65.3
<b>Content</b>	82.5	79.0	78.1	77.8	77.8	80.4	83.7	85.8	45.1	72.1	73.7	74.2
<b>Overall</b>	10.0	11.6	12.1	10.9	14.8	12.5	3.3	3.0	15.4	16.5	<u>19.6</u>	<b>20.4</b>

Table 2: Comparison of obfuscation methods measured by mean drop rate, grammar, meaning similarity, and overall (the mean product of the metrics), across STYLEREMIX and comparatively sized or larger baselines on each subset of AUTHORMIX. **Bold** and underline denote the highest and the second-highest score respectively in each row. All metrics displayed in the table are multiplied by 100 for easier viewing of significant figures.

merging four adapters with the base model (using AM) takes less than 5 seconds on average.

### 3 Experiments

#### 3.1 Datasets

We aimed to test how authorship obfuscation methods perform on a diverse array of author styles and domains. To this end, we develop a new benchmark dataset called AUTHORMIX, covering four distinct domains: presidential speeches, early-1900s fiction novels, scholarly articles, and diary-style blogs. Together, AUTHORMIX contains more than 30 k high-quality paragraphs from 14 authors.

For the presidential domain, we curate and clean speeches from George W. Bush, Barack Obama, and Donald Trump<sup>3</sup>. For novel domain, we choose a collection of early 1900s fiction writers with strong writing styles: Ernest Hemingway, F. Scott Fitzgerald, and Virginia Woolf. We choose these specific writers in an effort to limit the topic bias in the evaluation metrics.

Lastly, we alter two existing datasets to match the formality of our new domains: the Extended-Brennan Greenstad (Brennan et al., 2012), a collection of “scholarly” short (500-word) paragraphs

<sup>3</sup>We select these presidents/authors due to their diverse styles but similar eras to minimize content discrepancies.

gathered from Amazon Mechanical Turk (AMT), and the Blog Authorship corpus (Schler et al., 2006b), a collection of blogs (diary-style entries) that were posted to blog.com. More details can be found in Appendix D.3.

#### 3.2 STYLEREMIX Configurations

We compare three versions of STYLEREMIX: sequential, adapter merging, and LoraHub<sup>+</sup>. For sequential, to account for the order of the styles, we average over  $n = 3$  shuffled orders. The adapter merging method uses the static standard deviation mapping method described in Section 2.2. For these two methods, we select the best method per domain (based on the overall score) using the top  $k = 1, 2, 3, 4$  changed styles. Lastly, we run our customized LoraHub method (LoraHub<sup>+</sup>), matching the best styles per domain as the base adapter merging method for direct comparison.

#### 3.3 Baselines

We compare against both SOTA obfuscation methods and equal and larger-size LLMs using instructions. Full details can be found in Appendix D.

**Stylometric (Stylo)** We use the stylometric obfuscation technique presented by Karadzhov et al. (2017), which examines various statistical features

that characterize a writer’s style, such as sentence length and word frequency, and then modifies the text to align these features with an "average" value, which is established using a training dataset.

**Machine Translation (MT)** Keswani et al. (2016) introduce *round-trip machine translation* by translating a text from English to German, then to French, and then back to English. We use the new M2M translation models (Fan et al., 2020).

**Paraphraser (Paraphrase)** We use the T5-Large paraphraser introduced by Jung et al. (2024) which iteratively improves through self-distillation.

**JAMDEC (JD)** This method (Fisher et al., 2024) relies on a smaller LLM, GPT2-XL (Radford et al., 2019) to overgenerate many new rewrites given the keywords from the original text. It then uses a filter to select the best new rewrite. We run this method using the default settings, and a beam width of 10.

**Instruction-tuned LLMs** We compare against a suite of instruction-tuned LLMs including Llama-2-Chat (7B, 13B) (Touvron et al., 2023), Llama-3-Instruct (8B, 70B) (AI@Meta, 2024), and Gemma-Instruct (7B) (Team et al., 2024). For each model, we provide instruction to “rewrite” the given text. More comparisons of different models can be found in the Appendix B.2. Exact instructions used for generation can be found in Appendix D.

### 3.4 Automatic Evaluations

In line with previous work, we evaluate authorship obfuscation on four main criteria: obfuscation, content preservation, grammaticality, and overall<sup>4</sup>. See Appendix D.5 for more details.

**Obfuscation** Classifiers with various machine learning architectures have been used to measure obfuscation (Mahmood et al., 2019; Haroon et al., 2021; Fisher et al., 2024). Recent work in authorship obfuscation and style transfer often uses RoBERTa (Liu et al., 2019) classifiers (Xing et al., 2024a; Uchendu et al., 2021; Liu and May, 2024; Hallinan et al., 2023).

In line with this previous work, we fine-tune four RoBERTa large (355M) models, one for each domain in AUTHORMIX; classifiers achieve on average 94.0% accuracy in the evaluation set of

<sup>4</sup>All metrics are bounded between 0 and 1, which ensures the product has the same bounds. Although similarity scores are theoretically bounded from -1 to 1, we observe empirically across all datasets and methods that they are bounded more strictly between 0 and 1; see Appendix D.5 for more details

each respective domain (for full results, see Appendix D.5). Using these classifiers, we calculate the drop rate, which is the normalized decrease in the classifier’s accuracy when comparing its performance on the original texts to the obfuscated texts. The drop rate can be expressed as:

$$\text{Drop Rate} = \frac{acc_{\text{orig}} - acc_{\text{obf}}}{acc_{\text{orig}}}$$

We also report evaluation using an alternative metric to measure obfuscation based on universal authorship representations (Soto et al., 2021) in Appendix F; these results corroborate our main findings in §3.6.

**Content Preservation** We use the embedding similarity of the inputs and their obfuscations in Sentence Transformers (Reimers and Gurevych, 2019) to gauge semantic similarity.

**Language Quality** We measure grammaticality via the probability of being grammatically acceptable from TextAttack (Morris et al., 2020), a binary RoBERTa-large classifier fine-tuned on the Corpus of Linguistic Acceptability (Warstadt et al., 2018).

**Overall Task Score** The overall success of each obfuscation is measured by the product of the above three metrics: drop rate, similarity score, and CoLA score. This product ensures a high overall task score accurately reflects high scores in all three categories; it is also used in prior work in text rewriting (Krishna et al., 2020a; Hallinan et al., 2023; Patel et al., 2023; Xu et al., 2018)

### 3.5 Human Evaluation

We also conduct human evaluation to verify the quality of the obfuscations from the best STYLEMIX variant and comparably sized baselines; we omit . We randomly select  $n = 20$  texts from each author in AUTHORMIX for annotation via Amazon Mechanical Turk by three workers each. Following the setup of Fisher et al. (2024), we instruct each annotator to read both the original and obfuscated text, then respond to five questions rated on a three-point Likert scale (0, 0.5, or 1), measuring grammar, fluency, high content preservation, low content addition, and obfuscation. We discard evaluations where all annotators disagree on the label<sup>5</sup>. Lastly, we calculate an *overall* score using the weighted product of all five metrics. Further details can be found in Appendix E.

<sup>5</sup>Pairwise agreement is greater than 93% for all questions

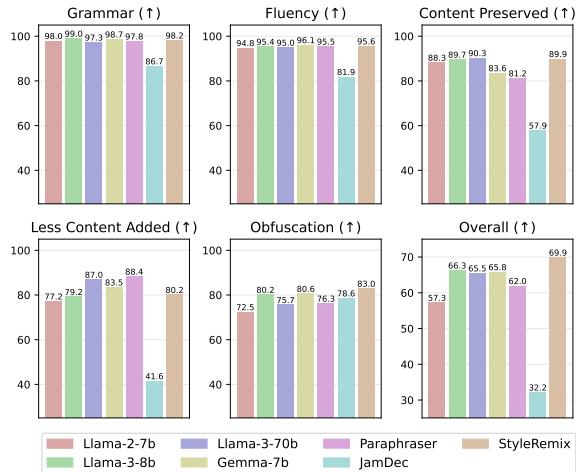


Figure 3: Human evaluation results for mean grammar, fluency, content preserved, less content added, and obfuscation. For each of the metrics, *higher* is better. We also compute the mean overall score, the product of grammar, content preserved, and less style similarity.

### 3.6 Main Results

**STYLEREMIX has the highest overall task score and obfuscation drop rate** Table 2 compares STYLEREMIX to LLMs of all sizes and the four SOTA authorship obfuscation methods on obfuscation, content preservation, grammar, and overall task score. STYLEREMIX has on average 5.5% higher overall task score and up to 12% higher obfuscation drop rate on all domains, with similar scores on content preservation and grammar. Impressively, STYLEREMIX is also more effective at obfuscation than LLMs of much larger magnitudes – such as Llama-2-Chat 13B and Llama-3-Instruct 70B – especially on the Blog domain with a  $\sim 10\%$  higher overall task score. This shows how a more targeted approach can outperform larger but more generic methods.

#### Human evaluation shows STYLEREMIX outputs the best overall and most obfuscated generations

Figure 3 shows the result of our human evaluation. All methods have strong grammar and fluency scores, with only slight differences between them. STYLEREMIX beats all other baselines in obfuscation with a score of 83.0%, and ranks second for content preservation, trailing by just 0.4 points behind the significantly larger Llama-3-Instruct 70B. Notably, STYLEREMIX even surpasses the larger Llama-3-Instruct 70B in obfuscation, content preservation, and grammar. In terms of less content added, STYLEREMIX still outperforms strong baselines like Llama-3-Instruct 8B, but loses slightly to

Method	Generation
<b>Original (Blog 3)</b>	100%! I was surprised, but not complaining lol. But yeah Mr. Curphey's now on crutches, hobbling around the classroom and still teaching us like his usual self.
MT	100 per cent! I was surprised, but I don't complain lol. But yes, Mr. Curphey is now in the <b>gardens to walk</b> through the classrooms and still teach us as his ordinary self.
Stylo.	<b>One hundred percent</b> ; i was surprised, but not complaining ll, as but yeah <b>Mr. murphy</b> 's now on crutches, <b>sobbing</b> around the classroom and always teaching us like his usual self.
JD	100%! I was surprised, but not complaining lol. But yeah Mr. Curphey's now on crutches, hobbling around the classroom and still teaching us like his usual self.
Para.	Mr. Curphey is currently using crutches in the classroom, but he is still teaching us with his usual enthusiasm <b>and enthusiasm</b> . I was taken aback, but I didn't feel <b>any remorse</b> .
Llama-2 Chat (7B)	100%! I was <b>pleasantly</b> surprised, but not complaining. However, Mr. Curphey is now using crutches to move around the classroom, still actively teaching us with his usual dedication.
Llama-3 Inst (8B)	100%! I was taken aback, but I wasn't about to grumble about it, haha! But yes, Mr. Curphey's now on crutches, limping around the classroom, and yet he's still teaching us as himself, business as usual.
Gemma Inst (7B)	100%! I was surprised, but not in a bad way. Mr. Curphey is still teaching us despite being on crutches.
STYLE REMIX	Initially, I experienced a <b>notable degree of surprise</b> , yet my responses remain <b>devoid of significant complaint</b> . Currently, Mr. Curphey, significantly, is navigating our classroom utilizing crutches, <b>effectively maintaining</b> his usual style of instruction, significantly.
<b>Original (Obama)</b>	I recognize that times have changed since I first spoke to this convention. The times have changed, and so have I. I'm no longer just a candidate. I'm the President.
MT	I recognize that the times have changed since I first spoke to this convention. The times have changed, and <b>that's how I have</b> . I'm not just a candidate anymore. I am the President.
Stylo.	I acknowledge that times have changed since I <b>1st</b> spoke to this convention, and the times have changed, and so have I. I am no longer just a candidate; and myself <b>'m</b> the President.
JD	It's important to remember where we came from, where we're going, and what we're trying to do," he said, adding, "I think it's important for all of us to be. We are the people who built this country, and we have a responsibility to make sure that we preserve it for our children and our children's children, and that's what we're going to do."
Para.	Since I first addressed this convention, <b>I have seen significant changes. As a result</b> , I have evolved from a candidate to the current President.
Llama-2 Chat (7B)	Since my initial conversation with this convention, time has passed and changes have occurred. As a result, <b>my perspective and circumstances have evolved</b> . I am no longer just a candidate, but the actual President.
Llama-3 Inst (8B)	Since my initial address to this esteemed convention, I have come to acknowledge that the times have indeed changed. So too have I. Gone are the days when I stood before you as a mere candidate, eager to make my case. Today, I stand before you as the President, <b>with a wealth of experience and insight gathered during my time in office</b> .
Gemma Inst (7B)	Since my initial address to this esteemed convention, the world has shifted, and I have evolved alongside it. Today, I stand before you not as a candidate, but as the President, <b>ready to tackle the challenges that lie ahead</b> .
STYLE REMIX	<b>Oh, how the world has transformed</b> since I first addressed you all here! <b>Indeed, the world has shifted</b> , and so have I. It's not just about being a candidate anymore—it's about <b>being the President</b> .

Table 3: Examples of obfuscations from baseline methods and STYLEREMIX from texts from AUTHORMIX-Blog and AUTHORMIX-Speech respectively. **Green** regions highlight where the method obfuscates well and **red** areas show grammar, content, or obfuscation issues.

methods like Gemma-Instruct 7B and Paraphrase; though these methods may be good abstaining from adding new content, this is likely a byproduct of their generations being too succinct and failing to preserve information or obfuscate, as shown by their low human evaluations on these two metrics.

For overall score, which captures aspects of fluency, content preservation, and obfuscation, STYLEREMIX performs the best, achieving an overall score of 69.9%; the next highest scoring method is Llama-3-Instruct 8B with a score of 66.3%, a significant dropoff. Each individual metric must be high to achieve a high product; this indicates that our method on averages produces the obfuscations

with the best overall quality, balancing between grammar, content preservation, and obfuscation, rather than optimizing for just one dimension.

### Qualitatively, STYLEREMIX generates more flexible, directed obfuscations compared to other methods

Qualitative results demonstrate that, as designed, STYLEREMIX provides a strong, personalized obfuscation compared to the more general results of other methods and models. Table 3 presents two randomly<sup>6</sup> selected texts along with the generations from various models and methods. Consistent with previous work (Fisher et al., 2024), the rule-based methods (MT and Stylo result in poor grammar or loss of content. Conversely, methods based on LLMs tend to maintain grammar and content preservation more effectively.

The most significant difference is evident in the *style* of the generated text. Other methods sometimes struggle to clearly obfuscate and instead generally mimic the original author’s style or default to a more formal “model”-like writing style. In contrast, STYLEREMIX stands out by providing a more personalized and targeted obfuscation. For instance, in the Blog example (top), STYLEREMIX generates text that is more formal, uses higher-grade level language, and is longer compared to the original text. Meanwhile, in the Speech example (bottom), it adopts a more sarcastic, less formal tone, and incorporates more function words.

We also find that this multi-style mixture approach often results in noticeably different sentence structures and punctuation. For example, in the speech text (bottom), the order of the first sentence is reversed compared to the original, a feature not observed in any other generation. Additional generations are available in Appendix B.6.

To further highlight the steerability of STYLEREMIX, we display a randomly selected text from AUTHORMIX-Speech and random generations created using an adapter in the optimal steering direction<sup>7</sup> for each of the seven style axes in Figure 2. Each generation demonstrates how the choice of adapter significantly transforms the text and influences the type of obfuscation.

## 3.7 Ablations and Other Studies

### Our automatic method of style selection results in better obfuscation than random selection

<sup>6</sup>Texts were filtered by a length threshold.

<sup>7</sup>The optimal direction is calculated based on the automatic style selection method in Section 2.2

# Styles	1	2	3	4	5	6	7
Speeches	17.0	17.7	<b>21.2</b>	19.2	6.0	17.0	11.4
Novels	8.6	11.2	13.0	14.4	16.3	11.2	<b>21.8</b>
Scholar	1.1	1.8	2.3	3.4	0.8	6.0	<b>16.9</b>
Blog	13.1	16.5	<b>19.6</b>	18.9	12.1	10.5	6.4

Table 4: Overall task score on the base adapter merging method using different number of style adapters. We compare the overall task score using 1 – 7 style adapters. For all automatic evaluations see Table 6

Although STYLEREMIX can be used with any arbitrary method of choosing the style axes to change, we do find that choosing based on difference between the average style vector and the author vector improves obfuscation on average by 6% over random selection of the same number of weights. We note that the grammar and content remained about equal. More details can be found in Appendix B.1

### Shuffling style adapters when using STYLEREMIX-Sequential leads to some variation

For STYLEREMIX-Sequential we experiment with shuffling the order of the chosen style adapters over  $n = 3$  random shuffling. We found that the order of the styles does have some effect on the obfuscation drop rate (standard deviation of 3% – 6%) but little effect on the grammar or content preservation (standard deviation of 1% – 2%). This was seen strongly when choosing 3+ styles and in domains with strong style differences among the authors (Speech and Blog). More details are in Appendix B.3

### Changing 5+ style axes decreases grammaticality

Table 4 shows how the overall task score changes the number of styles chosen to use the adapter merging method increases. At first, both obfuscation drop rate and overall score steadily increase as we increase the number of style adapters, which corresponds with changing more elements of the original text. However, for AUTHORMIX Speech, Scholar, and Blog, we see a sudden decrease in overall task score when using 5 styles. Investigating this, we found that using 5+ style adapters leads to an average of  $\sim 16\%$  decrease in grammar and a  $\sim 5\%$  decrease in overall score. More details can be found in Appendix B.4.

## 4 Related Work

**Authorship Obfuscation Methods** Traditional authorship obfuscation methods leverage stylistic insights, such as author invariant features,



to obfuscate texts (Karadzhov et al., 2017; Mansoorizadeh et al., 2016; Xing et al., 2024b). However, these methods have been shown to have issues with grammar and fluency due to their strict rule-based approach (Fisher et al., 2024).

To reduce this behavior, model-based approaches have been developed, such as Mutant-X, a genetic algorithm which utilizes an internal classifier to iteratively "mutate" a sentence (Mahmood et al., 2019). Later work improves on this with an *ensemble* of classifiers rather than a single one (Haroon et al., 2021) or via variational autoencoders as the base model to generate differentially private generations (Weggenmann et al., 2022). Most recently, Fisher et al. (2024) demonstrate the efficacy of smaller LLMs for authorship obfuscation through over-generation and filtering. However, this method's reliance on a heavy decoding algorithm to generate diverse candidates makes it impractical. Some obfuscation methods have also incorporated authorship information (Jones et al., 2022; Shetty et al., 2017). Although these both showed promise, they require extensive training and are only applicable in specific use cases.

**Parameter Efficient Learning** Parameter-efficient adapters, small modules tuned on top of a frozen large model for effective transfer learning, have been proposed for vision (Rebuffi et al., 2017) and NLP (Houlsby et al., 2019). Others have extended these methods by tuning specific layers and embeddings (Li and Liang, 2021; Lester et al., 2021), or by making the adapter matrices an addition to the original model weights themselves rather than additional injected layers (Hu et al., 2021; Lu et al., 2023).

Adjacent to parameter-efficient training strategies are **model merging** techniques, which seek to integrate model knowledge by combining their weights (Matena and Raffel, 2021); this is efficient and prevents additional inference cost. Merging has been explored extensively in previous work, to combine diverse, targeted domain models (Jang et al., 2023; Ramé et al., 2023), or over the same model trained with different seeds or hyperparameters to improve robustness (Wortsman et al., 2022; Ramé et al., 2022). Model merging has even been explored with parameter-efficient adapters like LoRA (Huang et al., 2023). Other lines of work expand on merging techniques, creating strategies beyond simply averaging model weights. (Yadav et al., 2023b; Stoica et al., 2023; Yu et al., 2023).

**Controllable Generation** Previous work introduces methods to control the content of a generation (Lu et al., 2021) or steer the style of the generation (Liu et al., 2021; Lu et al., 2023). However these types of controllable generation are less practical for authorship obfuscation, which requires a steerability of the content and the style.

**Style Transfer** Style transfer techniques have utilized both simple models (Hu et al., 2022) and more advanced machine learning models (Jin et al., 2022; Hallinan et al., 2023). Most approaches depend on training generation models with a dataset, which can be natural or synthetic (Jin et al., 2022). These methods also interpret "style" in various ways, ranging from comprehensive notions like specific authors (e.g., Shakespeare and Hemingway) (Krishna et al., 2020b) to particular stylistic elements (e.g., formality and sentiment) (Fu et al., 2018). Notably, recent, effective style transfer techniques resemble the approach we use, which involves fine-tuning a LLM (Hallinan et al., 2023; Krishna et al., 2020b).

## 5 Conclusion

In this work, we introduce `STYLEMIX`, a novel and interpretable method for authorship obfuscation. By targeting specific fine-grained stylistic elements and leveraging Low Rank Adaptation (LoRA) modules, `STYLEMIX` provides a more interpretable and controllable approach than existing methods based on large language models or other state-of-the-art techniques while still maintaining excellent performance. We show our new method outperforms a suite of strong, state-of-the-art baselines in four diverse domains overall in both automatic and human evaluation.

Additionally, as part of this work, we release a new authorship attribution dataset, `AUTHORMIX`. This dataset includes two new domains: presidential speeches and fiction books, which were carefully selected to ensure a high degree of topic matching, thereby enriching the dataset's applicability and depth. Furthermore, we develop `DISC`, a collection of 16 parallel, human-validated datasets spanning various stylistic dimensions, which can be employed in future research to further explore and refine the nuances of stylistic text manipulation. These resources aim to advance the field of authorship obfuscation and provide a solid foundation for subsequent studies.

## 6 Limitations and Ethical Considerations

One limitation of STYLEREMIX is the requirement of needing trained LoRA adapters and the corresponding style datasets for their training. This necessitates an additional pre-obfuscation step involving separate style corpi and computational training time. However, this is a one-time expense, and the same style adapters can be utilized for multiple authors. In return, users benefit from a more interpretable method for authorship obfuscation.

During obfuscation, STYLEREMIX does require a slightly higher computational time and memory due to the extra style LoRA adapters compared to using a finetuned model with instructions. For the sequential version of STYLEREMIX, the computational time is multiplied by the number of styles. The base adapter merging variation outperforms the sequential version and is also more efficient: the time is only increased by a small amount from merging adapters then adding them to the base model, rather than requiring multiple model forward-passes. In inference, no extra time is added, since LoRA weights are seamlessly added to the original model (Hu et al., 2021); see Appendix C.4 for further analysis. Finally, we note that for the adapter merging with LoraHub, there is also additional time over the base adapter merging to identify the optimal weights.

An additional limitation is that STYLEREMIX is developed and tested only on English. However, we believe that the framework (identifying style axes, training LoRA adapters, evaluating original text on these axes, and perturbing based on these evaluations) could be generalized to any language. As such, we believe our pipeline can be effectively adapted to obfuscate other languages; we leave the exploration of these adaptations to future work.

Lastly, our work also has some potential risks. Though the intention of authorship obfuscation is to protect identities in sensitive situations, there is a possibility that malicious users could misuse our method. We acknowledge this as a potential risk for any authorship obfuscation method, which is inherent when creating these methods.

## 7 Acknowledgements

This research was supported by the DARPA MCS program through NIWC Pacific (N66001-19-2-4031), IARPA 2022-22072200003, and NSF IFML CCF-2019844.

## References

- AI@Meta. 2024. [Llama 3 model card](#).
- Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.
- Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. [Heuristic authorship obfuscation](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. [Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity](#). *ACM Transactions on Information and System Security (TISSEC)*, 15.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.
- Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024. [JAMDEC: Unsupervised authorship obfuscation using constrained decoding over small 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 1552–1581, Mexico City, Mexico. Association for Computational Linguistics.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. [Style transfer in text: Exploration and evaluation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Robert Gunning. 1952. The technique of clear writing. (*No Title*).
- Project Gutenberg. [\[link\]](#).
- Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. [Steer: Unified style transfer with expert reinforcement](#). *ArXiv*, abs/2311.07167.
- Muhammad Haroon, Muhammad Fareed Zaffar, Padmini Srinivasan, and Zubair Shafiq. 2021. [Avengers ensemble! improving transferability of authorship obfuscation](#). *CoRR*, abs/2109.07028.

- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for nlp](#). *ArXiv*, abs/1902.00751.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *ArXiv*, abs/2106.09685.
- Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. [Text style transfer: A review and experimental evaluation](#). *SIGKDD Explor. Newsl.*, 24(1):14–45.
- Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. [Lorahub: Efficient cross-task generalization via dynamic lora composition](#). *ArXiv*, abs/2307.13269.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. [Personalized soups: Personalized large language model alignment via post-hoc parameter merging](#). *ArXiv*, abs/2310.11564.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. [Deep learning for text style transfer: A survey](#). *Computational Linguistics*, 48(1):155–205.
- Keenan Jones, Jason R. C. Nurse, and Shujun Li. 2022. [Are you robert or roberta? deceiving online authorship attribution models using neural text generators](#). *Preprint*, arXiv:2203.09813.
- Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2024. [Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing](#). *Preprint*, arXiv:2305.16635.
- Georgi Karadzhov, Tsvetomila Mihaylova, Yassen Kiprova, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. The case for being average: A mediocrity approach to style masking and author obfuscation. *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–185.
- Yashwant Keswani, H. Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. In *Conference and Labs of the Evaluation Forum*.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020a. [Reformulating unsupervised style transfer as paraphrase generation](#). *ArXiv*, abs/2010.05700.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020b. [Reformulating unsupervised style transfer as paraphrase generation](#). *Preprint*, arXiv:2010.05700.
- Wojciech Kry  ci  ski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. [Booksum: A collection of datasets for long-form narrative summarization](#).
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). *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)*, abs/2101.00190.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. [DExperts: Decoding-time controlled text generation with experts and anti-experts](#). 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 6691–6706, Online. Association for Computational Linguistics.
- Jialin Liu, Antoine Moreau, Mike Preuss, Baptiste Roziere, Jeremy Rapin, Fabien Teytaud, and Olivier Teytaud. 2020. [Versatile black-box optimization](#). *Preprint*, arXiv:2004.14014.
- Shuai Liu and Jonathan May. 2024. [Style transfer with multi-iteration preference optimization](#). *ArXiv*, abs/2406.11581.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *ArXiv*, abs/1907.11692.
- Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Raghavi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian R. Fisher, Bill Yuchen Lin, Skyler Hallinan, Xiang Ren, Sean Welleck, and Yejin Choi. 2023. [Inference-time policy adapters \(ipa\): Tailoring extreme-scale lms without fine-tuning](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Neuro-Logic decoding: \(un\)supervised neural text generation with predicate logic constraints](#). In *Proceedings*



- of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.
- Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. [A girl has no name: Automated authorship obfuscation using mutant-x](#). *Proceedings on Privacy Enhancing Technologies*, 2019:54 – 71.
- Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminian, and Mehdy Eskandari. 2016. Author obfuscation using wordnet and language models. In *Conference and Labs of the Evaluation Forum*.
- Michael Matena and Colin Raffel. 2021. [Merging models with fisher-weighted averaging](#). *ArXiv*, abs/2111.09832.
- Tsvetomila Mihaylova, Georgi Karadzhov, Preslav Nakov, Yassen Kiprova, Georgi Georgiev, and Ivan Koychev. 2016. Su@ pan’2016: Author obfuscation—notebook for pan at clef 2016.
- University of Virginia Miller Center of Public Affairs. Presidential speeches: Downloadable data.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.
- John O’Hayre. *Gobbledygook Has Gotta Go*.
- OpenAI. 2023. [Gpt-4 turbo](#). Accessed: 2024-06-05.
- Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2023. [Low-resource authorship style transfer: Can non-famous authors be imitated?](#) *Preprint*, arXiv:2212.08986.
- Roger Peng and Nicolas Hengartner. 2002. [Quantitative analysis of literary style](#). *Roger D. Peng*, 56.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. [Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards](#). *ArXiv*, abs/2306.04488.
- Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. 2022. [Diverse weight averaging for out-of-distribution generalization](#). *ArXiv*, abs/2205.09739.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. [Learning multiple visual domains with residual adapters](#). *ArXiv*, abs/1705.08045.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006a. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006b. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.
- Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2017. A4nt: Author attribute anonymity by adversarial training of neural machine translation. In *USENIX Security Symposium*.
- Rafael A. Rivera Soto, Olivia Elizabeth Miano, Juanita Ordoñez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. [Learning universal authorship representations](#). In *Conference on Empirical Methods in Natural Language Processing*.
- George Stoica, Daniel Bolya, Jakob Bue Bjorner, Taylor N. Hearn, and Judy Hoffman. 2023. [Zipit! merging models from different tasks without training](#). *ArXiv*, abs/2305.03053.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and policy considerations for deep learning in nlp](#). *ArXiv*, abs/1906.02243.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon,



- Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. [Turingbench: A benchmark environment for turing test in the age of neural text generation](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- 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*, page 721–731, New York, NY, USA. Association for Computing Machinery.
- Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. [Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time](#). *ArXiv*, abs/2203.05482.
- Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024a. [Alison: Fast and effective stylistic authorship obfuscation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19315–19322.
- Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024b. [Alison: Fast and effective stylistic authorship obfuscation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:19315–19322.
- Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. [Unpaired sentiment-to-sentiment translation: A cyclic reinforcement learning approach](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023a. [Ties-merging: Resolving interference when merging models](#). *Preprint*, arXiv:2306.01708.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023b. [Ties-merging: Resolving interference when merging models](#). In *Neural Information Processing Systems*.
- Le Yu, Yu Bowen, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. [Language models are super mario: Absorbing abilities from homologous models as a free lunch](#). *ArXiv*, abs/2311.03099.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. [Language models are super mario: Absorbing abilities from homologous models as a free lunch](#). *Preprint*, arXiv:2311.03099.

## A Code and Artifacts

The code for STYLEREMIX is available on Github at <https://github.com/jfisher52/StyleRemix>. In addition, we release a demo of our method at <https://huggingface.co/spaces/hallisky/StyleRemix>, which allows for arbitrary inputs and rewrites with multiple LoRA adapters across multiple style axes.

Finally, we release the rest of our datasets and trained classifiers and LoRA adapters in a HuggingFace collection. Notably, the collection includes AUTHORMIX at <https://huggingface.co/datasets/hallisky/AuthorMix> and DiSC at <https://huggingface.co/datasets/hallisky/DiSC>.

## B Extended Ablations and Other Studies

### B.1 Random Selection of Styles

In Section 2.2, we describe a simple automatic method to select the style axes to change for each author. It requires creating an author vector, which is composed of the ten style axes automatic evaluations, and finding the difference for each author compared to the average vector of all authors in a domain. In order to test the efficacy of our style axes selection method, we compare the results of STYLEREMIX when selecting the styles axes in this way and randomly (over  $n = 3$  different seeds).

Figure 4 shows the average and standard deviation of the drop rate, grammar score, content preservation score and overall task score for each domain randomly choosing 1 – 4 styles (circles) and using our automatic method of style axes selection (stars). First, we notice that overall, the grammar and content preservation is mostly similar for both random and the automatic method. However, we do see a large difference in obfuscation drop rate, especially in speech (18% average) and Scholar (8 average). These datasets have more modern, similar styles, which might need a more targeted obfuscation rather than the novels (which are written in older English) and the blog (which are very informal).

### B.2 Comparing with Different LLMs

For the main experiment we showed the comparison with different like-sized LLMs. Here we provide more comparisons with Mistral V2 (Jiang et al., 2023) and Gemma (2B) (Team et al., 2024) to the three variations of STYLEREMIX. We show results for all three criteria as well as the overall task

	Mistral	Gemma	STYLEREMIX		
	V2	2B	Seq.	AM	AM + LoraHub <sup>+</sup>
<b>AUTHORMIX-Speech</b>					
<b>Drop Rate</b>	25.8	24.7	34.9	41.2	31.4
<b>Grammar</b>	67.6	70.6	61.7	66.5	63.9
<b>Content</b>	81.0	78.2	71.3	77.3	73.9
<b>Overall</b>	14.1	13.6	15.3	21.2	14.8
<b>AUTHORMIX-Novels</b>					
<b>Drop Rate</b>	12.0	13.5	19.3	28.6	35.6
<b>Grammar</b>	69.7	72.2	72.6	68.1	63.5
<b>Content</b>	80.1	78.2	83.7	76.1	72.9
<b>Overall</b>	6.7	7.6	11.8	14.8	16.5
<b>AUTHORMIX-Scholar</b>					
<b>Drop Rate</b>	0.8	1.5	1.8	9.2	11.5
<b>Grammar</b>	66.8	69.5	65.8	48.6	44.7
<b>Content</b>	88.9	87.3	78.0	75.3	68.8
<b>Overall</b>	2.3	2.8	3.6	3.4	3.5
<b>AUTHORMIX-Blog</b>					
<b>Drop Rate</b>	23.7	21.9	34.4	41.0	42.0
<b>Grammar</b>	68.3	71.3	66.7	64.9	65.3
<b>Content</b>	78.3	77.1	72.1	73.7	74.2
<b>Overall</b>	12.7	12.0	16.5	19.6	20.4

Table 5: Results of automatic evaluation on other LLMs and methods compared to STYLEREMIX.

score. We see continue to have the highest overall and obfuscation rate compared to these models.

### B.3 Shuffling Styles using the Sequential Method

One version of STYLEREMIX described in Section 2.2 is the sequential method, which runs the original text through each adapter sequentially. We hypothesized that the order in which the styles were rewritten might affect the final outcome. To test this, we randomly shuffled the order of the adapters of the styles axes over  $n = 3$  different seeds when changing 2 – 4 styles and tested automatic evaluations as we did in the main experiment.

Figure 5 shows the average and standard deviation for all the automatic evaluations for each domain and different number of styles changed. We first note that grammar and content preservation remains similar, with very low standard deviation. However, for specific domains, the obfuscation drop rate has a large variation between the three random shuffles. This most diverse obfuscation drop rates seen in Speech ( $\sim 14\%$  standard deviation) and Blog ( $\sim 9\%$  standard deviation). This indicates that the order of adapter in the sequential method could contribute to the overall efficacy of the method. Future work could experiment more with these findings.

Figure 5 shows the automatic evaluations when we shuffle 2 – 4 style axes adapters.

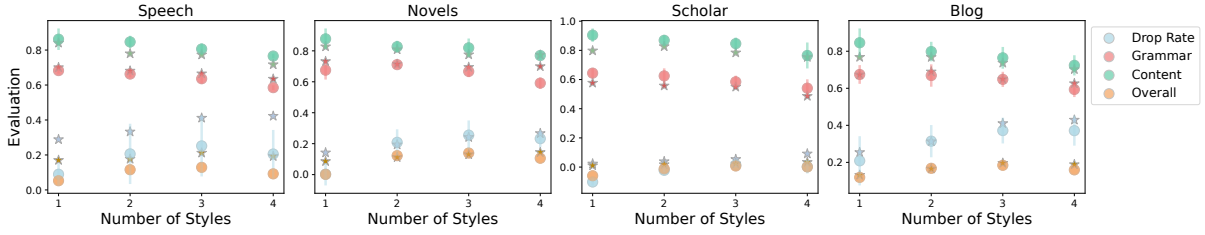


Figure 4: Base model merging with random styles,  $n = 3$

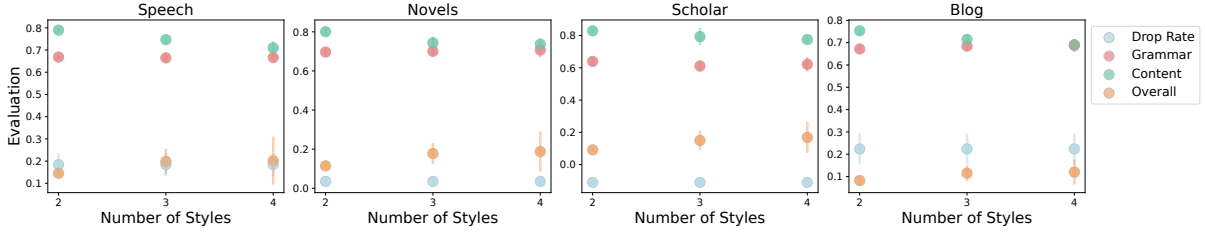


Figure 5: Seq. shuffle  $n = 3$

### B.4 Number of Styles Change

In STYLEREMIX the user can decide how many style adapters to use during obfuscation. We tested how obfuscation drop rate, grammar, and content preservation is affected when more style adapter are added. For this experiment, we used the base model adapter method and selected 1 – 7 styles using the difference from the author vector to the average domain vector.

Table 6 shows all the automatic evaluations for each number of style. At first, we see a stead increase in both obfuscation drop rate and overall score as we increase style adapters. This corresponds with changing more elements of the original text. However, as mentioned in our main paper, we see on average a 5% decrease in overall task score when using 4 to 5 style adapters. Then, as the number of style adapter increase, we see a stead decrease in content preservation and grammar. This correlates with a qualitative decrease in generations seen as we increase the styles over 5.

### B.5 Author Style Vector Analysis

In the pre-obfuscation phase, we choose 7 specific style axes to train the LoRA adapters; length, use of function words, grade level, voice, use of sarcasm, formality, and writing intent. Some of these style axes have rule-based evaluations, and others have classifier-based evaluations. We used these automatic evaluations to create a unique author vector for each author in a domain and use the difference in this vector compared to other authors in the same

# of Styles	1	2	3	4	5	6	7
<b>AUTHORMIX-Speech</b>							
<b>Drop Rate</b>	28.9	33.3	41.2	42.3	13.6	47.4	44.6
<b>Grammar</b>	70.0	68.1	66.5	63.5	61.4	52.9	46.1
<b>Content</b>	84.1	78.0	77.3	71.7	72.2	67.7	55.4
<b>Overall</b>	17.0	17.7	21.2	19.2	6.0	17.0	11.4
<b>AUTHORMIX-Novels</b>							
<b>Drop Rate</b>	14.2	19.3	24.2	26.7	36.1	32.9	83.7
<b>Grammar</b>	73.2	71.4	69.5	69.9	61.3	49.8	50.4
<b>Content</b>	82.7	80.9	77.6	77.4	73.7	68.4	51.6
<b>Overall</b>	8.6	11.2	13.0	14.4	16.3	11.2	21.8
<b>AUTHORMIX-Scholar</b>							
<b>Drop Rate</b>	2.3	3.8	5.3	9.2	2.3	50.4	73.5
<b>Grammar</b>	57.7	55.7	54.9	48.6	48.1	38.6	48.4
<b>Content</b>	79.8	82.7	78.4	75.3	71.7	30.8	47.5
<b>Overall</b>	1.1	1.8	2.3	3.4	0.8	6.0	16.9
<b>AUTHORMIX-Blog</b>							
<b>Drop Rate</b>	25.4	31.2	41.0	42.9	34.3	38.3	41.4
<b>Grammar</b>	67.3	68.9	64.9	62.6	55.3	46.4	44.0
<b>Content</b>	76.8	76.8	73.7	70.2	63.7	59.1	35.3
<b>Overall</b>	13.1	16.5	19.6	18.9	12.1	10.5	6.4

Table 6: Results of automatic evaluation on the base adapter merging method using different number of style adapters. We show the obfuscation drop rate, grammar, content preservation, and overall task score using 1 – 7 style adapters.

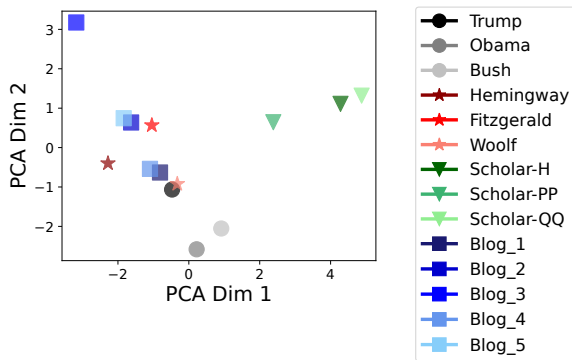


Figure 6: PCA, clustering analysis

domain to choose the styles axes to change during obfuscation. Although these selected style axes are just a subsample of suitable options, we wanted to explore how well these author vectors separate the authors in our test data set.

To analyze this, we first created an author vector for each author by taking the average of each automatic evaluation over the paragraphs in the authors test set. This resulted in 14 (authors), vectors with 7 (style axes) entries each. We then performed a principle component analysis (PCA) to reduce the size of the vector dimension to explain at least 90% of the variance in the data (it went from 7 to 4 dimensions). We note that the first two dimensions account for 70% of the variation.

Figure 6 compares all the authors (across all domains) using the first and second component of the PCA. First, we notice that the Scholar (triangles) and Speech (circle) domains have distinct clusters away from the other two domains. The most spread out domain is Blog (square) with one author quite different from the rest. Lastly, we see that the novel (start) dataset is closely clustered together, but are quite similar to 4 of the blog authors. We note that four of the blog authors have more storytelling writing styles, while the last one has a more diary-like, very informal writing style. This seems consistent then that it would cluster similarly as novels.

Overall, this analysis showed starting evidence that our style axes vectors were able to separate the diverse writing styles. Future research could continue to explore the types of style axes that are most important when obfuscating.

## B.6 More Qualitative Examples

In Appendix G we provide more examples from each author in the AUTHORMIX. We note that we

selected these samples by *randomly* selecting 3, paragraphs of less than 45 words for each author and then selecting the example from these three. For STYLEREMIX, we used the base model adapter method with 3 style adapters. From these examples, especially the Blog and Novels, we see the qualitative benefits of STYLEREMIX and it’s flexibility to adapt to different original author styles.

## B.7 Tradeoff between Obfuscation, Content Preservation, and Grammar

We want to note that there is a natural trade-off between authorship obfuscation, content preservation, and grammar. For example, a naive copying baseline would have high grammar and perfect content preservation but low obfuscation. On the other extreme, a complete gibberish output would score very low on grammar and content but high on obfuscation.

This phenomenon is well-documented in the context of style transfer. To assess the overall quality of generations where there are multiple objectives, previous work in style transfer (Krishna et al., 2020b; Hallinan et al., 2023; Xu et al., 2018; Patel et al., 2023) proposes taking the *product* (or geometric mean) of the metrics (instead of drop rate, for style transfer we have target style strength, still bounded). The intuition is that these style transfer systems should *jointly optimize all metrics* rather than just one or two; this is reflected in taking a product.

In line with this past style transfer work, we choose to use the product, as we think high-quality obfuscations should jointly prioritize the three metrics of fluency, meaning similarity, and obfuscation, and so that we do not encourage systems that only optimize one or two of these metrics. However, we note that certain aspects might be more important for users than others, and they might not want to use an equally weighted total.

However, as an alternative overall metric, we *present the overall score as an equally weighted average* of the drop rate, grammar, and content score below, rather than a product. See these results in Table 7. We note that this was used in other authorship obfuscation papers as an overall total metric as well (Fisher et al., 2024). Here, we again see **STYLEREMIX performs best overall in 3 of the 4 datasets**. Note that the decrease in performance on Scholar split of AUTHORMIX is due to the very low obfuscation rate among all methods, which results in only a difference of 5% between



STYLEREMIX and the main method.

## C Method Details

### C.1 Style Axes Selection and Evaluation

We choose seven different style axes. The first three style axes have rule-based evaluation; length, use of function words, and grade level. For length, we evaluate using the average words per sentence and for function words we use the number of function words. Additionally, we incorporated "grade level," which primarily measures the number of syllables. Since this measure can vary slightly, we averaged three similar metrics: the Flesch-Kincaid (FK; Flesch, 1948), Linsear Write (L; O'Hayre), and the Gunning Fog Index (GF; Gunning, 1952) metrics. The exact formulas are given below; for more details, see <https://github.com/textstat/textstat>.

- **Flesch-Kincaid** is computed via:

$$KF = 0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

- **Linsear Write** is computed by:

1. Take a 100-word sample from the text
2. Make a score starting with 0. For every "easy" word ( $\leq 2$  syllables), add 1 point. Otherwise add 3 points ("hard" words have  $\geq 3$  syllables).
3. Divide points by number of sentences in the 100-word sample.
4. Divide by 2 if the points  $< 20$ , otherwise divide by 2 and subtract 1.

- **Gunning Fog** is computed by selecting a passage around 100-words long, then applying the following formula:

$$GF = 0.4 \left[ \left( \frac{\text{words}}{\text{sentences}} \right) + 100 \left( \frac{\text{complex words}}{\text{words}} \right) \right]$$

where complex words are words with three or more syllables.

The next four style axes have model-based evaluation; use of sarcasm, voice (active or passive), formality, and writing intent (descriptive, expository, narrative, and persuasive). Although these were chosen arbitrarily, we believe they do reflect some unique aspects of authorship style. However, these styles do require a unique classifier to automatically evaluate a text. For formality we used a RoBERTa-base (Liu et al., 2019) based formality classifier (Babakov et al., 2023),

found at <https://huggingface.co/s-nlp/roberta-base-formality-ranker>.

However, for the other three axes (voice, sarcasm, and writing intent) there was not a reliable, existing model, so we trained our own classifiers. We follow the same procedure to make DISC detailed in Section 2.1, but 1) with *different* base training data, to ensure that there is no overlap between the classifier and adapter data and 2) only for the following style elements: voice passive, voice active, sarcasm less, sarcasm more, and persuasive, expository, narrative, and descriptive. With the new datasets of length 1500 for each style element, we then train RoBERTa-large (Liu et al., 2019) discriminators for the voice, sarcasm, and writing intent categories, splitting the train into 85% train and 15% dev set. We set the seed to 0 and train with a batch size of 128, learning rate of  $5e-5$ , and for 5 epochs.

For all models, we choose the checkpoint with the best evaluation accuracy product (to ensure high accuracy for all classes); this corresponded to 100%, 99.1%, 45.5% for sarcasm, voice, and type respectively. Each model took approximately 1 hour to train using 1 NVIDIA A100 GPU with 80 GB of VRAM.

### C.2 DISC Training Data and Evaluations

We use GPT4-Turbo (OpenAI, 2023) to distill the style axes into 16 parallel training sets. We collect 1500 paragraphs from Wikipedia, books and plays, and blogs, then prompt GPT4 with the following: "Rewrite the following paragraph to include the same content but specific prompt\n Paragraph: paragraph \n Rewrite: " where paragraph is the original data. Depending on the target style, we change the specific prompt to:

- **Length short:** "being more succinct"
- **Length long:** "being more verbose."
- **Lower grade-level:** "using language an early elementary school student can understand."
- **Higher grade-level:** "use high school reading level or above."
- **More function words:** "using far less function words (i.e. pronouns, determiners, and conjunctions)."
- **Less function words:** "using far more function words (i.e. pronouns, determiners, and conjunctions)."

Dataset	Llama-2-Chat		Llama-3-Inst		Gemma-Inst	Paraphrase	MT	Stylo	JD	StyleRemix		
	7B	13B	8B	70B	7B					Seq.	AM	AM+LoraHub+
Speech	56.60	57.30	55.17	55.73	56.50	59.73	51.43	47.47	47.43	55.97	<b>61.67</b>	56.40
Novels	55.63	56.07	55.13	56.07	55.07	53.47	46.17	46.13	48.23	<b>58.53</b>	57.60	57.33
Scholar	52.27	52.03	51.53	51.03	51.40	<b>53.73</b>	49.60	40.47	43.00	48.53	44.37	41.67
Blog	56.20	56.47	57.07	56.90	58.10	57.47	45.00	42.33	54.03	57.73	59.87	<b>60.50</b>

Table 7: Results from Table 2, but now using an **equal-weighted average** as the overall score rather than the product of metrics, across all domains and models.

- **More sarcasm:** “with more sarcasm.”
- **Less sarcasm:** “with less sarcasm.”
- **More formal:** “with more formal language.”
- **More informal:** “with more formal language.”
- **Active voice:** “with active voice.”
- **Passive voice:** “with passive voice.”
- **Persuasive writing style:** “with persuasive writing style.”
- **Expository writing style:** “with expository writing style.”
- **Narrative writing style:** “with narrative writing style.”
- **Descriptive writing style:** “with descriptive writing style.”

We use sampling with a temperature of 1.0. As a result of this prompting, we achieve  $1500 \cdot 16 = 24000$  generations spanning 16 unique style directions from GPT-4.

We then validate the quality of this data. For axes with available automatic metrics, specifically length, function words, grade level, and formality, we run their respective metrics on the original texts, and on the GPT-4 generations in both directions, i.e., we run the formality classifier on the original texts, and on both the more and less formal GPT-4 generations. For the axes without automatic evaluation, we instead randomly evaluate 10% of them. Specifically, we randomly combine generated data from the same style axis but different directions (such as more and less sarcasm), and ask annotators (three NLP experts) to label if the style axis is high or low (or the specific type for Writing Type), then compute the accuracy.

Table 1 shows the results. For the metrics that we can automatically evaluate, our generated data

captures the desired axes and directions well; for example, the texts steered towards higher length have the highest average number of words per sentence. For sarcasm and voice, human evaluations of 97.7% and 93.7% respectively indicate that the generations match the targeted directions. For writing intent, the human evaluation accuracy is 77.7% which is still a good number as the task of discriminating between four classes is inherently more complex.

### C.3 Style Adapter Training

We train LoRA adapters (Hu et al., 2021) using each of the 16 generated parallel datasets. Specifically, we train Llama-3 8B (base model) on the following prompt for each of the datasets:

```
<bos>### Original:{original} \n ###
Rewrite: {rewrite} <eos>
```

where original and rewrite denote the original text and rewrite is the text we generated from GPT-4. Note that the format we train on is the same for all parallel datasets to make future model merging more effective.

We train the 16 LoRA modules each for 5 epochs with a seed of 0, batch size of 6, and a max sequence length of 512; we choose the checkpoint with the best eval loss and have an early stopping criteria of 5. For LoRA parameters, we use  $r = 32$ , the rank of the matrix, and the alpha and dropout values of 32 and 0.01 respectively. Overall, each LoRA adapter involves training 13 million parameters each, about 0.16% of the total parameters in Llama-3 8B.

All of our models train well over time on both train and eval loss; please see our repository for exact training curves and loss numbers for the 16 models. We train each of the models on a single A100 80GB GPU for about 2 hours each.

### C.4 Concatenating Style Adapters

Given a model with weights  $W_0$  of dimension  $d \times d$  LoRA freezes  $W_0$  and trains two matrices:  $A$  of

size  $r \times d$  and  $B$  of size  $d \times r$ . At inference, we use the new weights of  $W = W_0 + BA$ .

In the situation when we have  $n$  LoRA adapters, parameterized by  $A_1 \dots A_n$  and  $B_1 \dots B_n$  and want to ensemble them for inference, we use **concatenation**. Specifically, we concatenate each of the  $A_1 \dots A_n$  matrices resulting in a matrix  $A_{1 \dots n}$  of size  $nr \times d$ . Similarly, we concatenate each of  $B_1 \dots B_n$  resulting in a matrix  $B_{1 \dots n}$  of size  $d \times nr$ . We then can combine the matrices the same way to get new weights of  $W = W_0 + B_{1 \dots n} A_{1 \dots n}$ . Notably, we have no additional inference latency by concatenating the vectors, only a slightly increased fixed matrix multiplication cost.

### C.5 Style and Weight Selection

As described in our paper, we developed an automatic method for selecting the style axes to change, direction, and weights of the adapters. First, we create an author vector for each author in a domain, which is a vector with 10 automatic evaluations; average words/sentences, average number of function words, average grade level (using FK, L, GF) (Flesch, 1948; O’Hayre; Gunning, 1952), average likelihood score from formality classifier (Babakov et al., 2023), average likelihood score from sarcasm classifier (see Appendix C.1 for more details), average likelihood from a voice classifier (see Appendix C.1 for more details), average classification into each of the four writing intents. We label this vector for author  $i$  as  $x_i \in \mathbb{R}^{10}$ .

In order to select the  $k$  number of styles axes to change, we use the other authors in the same domain as a baseline. Specifically, we average the values from all authors in the domain and find the styles of author  $i$  that are furthest from this average vector. More specifically, we use the following formula:

$$\text{styles to change} = \text{top}_k \left( \left| x_i - \sum_{j=1}^m x_j \right| \right),$$

where we have  $m$  total authors in the domain and  $\text{top}_k(y)$  is a function which selects the rows of  $y$  with the highest values. Similarly, we use the sign of this difference to decide on the direction of the change. For example, if the sign of the difference is negative, then the author’s style value is lower than the average and we will implement a higher direction (driving the style up to average).

Once the styles axes are selected, we use different methods for choosing the adapter weights

for each style axes. First, we also use the author difference vector to select the weight of the adapter. To do this, we calculate the number of standard deviation the author’s value is from the average vector. We then use this metric to map to a static weight; see Table 12. We note that these weights were selected in line with past work (Huang et al., 2023).

Second, we employ a non-gradient based optimization method called LoraHub (Huang et al., 2023). This method uses a few validation examples to optimize the values. For this method, we developed our our loss function which is the sum of the chosen style axes automatic evaluations as well as the grammar. Specifically,

$$L = \sum_{v_i \in \text{selected axes}} \begin{cases} v_i & v_i \leq \frac{1}{m} \sum_{j=1}^m x_i \\ 1 - v_i & v_i > \frac{1}{m} \sum_{j=1}^m x_i \end{cases}$$

where  $v_i$  represents the style value for a selected style axis of the obfuscated text and the grammar score. In Table 9, we show the difference between the base initial weights, chosen using the static method, to the once optimized using LoraHub.

# of Std. Deviations	Adapter Weight
0-1	0.7
1-2	0.9
2-3	1.2
3+	1.5

Table 8: This shows the static mapping used in the base adapter merging method. We use the number of standard of deviations an authors automatic style score is from the average style score of all authors in that domain. The static values were chosen base on past work (Huang et al., 2023)

## D Experimental Details

In this section we provide full details of the experimentation used in this paper. We start with the dataset in Appendix D.3, method implementations for each method in Appendix D.4, and our evaluation methodology in Appendix D.5.

### D.1 Software

We used Python 3.10.13, Pytorch 2.1.2, Hugging-Face Transformers 4.39.3. and NLTK 3.8.1. All code is licensed under the Apache License 2.0.

### D.2 Hardware

All experiments were run on either a single NVIDIA A100 GPU or 4 NVIDIA A100 GPUs

Author	Styles Axes	Base Weights	LoraHub Weights
<b>3 Style Adapters</b>			
<b>Trump</b>	['grade level', 'length', 'sarcasm']	[0.9, 0.9, 0.9]	[1.18, 0.96, 0.91]
<b>Obama</b>	['length', 'sarcasm', 'persuasive']	[0.7, 0.7, 0.7]	[0.68, 0.74, 0.75]
<b>Bush</b>	['sarcasm', 'formal', 'grade level']	[0.7, 0.7, 0.7]	[0.71, 0.56, 0.55]
<b>Hemingway</b>	['grade level', 'sarcasm', 'expository']	[0.9, 0.9, 0.7]	[1.16, 0.91, 0.71]
<b>Fitzgerald</b>	['descriptive', 'grade level', 'sarcasm']	[0.7, 0.7, 0.7]	[0.65, 0.58, 0.41]
<b>Woolf</b>	['expository', 'formal', 'grade level']	[0.9, 0.7, 0.7]	[1.17, 0.64, 0.95]
<b>Scholar-H</b>	['descriptive', 'voice', 'sarcasm']	[1.5, 0.7, 0.9]	[0.92, 0.28, 0.64]
<b>Scholar-PP</b>	['descriptive', 'grade level', 'voice']	[1.5, 0.7, 0.9]	[1.42, 0.72, 0.93]
<b>Scholar-QQ</b>	['length', 'grade level', 'narrative']	[0.9, 0.9, 1.5]	[1.16, 0.90, 1.46]
<b>Blog-1</b>	['expository', 'grade level', 'formal']	[0.9, 0.9, 0.7]	[0.90, 0.90, 0.95]
<b>Blog-2</b>	['length', 'expository', 'formal']	[0.7, 0.7, 0.7]	[0.93, 0.65, 0.68]
<b>Blog-3</b>	['sarcasm', 'descriptive', 'formal']	[0.9, 0.7, 0.9]	[0.78, 0.55, 0.74]
<b>Blog-4</b>	['formal', 'sarcasm', 'narrative']	[0.7, 0.7, 0.7]	[0.68, 0.45, 0.67]
<b>Blog-5</b>	['formal', 'voice', 'expository']	[0.7, 0.9, 0.7]	[0.61, 0.77, 0.50]
<b>3 Style Adapters</b>			
<b>Trump</b>	['length', 'grade level', 'persuasive', 'sarcasm']	[0.9, 0.9, 0.9, 0.9]	[1.27, 1.15, 0.88, 0.85]
<b>Obama</b>	['grade level', 'sarcasm', 'persuasive', 'length']	[0.7, 0.7, 0.7, 0.7]	[0.70, 0.70, 0.70, 0.70]
<b>Bush</b>	['formal', 'descriptive', 'grade level', 'sarcasm']	[0.7, 0.9, 0.7, 0.7]	[0.32, 0.07, 0.34, 1.05]
<b>Hemingway</b>	['sarcasm', 'grade level', 'expository', 'length']	[0.9, 0.9, 0.7, 0.9]	[0.98, 0.80, 0.66, 0.96]
<b>Fitzgerald</b>	['sarcasm', 'descriptive', 'grade level', 'length']	[0.7, 0.7, 0.7, 0.7]	[0.73, 0.70, 0.67, 0.72]
<b>Woolf</b>	['length', 'grade level', 'formal', 'narrative']	[0.7, 0.7, 0.7, 0.9]	[0.06, 0.30, 0.77, 0.24]
<b>Scholar-H</b>	['sarcasm', 'expository', 'voice', 'formal']	[0.9, 1.5, 0.7, 0.7]	[1.44, 1.36, 0.60, 0.74]
<b>Scholar-PP</b>	['formal', 'grade level', 'descriptive', 'voice']	[0.9, 0.7, 1.5, 0.9]	[1.24, 0.55, 1.47, 0.59]
<b>Scholar-QQ</b>	['length', 'narrative', 'formal', 'grade level']	[0.9, 1.5, 0.7, 0.9]	[0.91, 1.25, 0.70, 0.90]
<b>Blog-1</b>	['formal', 'narrative', 'length', 'grade level']	[0.7, 0.9, 0.9, 0.9]	[1.07, 1.16, 0.80, 0.67]
<b>Blog-2</b>	['expository', 'length', 'formal', 'sarcasm']	[0.7, 0.7, 0.7, 0.7]	[0.76, 0.70, 0.71, 0.66]
<b>Blog-3</b>	['formal', 'grade level', 'sarcasm', 'descriptive']	[0.9, 0.9, 0.9, 0.7]	[1.15, 0.90, 0.90, 0.70]
<b>Blog-4</b>	['narrative', 'formal', 'sarcasm', 'length']	[0.7, 0.7, 0.7, 0.7]	[0.58, 0.28, 0.46, 0.95]
<b>Blog-5</b>	['descriptive', 'voice', 'grade level', 'formal']	[0.7, 0.9, 0.7, 0.7]	[0.69, 0.70, 0.58, 0.59]

Table 9: Comparison of the initial base weights, chosen using the standard deviation to static mapping, and the optimized LoraHub weights, found using our customized loss function. We show the style axes changed, the base weights and the LoraHub weights for each author in each domain.



with 80GB memory. We estimate our total computational use to be approximately 80 GPU hours.

### D.3 Data

Dataset	Author	Train	Eval	Test	Total
Speeches	Trump	6,443	1,596	2,677	10,716
	Obama	810	189	331	1,330
	Bush	617	139	251	1,007
Novels	Hemingway	1,516	504	504	2,524
	Fitzgerald	2,658	885	885	4,428
	Woolf	1,469	488	488	2,445
Scholarly	H	91	-	45	136
	PP	110	-	85	195
	QQ	85	-	67	152
Blog	1	3,399	-	677	4,076
	2	1,073	-	143	1,216
	3	1,064	-	210	1,274
	4	595	-	217	812
	5	396	-	142	538

Table 10: Details of AUTHORMIX, including the number of samples for the test/eval/train for each domain.

As mentioned, we wanted to use a test dataset which had a wide range of diverse authorship styles as well as domains. For this reason, we decided to create a new data set on authorship obfuscation called AUTHORMIX. This dataset is composed for four domains; presidential speeches, early 1900s fiction novels, scholarly articles, and dairy-style blog entries. Altogether, AUTHORMIX contains over 30,000 high-quality paragraphs from 14 authors.

For the presidential domain, we curate and clean a novel collection of high-quality presidential speeches from George W. Bush ( $n = 38$ ), Barack Obama ( $n = 29$ ), and Donald Trump ( $n = 26$ )<sup>8</sup>, transcribed by the Miller Center (Miller Center of Public Affairs)<sup>9</sup> at the University of Virginia. We broke the speeches naturally into paragraphs and then selected all paragraphs between 2 – 5 sentences. This resulted in a total of  $n = 13K$  paragraphs.

Similarly, we also decided to develop a new collection of early 1900s fiction writers from the with strong writing styles, therefore we choose text from books by Ernest Hemingway, F. Scott Fitzgerald, and Virginia Woolf which were collected from Project Gutenberg (Gutenberg). We selected the top 4 most popular books on Project Gutenberg

<sup>8</sup>These presidents were selected due to their diverse styles but similar time periods, which minimizes content discrepancies.

<sup>9</sup><https://data.millercenter.org>

for each author and then again, used the natural paragraphs from each author. We selected all paragraphs between 2 – 5 sentences. This resulted in a total of  $n = 9K$  paragraphs.

Lastly, we altered the existing data from two current datasets, the Extended-Brennan Greenstad (Brennan et al., 2012) which is a collection of “scholarly” short (500-word) paragraphs gathered from Amazon Mechanical Turk (AMT) and the Blog Authorship corpus (Schler et al., 2006b), a collection of blogs (diary-style entries) that were posted to blog.com. We note, these datasets match those used in (Haroon et al., 2021), (Mahmood et al., 2019), and (Fisher et al., 2024). For the AMT dataset, we used authors "h", "pp", and "qq" and we artificially created paragraphs by chunking the text into a random collection of 2-5 sentences (as the text is not naturally broken into paragraphs). For the Blog dataset, we used authors "5546", "11518", "25872", "30102", "30407", we used the natural paragraphs. Then, to match the speech and novel domains, we edited to include all paragraphs between 2 – 5 sentences and 3 words. This resulted in  $n = 500$  and  $n = 8K$  paragraphs for the AMT and Blog accordingly.

#### D.3.1 Artifact Terms of Use

Our artifacts allow for fair use under Project Gutenberg (Gutenberg): [https://gutenberg.org/policy/terms\\_of\\_use.html](https://gutenberg.org/policy/terms_of_use.html)

### D.4 Method Implementation

#### D.4.1 Baselines

For each baseline, we use the optimal set of hyperparameters reported in its respective paper.

**Stylometric (Stylo)** We used (Karadzhov et al., 2017) method for AO using stylometric methods, which was originally proposed in the PAN-2016 Author Masking Shared Task competition (Mihaylova et al., 2016). This method calculates metrics for 12 features that are indicative of style, then modifies the text, so these metrics align with an "average" value. The "averages" were calculated using a combination of training sets including the PAN-2016 Author Obfuscation task (Mihaylova et al., 2016) and public domain books from Project Gutenberg (Gutenberg). Examples of the metrics this method uses include the average number of words per sentence, word frequency, and the use of uppercase letters. Changes employed include actions such as sentence splitting and merging, sub-

stitution of words with synonyms, and alterations in spelling. For a full list of metrics and proposed changes, see the (Karadzhov et al., 2017). To further enhance the obfuscation process, the method introduces "noise" by modifying words that differ between English and British English and introducing additional functional words. We make no changes to the hyperparameters used in the original method.

**Machine Translation (MT)** We used a round-trip machine translation method proposed by Keswani et al. (2016). In this method, they translate the original text from English to German, German to French, and then French back to English. We enhanced their method by use of the new M2M translation model (Fan et al., 2020), which does not rely on English as an intermediate language.

**JAMDEC (JD)** This method was proposed by Fisher et al. (2024) and uses a small language model, GPT2-XL (Radford et al., 2019), as the base model. For this method, they use a three stage approach where they extract the keywords of text (to guide generation to have the same content), over-generate using diverse constrained beam search, and then filter based on grammar and content overlap. We used this model’s default parameters, with a beam width of 10, and only using the likelihood keyword extractors, which was recommended to be just as effective but take less time. More details of this methods’ implementation can be found (Fisher et al., 2024).

**Paraphrasing** We used the paraphrasing model from Jung et al. (2024). This model uses Google T5 (Raffel et al., 2020) as the base and is finetuned on the dataset DIMPLE, which is a dataset of 4M high-quality pairs of paraphrases.

**Instruction LLMs** Lastly, we wanted to compare with LLMs of similar and bigger sizes. For these, we opted to use instruction tuned models which could easily follow instruction to rewrite the text. For each model, we used a temperature of 1.0 and a top-p of 0.9. Table 11 shows the exact prompts used to generate the generations from each of the baseline LLMs.

#### D.4.2 STYLEREMIX

**Style and Weight Selection** We used the described automatic style and weight selection described in the paper and in Appendix C. We note

that almost all values were less than 3 standard deviations, with the majority between 0 – 2.

**Adapter Merging** We used three different ablation of our methods; sequential, adapter merging base, and adapter merging LoraHub+. For the sequential method, we averaged results over  $n = 3$  random shuffling of style axes orders. For the adapter merging base method we used the weight adapter found from mapping using the standard deviations.

For the adapter merging LoraHub+, we build on the prior LoraHub method (Huang et al., 2023). We used the weights selected using our mapping method as the initial values and then used a non-gradient based optimization (Liu et al., 2020) over a new customized loss function. The loss function adds together the automatic evaluations from the author vector (described in Appendix C.5) for the specific style axes that are being considered for merging. Note, that since we are optimizing by finding the lowest loss, if the direction of the style axes is "higher" we take  $1 - value$  and if the direction is "lower" we just add the *value*. Lastly, we also add the grammar score into the loss to maintain good fluency. Then, a non-gradient based optimization method is use (Liu et al., 2020). Note, we use non-gradient based due to the large number of parameters of the model. We provide a comparison of the base weights chosen and the optimized LoraHub weights in Table 9.

**Hyperparameter Selection** To tune the hyperparameters of STYLEREMIX, we use the validation split of AUTHORMIX.

#### D.5 Evaluation Methodology and Other Details

**Obfuscation: Classifier** We train classifiers over each of the four domains in AUTHORMIX to measure obfuscation during evaluation using their respective training and development sets. Specifically, for each of AUTHORMIX- { speech, novels, scholar, blog }, we train a RoBERTa-Large classifier (Liu et al., 2019) with a learning rate of  $5e-5$ , batch size of 64, seed of 0, a max length of 256, and for 10 epochs. We set an early stopping threshold of 5, and choose the best checkpoint based on the best evaluation accuracy product (to ensure high accuracy for all classes).

Overall, our final evaluation accuracy products for AUTHORMIX- {speech, novels, scholar, blog }

Model	Instruction
Llama 2	"[INST] «SYS»\n You are a helpful assistant.\n \n «/SYS» \nPlease rewrite the following: <paragraph>[/INST] Rewrite: "
Llama 3	"[INST] «SYS»\n You are a helpful assistant.\n \n«/SYS» \nPlease rewrite the following: <paragraph>[/INST] Rewrite: "
Gemma	"You are a helpful assistant.\n \nPlease rewrite the following: <paragraph> Rewrite: "
Mistral	"<s>[INST] You are a helpful assistant.\n \nPlease rewrite the following: <paragraph> [/INST] Rewrite: "

Table 11: The instruction used for prompting the LLMs used as baselines.

# of Std. Deviations	Adapter Weight
0-1	0.7
1-2	0.9
2-3	1.2
3+	1.5

Table 12: This shows the static mapping used in the base adapter merging method. We use the number of standard of deviations an authors automatic style score is from the average style score of all authors in that domain. The static values were chosen base on past work (Huang et al., 2023)

are 74.5%, 85.6%, 100%, and 70.8%, while the average overall accuracies are 90.6%, 95.8%, 100%, and 93.3%. Further training details including loss functions can be found in our repository. We train each of these models with a single NVIDIA A100 80 GB GPU for approximately 2 hours.

**Content Preservation: Cosine Similarity** We compute neural text embeddings on the inputs and their obfuscations in Sentence Transformers (Reimers and Gurevych, 2019). Next, we use the cosine similarity between the two vectors to gauge semantic similarity and get a approximation of content preservatinon. Note that though the cosine similarity can output values from -1 to 1, we find on all of our validation dataset (across all datasets and methods) **all** similarities between inputs and their obfuscations are non-negative, with a bound of 0 to 1. If the similarity metric were to, in a very rare case, have a negative value, we would set the value to 0 so that we could have a still meaningful overall product of metrics; however, we never observe this.

**Grammar: CoLA** To ensure both fluency and grammaticality, we use TextAttack (Morris et al., 2020), a RoBERTa-large model (Liu et al., 2019) fine-tuned on the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018) which includes 10,600 sentences with binary annotations for linguistic acceptability.

## E Human Evaluation

We omit the MT and Stylo methods from human evaluation as Fisher et al. (2024) show that JamDec outperforms them in previous work for both automatic and human evaluation. We also omit 13B and 70B models for fair comparison. Finally, we report human evaluation for the best performing STYLEREMIX.

We used workers from Amazon Mechanical Turk (AMT) who voluntarily opt-in to the task to annotate  $n = 20$  text from each author. Each text was annotated by  $n = 3$  authors, who were paid at a rate of 15\$/hour. Our annotators are from English-speaking countries. A screenshot of the interface is shown in Figure 7

Our agreement numbers for the five metrics we collect, grammaticality, fluency, content preservation, low content addition, and obfuscation are 99.8, 98.2, 95.3, 93.8, and 93.5% respectively.

To compute the overall score, it would be unfair to take the product of our metrics (grammaticality, fluency, content preservation, low content addition, and obfuscation), since content preservation and fluency have two metrics representing their category, while obfuscation has only one; rewrites that simply output fluent, content-preserving texts would score higher than more balanced obfuscations which sacrifice small amounts of content preservation. Instead, we take a *weighted* product. Our overall product is computed as:

$$\text{overall} = \frac{\text{grammaticality} + \text{fluency}}{2} \times \frac{\text{content preservation} + \text{low content added}}{2} \times \text{obfuscation}$$

This product fairly takes the product of the three major categories (fluency, content preservation, and obfuscation), which aligns with our automatic metrics.

## F Alternative Obfuscation Evaluation Metrics

To verify the obfuscation effectiveness of STYLEREMIX, we run an alternative evaluation to measure drop rate using the method from Learning Universal Authorship Representations (LUAR; Soto et al., 2021) and train models to learn authorship embeddings for each of the four domains in our AuthorMix dataset (speech, novels, scholar, and blog) using the training data. We use the default hyperparameters from the codebase (ie, training for 20 epochs, using sentence-transformers/paraphrase-distilroberta-base-v1 as the base model, etc).

Next, we create authorship embeddings for all authors by passing their validation data into the trained models with their respective domain where they are aggregated, resulting in a single embedding for each author. To perform authorship attribution and obtain predictions for a set of input data of size  $N$  over some domain (such as speech), we first pass the input data through the trained model to extract individual embeddings before they are aggregated, resulting in  $N$  input embeddings. For each of these input embeddings, we calculate the cosine similarity with each validation authorship embedding in the specified domain. The predicted authorship style is the one with the highest cosine similarity.

As with our RoBERTa models, we can compute the drop rate for the LUAR method. Recall that the drop rate is the drop in accuracy of the classifier evaluated on the original text and the obfuscated text, where accuracy represents how many of the text the classifier correctly identified the author. For each domain in AuthorMix, we obtain the initial classification accuracy with the LUAR method using the test set. Then, we calculate the LUAR drop rate for our StyleRemix methods and for the baselines.

In Table 13, we display Table 2 from the main obfuscation results, but now add two new rows which we compute: [NEW] Drop Rate w/ LUAR, the drop rate with LUAR, and [NEW] Overall (using Drop Rate w/ LUAR), the overall obfuscation quality, equivalent to the product of the grammar, content and Drop Rate w/ LUAR which we compute.

Across all datasets, the new authorship attribution results with LUAR aligns with our RoBERTa based results and reinforces the strong obfuscation results of StyleRemix over baselines. Specifi-

cally, across all datasets, StyleRemix has the highest LUAR drop rate on the speech, novels, and scholar datasets, and the second-highest LUAR drop rate on the blog dataset, beating much larger baselines like Llama-3-70b-Inst. This is the same as the previous obfuscation results and underlines the effectiveness of StyleRemix for obfuscation. Furthermore, the new overall metric with LUAR drop rate confirms the results from the original overall metric: over all datasets, StyleRemix generates the best overall obfuscations, beating all baselines.

Overall, our additional evaluation using LUAR authorship attribution confirm the previous result obtained with the RoBERTa classifiers and demonstrates the excellent anonymization capabilities of StyleRemix.



Model Size	Llama-2-Chat		Llama-3-Inst		Gemma-Inst	Paraphrase	MT	Stylo	JD	STYLEREMIX		
	7B	13B	8B	70B	7B					Seq.	AM	AM + LoraHub*
AUTHORMIX- Speech												
Drop Rate	18.2	24.0	17.6	16.8	23.1	24.1	10.3	15.1	29.2	34.9	41.2	31.4
[NEW] Drop Rate w/ LUAR	8.3	7.4	4.7	6.8	5.3	3.2	0.0	7.7	9.2	3.3	23.9	12.2
Grammar	67.8	67.1	67.1	70.2	67.8	71.2	54.9	37.8	56.7	61.7	66.5	63.9
Content	83.8	80.8	80.8	80.2	78.6	83.9	89.1	89.5	56.4	71.3	77.3	73.9
Overall	10.3	13.0	9.5	9.5	12.3	14.4	5.1	5.1	9.4	<u>15.3</u>	<b>21.2</b>	14.8
[NEW] Overall (using Drop Rate w/ LUAR)	4.7	4.0	2.5	3.8	2.8	1.5	0.0	2.6	2.9	1.5	<b>12.3</b>	<u>5.8</u>
AUTHORMIX- Novels												
Drop Rate	12.2	13.7	9.2	11.3	13.3	10.8	7.0	13.5	24.9	19.3	28.6	35.6
[NEW] Drop Rate w/ LUAR	3.6	5.0	5.7	3.6	6.4	5.3	2.2	10.4	16.5	8.8	17.9	31.7
Grammar	71.8	73.8	73.1	75.4	70.0	68.3	46.3	36.8	61.2	72.6	68.1	63.5
Content	82.9	80.7	83.1	81.5	81.9	81.3	85.2	88.1	58.6	83.7	76.1	72.9
Overall	7.3	8.2	5.6	6.9	7.6	6.0	2.8	4.4	8.9	11.8	<u>14.8</u>	<b>16.5</b>
[NEW] Overall (using Drop Rate w/ LUAR)	2.1	3.0	3.5	2.2	3.7	2.9	0.9	3.4	5.9	5.3	<u>9.3</u>	<b>14.7</b>
AUTHORMIX- Scholar												
Drop Rate	0.8	1.5	1.6	2.5	0.0	0.8	1.5	4.6	6.1	1.8	9.2	11.5
[NEW] Drop Rate w/ LUAR	6.1	2.6	5.2	6.1	6.9	0.9	0.0	1.8	5.2	10.4	10.4	13.0
Grammar	64.3	64.9	64.1	66.6	65.3	69.1	54.5	31.0	62.3	65.8	48.6	44.7
Content	91.7	89.7	88.9	84.0	88.9	91.3	92.8	85.8	60.6	78.0	75.3	68.8
Overall	0.5	0.9	0.9	1.4	0.0	0.5	0.8	1.2	2.3	0.9	<u>3.4</u>	<b>3.5</b>
[NEW] Overall (using Drop Rate w/ LUAR)	3.6	1.5	3.0	3.4	<u>4.0</u>	0.6	0.0	0.5	2.0	<b>5.3</b>	3.8	<u>4.0</u>
AUTHORMIX- Blog												
Drop Rate	17.7	21.3	21.8	18.9	27.5	22.2	9.4	12.1	56.4	34.4	41.0	42.0
[NEW] Drop Rate w/ LUAR	7.2	9.4	7.5	5.9	12.3	14.2	6.0	9.1	19.4	12.7	16.2	14.2
Grammar	68.4	69.1	71.3	74.0	69.0	69.8	41.9	29.1	60.6	66.7	64.9	65.3
Content	82.5	79.0	78.1	77.8	77.8	80.4	83.7	85.8	45.1	72.1	73.7	74.2
Overall	10.0	11.6	12.1	10.9	14.8	12.5	3.3	3.0	15.4	16.5	<u>19.6</u>	<b>20.4</b>
[NEW] Overall (using Drop Rate w/ LUAR)	4.1	5.1	4.2	3.4	6.6	8.0	2.1	2.3	5.3	6.1	<b>7.7</b>	<u>6.9</u>

Table 13: Comparison of obfuscation methods measured by 1) mean drop rate, grammar, meaning similarity, and overall (the same metrics and results as in Table 2) and 2) additionally with an alternative, LUAR-based drop rate metric and a new overall score computed with this drop rate. **Bold** and underline denote the highest and the second-highest score respectively in each row. All metrics displayed in the table are multiplied by 100 for easier viewing of significant figures.

Instructions (click to expand/collapse)

**Thanks for participating in this HIT! Please read the instructions carefully.**

In this HIT, you'll be asked to give feedback on the effectiveness for a method to hide (obfuscate) a specific authors' writing style. You will be given the **original text** (written by Author A) and a **rewritten text** which aims to hide (obfuscate) Author A's writing style.

Please consider the following attributes of the **rewritten text** in comparison to the **original text**.

**Characteristics of a good rewritten text:**

- Sensible:** The **rewritten text** should be grammatically correct and make logical sense.
- Content:** All content from the **original text** should be present in the **rewritten text**. The **rewritten text** should NOT be a paraphrase or summary of the **original text**, but instead contain all the original content and sentiment. However, there should also not be any new information in the **rewritten text** that was not conveyed in the **original text**.
- Style:** The **rewritten text** should be stylistically different from the **original text**. In other words, you should have a hard time identifying that the **rewritten text** was written by Author A.

**You will be asked the following 5 questions to evaluate the quality of the rewritten text:**

- Grammar:** How **grammatically** correct is the **rewritten text**?
  - Tip: Does the rewritten text have good grammar?
- Fluency:** How **fluent** (natural sounding) is the **rewritten text**?
- Content Preservation:** How much **content** is **preserved** in the **rewritten text** compared to the **original text**?
  - Tip: This means the rewritten text should contain all the important information (e.g., names, places, actions) from the original text.
  - Tip: The rewritten text **should NOT** be a summary or paraphrase of the original text.
- Content Addition:** Is there **new content added** in the **rewritten text** not in the **original text**?
  - Tip: The rewritten text should **NOT** add significant new information (e.g., names, places, actions) that is not in the original text, or change any information.
- Style:** How **similar** is the **style** between the **rewritten text** and the **original text**?
  - Tip: Style can compose of many factors including word choice, punctuation, use of slang, sentence structure, etc.
  - Tip: Having different styles means that you would not guess that Author A wrote the rewritten text.

Examples (click to expand/collapse)

**Example 1:**

**Original Text:**  
I was wondering if you could recommend any good cheese? I am having a dinner party and would love to serve some as an appetizer.

**Rewritten Text 1:**  
I hope recommend fine cheese? We had a dinner partie and would love to give to people.  
**Grammar:** Bad **Fluency:** Bad **Content Preservation:** Fair **Content Addition:** Perfect / Good **Style:** Fair

**Rewritten Text 2:**  
Is there cheese you could offer a recommendation for? Tonight, there is a dinner party I am hosting and giving some at the start would be good.  
**Grammar:** Fair **Fluency:** Bad **Content Preservation:** Perfect / Good **Content Addition:** Perfect / Good **Style:** Fair

**Example 2:**

**Original Text:**  
A recent NC State University graduate won the cheese-rolling women's championship in 2022. She said she "practiced for hours, rolling down many hills in preparation.

**Rewritten Text 1:**  
In 2022 a NC State University student (who had just graduated) won the cheese-rolling women's championship.  
**Grammar:** Perfect / Good **Fluency:** Perfect / Good **Content Preservation:** Bad **Content Addition:** Perfect / Good **Style:** Fair

**Rewritten Text 2:**  
In 2022 an NC State University recent graduate won the cheese-rolling women's championship in Gloucestershire, England. As a proud ex-wellieball player, she said she "practiced for hours" by rolling down hills.  
**Grammar:** Perfect / Good **Fluency:** Perfect / Good **Content Preservation:** Perfect / Good **Content Addition:** Bad **Style:** Fair

**Example 3:**

**Original Text:**  
I am at the moment writing a lengthy indictment against our century. When my brain begins to reel from my literary labors, I make an occasional cheese dip.

**Rewritten Text 1:**  
I am at the moment drafting a lengthy indictment against our era. When my brain begins to weaken from all my literary labors, I sometimes make an aromatic cheese dip.  
**Grammar:** Perfect / Good **Fluency:** Perfect / Good **Content Preservation:** Perfect / Good **Content Addition:** Perfect / Good **Style:** Bad

**Rewritten Text 2:**  
I'm currently busting my brain writing a massive rant about how messed up our century is. But hey, when my head starts spinning from all that heavy thinking and writing, I take a breather and whip up some awesome cheese dip. Gotta keep the snacks game strong, you know?  
**Grammar:** Perfect / Good **Fluency:** Perfect / Good **Content Preservation:** Perfect / Good **Content Addition:** Perfect / Good **Style:** Perfect / Good

**Task**

<b>Original Text</b> <code>\$(original_text)</code>	<b>Rewritten Text</b> <code>\$(obfuscated_text)</code>
--------------------------------------------------------	-----------------------------------------------------------

- Q1. Grammar:** How **grammatically** correct is the **rewritten text**?  
Tip: Does the rewritten text have good grammar?
  - Perfect / Good It has no grammar mistakes, or very minor grammar issue that doesn't interfere with reading.
  - Fair It has noticeable grammar issues.
  - Bad It has major grammar errors that interfere with reading significantly.
- Q2. Fluency:** How **fluent** (natural sounding) is the **rewritten text**?  
Tip: Does the rewritten text convey meaning fluently (natural sounding)?
  - Perfect / Good It is mostly fluent. It was mostly easy to read.
  - Fair It is less fluent. It was a bit difficult to read.
  - Bad It is not fluent. It was very difficult to read.
- Q3. Content:** How much **content** is **preserved** in the **rewritten text** compared to the **original text**?  
Tip: Is all the content from the original text contained in the rewritten text?
  - Perfect / Good The content is completely consistent. It leaves out no information.
  - Fair The content is mostly consistent. It leaves out some information, but the meaning is still related.
  - Bad The content is very inconsistent and has lost a lot of the original meaning.
- Q4. Content:** Is there **new content added** in the **rewritten text** not in the **original text**?  
Tip: Is there new content in the rewritten text that is not in the original text?
  - Perfect / Good The content is completely consistent. It adds no new information.
  - Fair The content is mostly consistent. It adds some information, but the meaning is still related.
  - Bad The content is very inconsistent and has added a lot of new meaning.
- Q5. Style:** How **similar** is the **style** between the **rewritten text** and the **original text**?  
Tip: Does it seem like the rewritten text was written by the same author as the original text?
  - Perfect / Good The authors of the two text are definitely **different**.
  - Fair I have some doubt the the two texts are written by the same author.
  - Bad The authors of the two text are definitely the **same**.

Figure 7: Instructions used in the human evaluation

## G Further Examples

AUTHORMIX-Speech	
<b>Author</b>	Trump
<b>Original</b>	Americans know that in a world where others seek conquest and domination, our nation must be strong in wealth, in might, and in spirit. That is why the United States vigorously defends the traditions and customs that have made us who we are.
<b>MT</b>	The Americans know that in a world where the other is seeking conquest and domination, our people must be strong in wealth, power and spirit. That’s why the United States defends the traditions and habits that have made us know who we are.
<b>Stylo</b>	However, Americans know that in a world where others seek conquest and domination, our nation must be strong in wealth, in might, and in spirit. Lastly, That is why the United States vigorously depends the traditions and customs that have made us who we are.
<b>JAMDEC</b>	Americans know that in a world where others seek conquest and domination, our nation must be strong in wealth, in might, and in spirit. That is why the United States vigorously defends the traditions and customs that have made us who we are.
<b>Paraphr</b>	In a world where others strive for dominance and control, Americans recognize the importance of being strong in wealth, strength, and spirit. As a result, the United States vigorously protects the cultural and historical heritage that has contributed to our identity.
<b>Llama2 (7B)</b>	As Americans, we recognize the importance of a strong and resilient nation in a world where power-hungry individuals and nations seek to conquer and dominate. To protect our identity and values, the United States fiercely guards the traditions and customs that define us as a nation. These beliefs are the bedrock upon which our great country was built, and they continue to serve as the foundation for our continued prosperity and influence in the world today.
<b>Llama3 (8B)</b>	Americans understand that in a world where some nations strive for dominance, it is crucial that our own nation remains robust in its economic, military, and moral strength. This is why the United States actively preserves and upholds the values, customs, and traditions that have shaped our identity and character.
<b>Mistral V2</b>	Americans recognize the importance of maintaining strength in wealth, military power, and cultural values in a global landscape where others may strive for conquest and dominance. Consequently, the United States is committed to safeguarding its heritage and customs that define its identity.
<b>Gemma (7B)</b>	In a world where others strive for dominance, Americans believe their nation must maintain strength in wealth, military power, and morale. To uphold their heritage and identity, the United States stands firm in protecting its cherished traditions and customs.
<b>STYLEREMIX</b>	Americans hold the conviction that, in a universe where many nations pursue expansion and hegemony, our country must maintain its robustness in wealth, military prowess, and unwavering resolve. Hence, the United States zealously protects the very traditions and values that have shaped our distinct identity.
<b>Author</b>	Obama
<b>Original</b>	Thank you. God bless you. And God bless this country we love.
<b>MT</b>	Thanks for you. God blessed you. And God blesses this country that we love.
<b>Stylo</b>	Thank you, god bless you; and and God bless this country we love.

<b>JAMDEC</b>	And because we are on the cusp of a new era in human history, one that will be defined by our ability to boldly go where no one has gone before, to reach for new frontiers, to seek out new life and new civilizations, and to learn from the mistakes and successes of the past so that we might better prepare ourselves for the challenges of And because, in the words of President Kennedy, we choose to go to the moon in this decade and do the other things, not because they were easy, but because they are hard, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one which we intend to win, and the others, too."
<b>Paraphr</b>	Thank you, God bless you, and God bless this nation we cherish.
<b>Llama2 (7B)</b>	Thank you. May God bless you and our country.
<b>Llama3 (8B)</b>	I would like to express my gratitude to you. May God grant you His blessings. May God also continue to bless our beloved nation.
<b>Mistral V2</b>	I appreciate your kind words. May God's blessings be upon you. And may God continue to bless the country we cherish deeply.
<b>Gemma (7B)</b>	Thank you. May God bless you and this beloved nation.
<b>STYLEREMIX</b>	Thank you. May God bless you, and may He continue to bless this nation that we hold dear.
<b>Author</b>	Bush
<b>Original</b>	His tax – his policies of tax and spend – of expanding government rather than expanding opportunity – are the policies of the past. We are on the path to the future – and we're not turning back.
<b>MT</b>	His taxes ,his tax and expenditure policy ,the extension of the government rather than the extension of possibilities ,are the policies of the past. We are on the way to the future ,and we will not return.
<b>Stylo</b>	His tax– his policies in tax and spend– in expanding government rather than expanding opportunity– are the policies in the past. We are on the path to the future– and we 're not turning back.
<b>JAMDEC</b>	His tax – his policies of tax and spend – of expanding government rather than expanding opportunity – are the policies of the past. We are on the path to the future – and we're not turning back.
<b>Paraphr</b>	The current tax policies, which prioritize government expansion over promoting economic growth, are a reflection of the past. However, we are on the path to the future and we will not retreat.
<b>Llama2 (7B)</b>	His tax and spend policies, which focused on expanding the government rather than creating opportunities for growth, are a relic of the past. We are moving forward with a renewed commitment to innovation and progress, leaving behind the stagnant policies of the past.
<b>Llama3 (8B)</b>	His tax philosophy of spend and tax, which prioritizes government expansion over creating opportunities, is a relic of the past. We're now on a trajectory towards a brighter future and we won't be reversing course.
<b>Mistral V2</b>	His outdated tax and spending policies, which focus more on expanding government rather than creating opportunities, are no longer relevant. We are moving forward towards the future and will not revert to these old ways.
<b>Gemma (7B)</b>	His tax and spending policies, which aim to expand government rather than foster opportunity, reflect outdated ideologies. We are committed to embracing a forward-looking vision that prioritizes progress and innovation.
<b>STYLEREMIX</b>	He's all about taxing up a storm and spending it all, plus, he's all about growing government big instead of creating more chances for everyone. That's really old-school thinking. We're all about moving forward and not looking back.

---

**AUTHORMIX-Novels**

---

<b>Author</b>	Hemingway
---------------	-----------



<b>Original</b>	"Don't worry about money," Mike said. , "You can pay for the car, Jake, and I'll send you my share."
<b>MT</b>	"They don't worry about money,"Mike said. , "You can pay for the car, Jake, and I'll tell you my part."
<b>Stylo</b>	" Don ' t worry about money, "Mike said, and , " You can pay for the wagon, Jake, and I ' ll send you my party, and "
<b>JAMDEC</b>	"We can pay the family and move on with our lives, and we won't have to worry about where we're going to live or who's going to take care of us when we're gone.
<b>Paraphr</b>	Mike advised Jake not to worry about finances. He suggested that Jake could pay for the car himself, and he would then send him his share of the cost.
<b>Llama2 (7B)</b>	"Don't worry about money," Mike said. "I'll cover my share of the cost, and you can pay for Jake's."
<b>Llama3 (8B)</b>	Mike reassured Jake, "Don't stress about the cost, Jake. You take care of buying the car, and I'll cover my portion of the expense."
<b>Mistral V2</b>	Mike assured Jake, "You don't need to be concerned about the cost of the car. I'll cover my portion and send it to you later."
<b>Gemma (7B)</b>	"Don't worry about the cost,"Mike said. , "I'll cover the car, Jake, and I'll transfer my share to you later."
<b>STYLEREMIX</b>	Mike addressed the financial concerns with a seemingly casual remark. "Don't stress over finances," he assured, "Jake, you handle the car expenses, and I'll cover my portion." This approach, seemingly nonchalant about the financial management, was delivered with a casual air.
<b>Author</b>	Fitzgerald
<b>Original</b>	She liked him immediately, reddish-brown mustache under watery blue eyes that had something in them that these other eyes lacked, some quality of appreciation. They exchanged stray sentences through dinner, and she made up her mind to see him again.
<b>MT</b>	She loved her immediately ,a red-brown moustache under the blue eyes of water that had something in them, which was missing in those other eyes, a kind of appreciation. They exchanged stray for dinner, and she awakened her mind to see him again.
<b>Stylo</b>	She liked her immediately, a reddish-brown mustache, under watery blue eyes that had something of them that these other eyes lacked, some quality of appreciation. They exchanged stray sentences through supper, and she made up her mind to see it again.
<b>JAMDEC</b>	
<b>Paraphr</b>	She immediately fell in love with him, noticing his distinctive reddish-brown mustache beneath his watery blue eyes, which had a unique quality that set him apart from other eyes. They exchanged brief sentences during dinner, and she made up her mind to see him again.
<b>Llama2 (7B)</b>	She was immediately taken by him, noticing the reddish-brown mustache framing his watery blue eyes. There was something unique about those eyes, a quality of thoughtfulness that set them apart from the other eyes she had seen. They exchanged brief words throughout dinner, and she found herself making plans to see him again.
<b>Llama3 (8B)</b>	As soon as she met him, she was drawn to him - the distinctive reddish-brown mustache and watery blue eyes that seemed to hold a certain depth of understanding. Throughout dinner, they exchanged brief, scattered phrases, and she found herself resolved to reconnect with him again in the future.
<b>Mistral V2</b>	From the first moment, she was drawn to him. His reddish-brown mustache framed watery blue eyes that held a unique quality of appreciation. During dinner, they engaged in brief conversations, and she found herself looking forward to another encounter with him.

<b>Gemma (7B)</b>	She instantly admired his reddish-brown mustache framing his watery blue eyes, which held a depth that set them apart from the others. They exchanged fleeting words throughout dinner, and she resolved to meet him again.
<b>STYLEREMIX</b>	She felt an instant connection with him. His reddish-brown mustache framed his watery blue eyes, which seemed to hold a special kind of appreciation, different from others she had seen. During dinner, they shared a few casual conversations, and she decided she wanted to meet him again.
<b>Author</b>	Woolf
<b>Original</b>	"But I must look so queer!" she cried, running over to the glass and looking first this side then that. Then she snatched it off again, for there was a tap at the door. Could it be Sir William Bradshaw? Had he sent already?
<b>MT</b>	"But I need to look so beautiful!" she cried, blown to the glass and looked first on that side, then on that. Then she pushed him, because there was a ball at the door. Can he be Sir William Bradshaw? Has he already sent?
<b>Stylo</b>	"But I must listen so queer; and "she cried, running over to the drink and looking first this side then that, then she snatched it off once again, for there was a tap at the door; could it be Sir William Bradshaw, had he sent already?
<b>JAMDEC</b>	she cried, looking at herself in the mirror. "I don't know what's the matter with me. "What is the matter with me?" she asked, looking at herself in the glass. She looked again, and there he was, sitting in a chair by the table.
<b>Paraphr</b>	She exclaimed, "I must appear unusual!" as she ran towards the glass and carefully examined it, first on one side and then the other. Then she quickly snatched it off again, as there was a tap at the door. Could it be Sir William Bradshaw? Had he already sent something?
<b>Llama2 (7B)</b>	"But I must look so strange!" she exclaimed, rushing over to the mirror and examining herself from different angles. She then quickly removed the mirror from the wall, as she heard a knock at the door. "Could it be Sir William Bradshaw? Has he arrived already?" she wondered aloud.
<b>Llama3 (8B)</b>	But I must look so peculiar! she exclaimed, hurrying to the mirror and scrutinizing herself from every angle. She then hastily removed the mirror, her attention distracted by a knock at the door. Could it be Sir William Bradshaw? Had he already arrived, sent by her aunt?
<b>Mistral V2</b>	"Oh dear, I must look quite strange!" she exclaimed, hurrying towards the mirror. She glanced at her reflection from various angles before hurriedly removing the mirror from the wall. A knock at the door interrupted her, and she wondered if it could be Sir William Bradshaw. Had he arrived already?
<b>Gemma (7B)</b>	In a sudden burst of anxiety, she exclaimed, "Oh, I must look so strange!" and rushed over to the glass. She peered cautiously through, casting her gaze first one way and then the other. Moments later, she snatched the glass away, for a tap at the door startled her. The unfamiliar sound made her wonder if it could be Sir William Bradshaw, and if he had already arrived.
<b>STYLEREMIX</b>	She suddenly felt awkward when she noticed herself in the mirror, turning this way and that to check her appearance. Just as she was about to examine herself further, someone knocked on the door. She wondered if it might be Sir William Bradshaw, wondering if he had already arrived.

#### AUTHORMIX-Scholar

<b>Author</b>	Scholar-h
<b>Original</b>	So what are the takeaways from the Nepal case study?
<b>MT</b>	So what are the Takeaways of the Nepal case study?
<b>Stylo</b>	So what are the takeaways from the Nepal case study?

<b>JAMDEC</b>	First of all, it is important to note that this is not the first time that the Nepalese government has had.
<b>Paraphr</b>	The Nepal case study has provided valuable insights.
<b>Llama2 (7B)</b>	What are the key takeaways from the Nepal case study?
<b>Llama3 (8B)</b>	Here are the key lessons learned from the Nepal case study: Please let me know if this meets your expectations or if you would like me to make any changes!
<b>Mistral V2</b>	What key insights can be gained from the Nepal case study?
<b>Gemma (7B)</b>	What are the key lessons learned from the Nepal case study?
<b>STYLEREMIX</b>	Imagine yourself in the heart of a Nepal case study, "what exactly are the lessons you're likely to glean? The narrative unfolds there, rich with nuances and insights, each experience potentially etched into the global lexicon.
<b>Author</b>	Scholar-pp
<b>Original</b>	Even if you can't eliminate identity theft, you can get instant fraud notifications any time your identity is used without your knowledge. Prevent identity theft with MyID.
<b>MT</b>	Even if you cannot eliminate identity theft, you can receive instant fraud messages at any time if your identity is used without your knowledge. Avoid identity theft with MyID.
<b>Stylo</b>	Conversely, Even if you can't eliminate identity theft, you can obtain instant fraud modifications any time your identity is used without your knowledge; prevent identity theft with MyID.
<b>JAMDEC</b>	If you have questions about your identity, or if you think someone is trying to use it without your permission, click here to get instant alerts from the Identity Theft Resource Center (ITRC) and the Federal Trade Commission (FTC) about identity theft and how to protect. Identity theft is the theft of your identity by someone who uses your identity without your permission, such as by using your Social Security number to open a bank account or apply for a credit card, or by opening up a new bank account in your name and using it to receive.
<b>Paraphr</b>	Despite the fact that it may not be possible to completely prevent identity theft, you can receive instant notifications of any fraudulent activity involving your identity without your knowledge. This helps to protect your identity and prevent any potential threats.
<b>Llama2 (7B)</b>	Although identity theft cannot be completely eliminated, MyID offers instant fraud notifications whenever your identity is used without your knowledge. Protect yourself from identity theft with MyID.
<b>Llama3 (8B)</b>	While it's impossible to completely eliminate the risk of identity theft, you can receive immediate alerts whenever your identity is used without your consent. Stay protected with MyID, your trusted identity monitoring solution.
<b>Mistral V2</b>	MyID allows you to receive immediate notifications if your identity is used without your consent, providing an extra layer of protection against identity theft. Although it's impossible to completely prevent identity theft, having this alert system can help you take swift action to minimize potential damage.
<b>Gemma (7B)</b>	Stay vigilant against identity theft with MyID. Receive instant fraud notifications whenever your identity is used without your knowledge, even if you can't completely eliminate the risk.
<b>STYLEREMIX</b>	Amidst the complexities of digital security, the safeguarding of your identity might be challenged beyond your control, yet, it is assured that you will be alerted to instances of suspected fraud almost instantaneously whenever your identity is invoked without your awareness. This protection is diligently offered through the mechanism of MyID, a strategic layer of defense, designed to help you actively prevent the encroachment of identity theft.
<b>Author</b>	Scholar-qq

<b>Original</b>	They further stated that an independent Angola under the MPLA would be dominated by urban mesticos and assimilados, effectively transferring power from one class of colonial masters to another.
<b>MT</b>	They continued to explain that an independent Angola under the MPLA is dominated by urban mesticos and assimilados, which effectively transfer power from one class of colonial masters to another.
<b>Stylo</b>	They further stated that an independent Angola under the MPLA would be dominated by urban mesticos and assimilados, effectively transferring power from one class of colonial masters to another.
<b>JAMDEC</b>	This was further stated in an article entitled "What would happen if Angola's urban elite became mestisos?" which was published in La Republica, the official mouthpiece of the Communist Party of Angola (MPLA).
<b>Paraphr</b>	According to the text, an independent Angola under the MPLA would be primarily governed by urban elites and assimilationists, effectively transferring power from one class of colonial leaders to another.
<b>Llama2 (7B)</b>	They further emphasized that an independent Angola under the MPLA would be dominated by urban mixed-race and assimilated individuals, effectively transferring power from one group of colonial oppressors to another.
<b>Llama3 (8B)</b>	According to them, if Angola became an independent country under the MPLA's leadership, it would likely be controlled by urban mesticos and assimilados, essentially replacing one group of colonial rulers with another.
<b>Mistral V2</b>	The MPLA's proposed independent Angola was predicted to be controlled by urban mesticos and assimilados, implying a shift in power from one colonial ruling class to another.
<b>Gemma (7B)</b>	According to their statement, an independent Angola under the MPLA would likely be controlled by urban dwellers who had mixed ancestry and those who had assimilated into Angolan society, effectively replacing one class of colonial masters with another.
<b>STYLEREMIX</b>	They explained that if Angola became independent, led by the MPLA, it might end up being mostly ruled by people who were born in cities and considered either mesticos or assimilados. This would mean that the power might just shift from one group of people who were once in charge, to another group who were also outsiders, but from a different background.

#### AUTHORMIX-Blog

<b>Author</b>	Blog-1
<b>Original</b>	Costco is a phenomenon you have to visit to believe. This is impulse buying taken to a new level.
<b>MT</b>	Costco is a phenomenon that you need to visit to believe. This is an impulse for the purchase to a new level.
<b>Stylo</b>	osteo 's a phenomenon you have to visit to believe, that is impulse buying adopted to a new level.
<b>JAMDEC</b>	Costco is a phenomenon you have to visit to believe. This is impulse buying taken to a new level.
<b>Paraphr</b>	Costco is a unique and remarkable experience that requires a visit to fully appreciate. It's like a shopping experience that takes the concept of impulse buying to a whole new level.



<b>Llama2 (7B)</b>	Costco is a truly unique and unforgettable experience that will leave you amazed and surprised at every turn. Known for its unparalleled impulse buying opportunities, this retail giant has taken the concept of spontaneous shopping to a whole new level. From its vast array of products to its unbeatable prices, a visit to Costco is an event that should not be missed. Whether you're a seasoned shopper or just looking for a fun day out, this incredible store is sure to delight and surprise you in ways you never thought possible. So come and experience it for yourself ,you won't be disappointed!
<b>Llama3 (8B)</b>	"Costco is an extraordinary experience that must be seen to be fully appreciated. It's a retailer that takes impulse buying to a whole new level, making it easy to get caught up in the excitement of discovery and purchase."
<b>Mistral V2</b>	Visiting Costco is an experience unlike any other, showcasing impulse buying at its peak.
<b>Gemma (7B)</b>	Costco's sprawling warehouses and endless aisles entice customers with an unparalleled shopping experience, urging them to explore the realm of bulk buying and spontaneous purchases.
<b>STYLEREMIX</b>	Costco is a place you really need to see to understand. It's a spot where you might just end up buying more than you planned.
<b>Author</b>	Blog-2
<b>Original</b>	If not, then not. "How long ago since you ate anything proper, hmm?"
<b>MT</b>	If not, then not. How long have you spent since you've eaten something good?
<b>Stylo</b>	If not, then not. "How a lengthy time ago since you ate anything proper, him?"
<b>JAMDEC</b>	And if not, well, then, I guess I'll just have to go back to the drawing board and figure out what to. I don't know what else to say, other than that I'm really happy that I got to be a part of it and.
<b>Paraphr</b>	"How long has it been since you last consumed a nutritious meal, hmm?"
<b>Llama2 (7B)</b>	If not, then not. How long has it been since you've had a proper meal?
<b>Llama3 (8B)</b>	It seems you haven't eaten a substantial meal recently, have you?
<b>Mistral V2</b>	If you haven't eaten a proper meal for some time now, hmm?
<b>Gemma (7B)</b>	If not, then not. How long ago was it since you ate something proper, hmmm?
<b>STYLEREMIX</b>	If not, then indeed, not. "How long has it been since you last consumed a proper meal?"
<b>Author</b>	Blog-3
<b>Original</b>	Quick lunch.....I wanted Jasper's today though. Their food is startin' to really be good.....mmmmm perogies and quesadillas....and pizza.....mmmmmmmmmmmm.....foooood... Bio class was pretty lame.
<b>MT</b>	Breakfast, but I wanted Jasper's today. Your diet begins to be really good...mmmm perogies and quesadillas....and pizza.....mmmmmm.....foooood... the biological class was pretty lame.
<b>Stylo</b>	Quick lunch ..... myself wanted paper 's today though; and their food is starting ' to really be good ..... mamma perogies also quesadillas .... and penza .... mmmmmmm-mmmmm ..... foooood... mio class was pretty lame.
<b>JAMDEC</b>	I don't know if I'll be able to get it, but I wanted to make sure I got it in time for lunch, so I made a quick lunch out of some of the things I had left over from the last time I made lunch, and a few things I found in the pantry that I thought would be good for lunch that I didn't have time to make lunch with, and then. I was going to be honest and say that I didn't really care for the food here, but it was Jasper's so I had to give it a try! I ordered the chicken and waffles and it was pretty good! The waffles were crispy on the outside and soft on the inside, and the chicken was cooked perfectly!! I would definitely come back here if I was in the area again!

<b>Paraphr</b>	The author wanted to have a quick lunch at Jasper's today, as their food is starting to be quite good. They are particularly fond of their perogies and quesadillas, as well as their pizza. The author found the bio class to be quite uninteresting.
<b>Llama2 (7B)</b>	"Ah, a quick lunch... I was actually hoping to visit Jasper's today, their food has really been impressing me lately. I can't get enough of their perogies, quesadillas, and pizza... mm mm mm... delicious! Unfortunately, bio class was quite dull today. "
<b>Llama3 (8B)</b>	I was craving Jasper's for lunch today, but unfortunately, I didn't get a chance to go. Their food has been getting better and better, and I've been eyeing their perogies, quesadillas, and pizza. Mmm, just thinking about it is making me hungry! Too bad I missed out. On a different note, bio class was pretty underwhelming today.
<b>Mistral V2</b>	I was considering having a quick lunch, but I had a craving for Jasper's today. Their food has significantly improved recently, and I find myself longing for their perogies, quesadillas, and pizza - oh, the deliciousness! However, my biology class left me feeling less than satisfied.
<b>Gemma (7B)</b>	My quick lunch today was at Jasper's. Their food is seriously good! I highly recommend their perogies, quesadillas, and pizza. Bio class was unfortunately pretty boring.
<b>STYLEREMIX</b>	Today's lunch session was brief and I had a particular desire to visit Jasper's. Their culinary offerings are notably improving, notably including their distinctive perogies, quesadillas, and pizza, which are truly exceptional. The experience was quite underwhelming during my bio class.
<b>Author</b>	Blog-4
<b>Original</b>	Nowadays, this phase I have been going through is making me face all these quite brutally. Because I choose it to be so. I have never been a woman who takes things swiftly (which is, yet another choice).
<b>MT</b>	Today, this phase I've done faces me with all this quite brutal. Because I chose to be like that. I've never been a woman who takes things quickly (i.e. another choice).
<b>Stylo</b>	Afterwards, Nowadays this: step I have been going through is making me face all these quite brutally; and because I choose it to be so, i have never been a wife which takes things swiftly ( which is yet another choice).
<b>JAMDEC</b>	I don't know what to do about it, but I'm going to try my best to get through this phase as best as I can, so that I can get on with my life. I don't know if I'll ever be able to look at myself in the mirror and be proud of what I look like, but at least I know that I'm not a. I choose to live my life as a woman, even though I know it's not a choice I'm allowed to make, it's a choice my body makes for me, and.
<b>Paraphr</b>	The current phase I am experiencing is causing me to confront these challenges with a brutal intensity, as I have never been a person who takes things lightly (which is another choice).
<b>Llama2 (7B)</b>	In recent times, I have been grappling with a phase that is proving particularly challenging. Rather than navigating it with ease, I find myself facing these challenges head-on with a sense of brutal directness. This is a choice I have made, one that I have always been hesitant to embrace. As someone who prefers to take things slowly and carefully, I have never been one to rush into situations without fully considering the consequences. This cautious approach has led to a certain degree of struggle, but it is a choice that I stand by nonetheless.
<b>Llama3 (8B)</b>	Lately, I've been confronting some harsh realities, and I have to admit that it's a deliberate choice I've made. I've never been one to rush through life's challenges, and that's a choice I've always made for myself.
<b>Mistral V2</b>	Currently, I'm experiencing this period in life quite intensely. I've decided to face its challenges head-on, although I've never been one to act quickly in the past.

<b>Gemma (7B)</b>	Currently, I am navigating through a phase that has forced me to confront various challenges with blunt honesty. This stems from my conscious decision to approach things at a deliberate pace, a characteristic that aligns with my intrinsic personality.
<b>STYLEREMIX</b>	These days, I'm really laying it all out there, hitting the tough stuff head-on, because that's just how I roll. It's definitely not my style to breeze through life; I've always been more of the slow-and-thorough type, another deliberate choice of mine.
<b>Author</b>	Blog-5
<b>Original</b>	Anyway, it looks great. I painted the same color, but in flat instead of eggshell, and it's a big improvement. I love the way flat paint looks.
<b>MT</b>	However, it seems great. I made the same color, but in flat instead of garlic, and it's a great improvement. I like how Flat Paint looks like.
<b>Stylo</b>	Anyway, it looks awesome, and i painted the same color colour, but in apartment instead in eggshell, and it ' s a big improved; and i love the way apartment flat paint looks.
<b>JAMDEC</b>	It looks like it's going to be a really good game. I can't wait to see what you come up with for the rest of the series, and I'm really looking forward to seeing what you do with the other characters in the series! Im going to try to get a few more of these in the.
<b>Paraphr</b>	The paint job turned out great. I used the same color, but in a flat finish instead of eggshell, and it's a significant improvement. I love the way flat paint looks.
<b>Llama2 (7B)</b>	"Ah, I see! Painting the same color but in a flat finish instead of eggshell has made a significant difference. I completely agree, flat paint can give a room a sleek and modern look that's simply stunning. It's great to hear that you're enjoying the new finish!"
<b>Llama3 (8B)</b>	To be honest, I'm really pleased with how it turned out. I opted to use the same colour, but in a flat finish instead of eggshell, and it's made a significant difference. I'm a big fan of the flat paint look, and I think it's a real improvement.
<b>Mistral V2</b>	The new look is excellent. I opted for flat paint instead of eggshell for the same color, and I'm thrilled with the outcome. Flat paint really enhances the appearance.
<b>Gemma (7B)</b>	In any case, it looks fantastic! I opted for flat paint instead of eggshell in the same color, and the difference is remarkable. I adore the seamless and sophisticated look of flat paint.
<b>STYLEREMIX</b>	The appearance has been significantly improved. The same color was chosen, but the application was altered from eggshell to flat, which has been found to significantly enhance the aesthetic. There is a preference for the appearance of flat paint.

---

More qualitative examples of different methods.