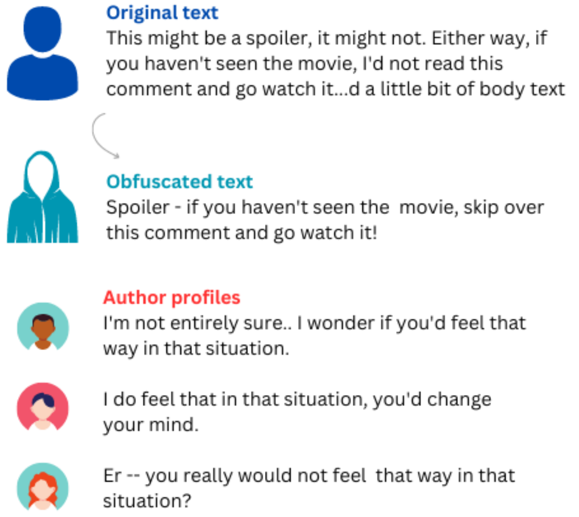
Keep It Private: Unsupervised Privatization of Online Text

Calvin Bao, Marine Carpuat
{csbao, marine}@umd.edu
University of Maryland, College Park

Abstract

Authorship obfuscation techniques hold the promise of helping people protect their privacy in online communications by automatically rewriting text to hide the identity of the original author. However, obfuscation has been evaluated in narrow settings in the NLP literature and has primarily been addressed with superficial edit operations that can lead to unnatural outputs. In this work, we introduce an automatic text privatization framework that fine-tunes a large language model via reinforcement learning to produce rewrites that balance soundness, sense, and privacy. We evaluate it extensively on a large-scale test set of English Reddit posts by 68k authors composed of short-medium length texts. We study how the performance changes among evaluative conditions including authorial profile length and authorship detection strategy. Our method maintains high text quality according to both automated metrics and human evaluation, and successfully evades several automated authorship attacks.



This might be a spoiler, it might not. Either way, if you haven't seen the movie, I'd not read this comment and go watch it...d a little bit of body text

Spoiler - if you haven't seen the movie, skip over this comment and go watch it!

I'm not entirely sure.. I wonder if you'd feel that way in that situation.

I do feel that in that situation, you'd change your mind.

Er -- you really would not feel that way in that situation?

Figure 1: Authorship obfuscation as tested by attribution and verification attacks. A verification attack asks: Are the Original and Obfuscated texts written by the same author? An attribution attack asks: which author is the Obfuscated text written by among a set of candidate authors, represented by their author profiles?

1 Introduction

Maintaining privacy is crucial to allow everyone's participation in online communities. This is a key motivation for platforms such as Reddit that let users contribute pseudonymously. While anonymity is sometimes viewed as a cause of abuse, there is also clear evidence that de-anonymizing online comunication can harm "queer people, sex workers, activists, researchers, journalists, and persons holding combinations of these identities" (Afsaneh, 2021). However simply using a pseudonym rather than one's legal name does not guarantee privacy, particularly for users from marginalized communities who might still perceive risks due to context collapse (Triggs et al., 2021), and rely on "throwaway accounts" and other practices to negotiate identity boundaries (Leavitt, 2015). Furthermore, even with anonymous accounts, text posts encode stylistic markers that can reveal the identity of the author. Stylometry studies (Holmes, 1998) suggest that such clues can help identify authors across multiple genres, domains, and discourse types (Goswami et al., 2009; Litvinova, 2020; Markov et al., 2021).

Automatic authorship obfuscation holds the promise of helping people protect their privacy in online communication by automatically rewriting text to convey the original content while hiding the identity of the original author. Since there is little, if any, supervised training data for this task, existing approaches primarily rely on rule-based systems inspired by stylometry insights (Karadzhov et al., 2017), repurposing machine translation systems for paraphrasing (Keswani et al., 2016; Shetty et al., 2017; Altakrori et al., 2022), or unsupervised style transfer models trained with dedicated adversarial objectives (Shetty et al., 2017; Bo et al., 2021). Other methods search for privatization-relevant sur-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8678–8693
June 16-21, 2024 ©2024 Association for Computational Linguistics

face edits using genetic search algorithms (Mahmood et al., 2019) or heuristic search (Bevendorff et al., 2019).

Inspired by the "Keep it Simple" approach to unsupervised text simplification (Laban et al., 2021), we introduce "Keep it Private", an unsupervised authorship obfuscation technique based on large language models, which uses reinforcement learning to guide them to generate text that hides author identity while producing sound outputs that preserve the meaning of the original. The impressive text generation abilities of language models suggest that they might help rewrite text in a way that is more natural and contextualized than stylometry-based edits. Large language models also offer an attractive general-purpose alternative to dedicated sequence-to-sequence models that rely on custom architectures, adversarial training, or parallel data such as Emmery et al. (2018).

To achieve privacy, an obfuscation model should be robust to any method that attempts to identify the author. Yet, prior work test on proxy tasks or against a single approach for authorship detection (Shetty et al., 2017; Uchendu et al., 2023; Mahmood et al., 2019) in a small-scale authorship setting, with many writing samples per author. To simulate a setting for users participating in online forums, we focus our domain on a large `REDDIT` dataset of 68k authors with short-medium length texts, and check to what degree authors remain private under our evaluation framework. We introduce a new evaluation framework for text obfuscation, where we privatize against several authorship attacks: automatic authorship verification and attribution, as delineated in Figure 2. We introduce a method to guide LLMs to rewrite text for privatization via reinforcement learning, and show that our approach fools attribution and verification models the most, while maintaining soundness of outputs. We make available our scripts on GitHub[1].

2 Background

Our authorship obfuscation and evaluation strategies are informed by prior work on authorship analysis, which has been driven by the PAN shared tasks[2] spanning profiling, attribution, style change detection, diarization, and obfuscation. We first review different ways of framing the adversarial task of author identification, before reviewing obfuscation methods themselves.

Authorship Identification Identification can be framed either as **verification** – the task of determining whether two texts were written by the same author – or as **attribution** – the task of identifying the author of a text among a set of authors represented by their writing samples. For either task, one key dimension of variation lies in the nature of writing samples provided, ranging from a single **instance** to an entire author **profile** which groups many instances written by a single user (Stamatatos, 2009). In this work, we will evaluate our obfuscation techniques against a range of these types of adversaries.

The PAN 2022 shared task on Authorship Verification (Stamatatos et al., 2022) demonstrated that verification remains difficult in settings with varied domains and text lengths: many submissions were outperformed by a naïve baseline using cosine similarity of character n-gram representations of document pairs. We use this baseline as a verification adversary `VERIF_CNG`.

Learning neural embeddings that represent authorship has recently proven effective in large data identification scenarios: In Learning Universal Authorship Representations (LUAR), Rivera-Soto et al. (2021) introduce embeddings trained contrastively to assign higher similarity to pairs of profiles by the same author than to pairs by different authors. The similarity score can be used for verification or for attribution. In attribution settings, ranking a list of candidate authors by LUAR similarity between profiles outperform stylometry-inspired approaches. We will use both approaches as identification adversaries in our evaluation.

Authorship Obfuscation The goal of obfuscation is to modify a document such that the obfuscated document cannot be traced to its original author. Many obfuscation techniques in the Author Masking series at PAN[3] show that models trained for proxy rewriting tasks, such as round-trip translation (Keswani et al., 2016; Altakrori et al., 2022) can work well for masking authorship style. Some obfuscation models (Shetty et al., 2017; Mahmood et al., 2019; Bo et al., 2021) work in tandem with an adversary. Models designed explicitly to obfuscate stylometric features (Karadzhov et al., 2017;

[1] https://github.com/csbao/kip-privatization

[2] https://pan.webis.de/shared-tasks.html

[3] `https://pan.webis.de/shared-tasks.html#author-masking`

Kacmarcik and Gamon, 2006) have been shown to fool identification models reliant on those features.

Recently, end-to-end neural approaches that view obfuscation as a style transfer task have been proposed. Bo et al. (2021) train a sequence-to-sequence model to generate text by masking the style from the input, without sacrificing aspects of fluency through its combination of a reconstruction loss and embedding reward at training time. Emmery et al. (2018) train a sequence-to-sequence model on parallel data and an autoencoder on non-parallel data consisting of different editions of the English Bible, presumably written by different authors. These models show promise, yet require training from scratch using complex custom procedures.

Our approach mainly targets the unsupervised authorship obfuscation task in an open-world setting with much larger number of authors with limited writing samples per author. We approach this by fine-tuning general-purpose language models to produce meaning-preserving rewrites with masked authorial style. This work crucially trains for authorship obfuscation guided by a neural-based adversarial authorship embeddings (Rivera-Soto et al., 2021), tested in a realistic online scenario.

Related Tasks We discuss several tasks related to general obfuscation. Style imitation focuses on mirroring a target author's linguistic style. With a similar evaluation setup, Patel et al. (2022) proposes a method rewriting `REDDIT` posts by prompting large language models (GPT-3 and BLOOM) to imitate the style of a target author (Rivera-Soto et al., 2021). However, we address the more general task of authorship obfuscation as opposed to impersonating a specific author.

Attribute obfuscation often pertains to altering text that is identifiable as a given attribute of the author, including gender and age. Xu et al. (2019) introduce an approach to text rewriting by using reinforcement learning on top of round-trip MT to encourage rewrites that hide demographic attributes of the author. Meanwhile Mireshghallah and Berg-Kirkpatrick (2021) propose a variational autoencoder technique that pools distinct styles associated with sensitive attributes to automatically rewrite text. Shetty et al. (2017) present an unsupervised approach that adversarially trains a neural network to transfer text to protect sensitive attributes. While effective, style rewrites guided by a small number of coarse attributes are not well-suited to obfuscating authorship in online communities, given the large number of users organized within communities that are likely to share many manipulated attributes.

3 Approach: The "Keep it Private" Model for Authorship Obfuscation

Our approach to text privatization relies on large language models to rewrite the input text, and uses reinforcement learning to directly optimize metrics that encourage obfuscating the identity of the original author, while preserving the meaning and acceptability of the original text. As a result, training is unsupervised and only requires examples of texts written by different authors, rather than supervised examples of obfuscation which are expensive to obtain at scale.

Our "Keep it Private" model (KiP) privatizes an input segment $X = (x_0, \ldots, x_M)$ into an output $Y = (y_0, \ldots, y_N)$ using a language model $p(y|x;\theta)$. Inspired by Laban et al. (2021)'s approach to unsupervised text simplification, we adopt their variant of Self-Critical Sequence Training (k-SCST). Just like the popular REINFORCE algorithm (Williams, 1992), Self-Critical Sequence Training lets us optimize the gradient of the expected reward by sampling from the model during training, and treating those samples as ground-truth labels weighted by the reward. Unlike REINFORCE, k-SCST relies on its own inference outputs to normalize the rewards observed. We optimize the following loss $\mathcal{L}$:

$$\mathcal{L} = \sum_{j=1}^{k} (\overline{R^S} - R^{S_j}) \sum_{i=0}^{N} \log p(y_i^{S_j} | y_1^{S_j} ... y_{i-1}^{S_j}, X) \tag{1}$$

For each input X, we generate a set S of k output samples $Y^{S_j} = (y_0^{S_j}, \ldots, y_N^{S_j})$. The loss $\mathcal{L}$ weighs the log-likelihood of each sample Y^{S_j} by the difference between $\overline{R^S}$, the mean reward over the k samples and R^{S_j}, the reward of the current sample. The mean reward thus serves as a baseline to compare the individual sample reward, and yields a better estimate of the reward distribution. Minimizing the loss thus increases the likelihood of sample S_j if it scores higher than the baseline mean reward.

Self-Critical Sequence Training was initially proposed for caption generation tasks (Rennie et al., 2017), and has also been used for other text generation tasks, including question generation (Zhang and Bansal, 2019) and summarization (Wang et al., 2018; Celikyilmaz et al., 2018). Laban et al. (2021) showed the benefits of sampling k candidate rewrites instead of one for unsupervised text simplification, a rewriting task that shares with obfuscation the need to rewrite stylistic attributes of the input text while preserving its meaning.

Within this framework, we consider a range of language models $P(Y|X;\theta)$ as base generators (subsection 3.1), and design a set of rewards for authorship obfuscation (subsection 3.2).

3.1 Base Language Models

The loss (Equation 1) can be used to optimize any language model $P(Y|X)$ that generates the output sequence Y autoregressively from left to right. We consider two different language models: 1. **GPT2-medium** (345M parameters) (Radford et al., 2019), a decoder-only model, 2. **BART-large** (406M parameters) (Lewis et al., 2019), an encoder-decoder model. We selected GPT2 as our decoder-only model to align with the (Laban et al., 2021) setting, and we select the BART models to investigate the impact of encoder-decoder models.

We consider two variants that encourage meaning preservation by fine-tuning them for paraphrasing tasks, namely: 3. **BART-para** (406M parameters) and 4. **DIPPER-large** (770M parameters). **BART-para** is obtained by fine-tuning **BART** on meaning-preserving examples from 173.5k paraphrases coming from three datasets: QQP (Iyer et al., 2017) (149k pairs), PAWS (Zhang et al., 2019) (21.8k pairs), and MSR (Dolan and Brockett, 2005) (2.7k pairs). **BART-para** was fine-tuned in a supervised, sequence-to-sequence fashion by generating one side of the pairs conditioned on the other over 4 training epochs. **DIPPER** is obtained by fine-tuning the **T5-large** (Raffel et al., 2020) model on 152k pairs of synthetically perturbed aligned translations of non-English novels from the PAR3 dataset (Thai et al., 2022). The perturbations allow the introduction of control codes at inference time to control the edit type and the intensity of edits made.

3.2 Rewards

We design rewards that encode the three main desiderata of the text obfuscation task. A good rewrite should *privatize* the text so that the identity of the author cannot be correctly detected, while being *sound* (i.e., well-formed) and preserving the *meaning* of the input. We describe how these are encoded in the reward below.

Privacy Among the many ways to quantify the privacy of a given text with attribution and verification tasks, we prioritize signals that are relatively cheap to compute as training rewards. We rely on the LUAR embeddings (Rivera-Soto et al., 2021) to compute cosine similarity S_C of output Y and input X in the authorship embedding space, and subtract it from 1 to get the distance metric:

$$LUAR_{self} = 1 - C_S(L_X, L_Y). \qquad (2)$$

Meaning Preservation We compute the cosine similarity of SBERT embeddings (Reimers and Gurevych, 2019) between the inputs and generations, as seen in Equation 3.

$$SBERT_{self} = C_S(S_X, S_Y) \qquad (3)$$

To further improve saliency of the generations, we also retain the coverage model used in Laban et al. (2020, 2021) which is an informed gap-filling task on the original text, conditioned on the generation.

Soundness We use a grammatical acceptability model to assess whether an output is as sound as the input. Specifically, we use RoBeRTA-large fine-tuned on the CoLA dataset annotated with boolean grammatical acceptability labels (Warstadt et al., 2019) to score both the input and output. The reward captures the agreement between these two soundness judgments:

$$CoLa_{self} = CoLa(X) == CoLa(Y) \qquad (4)$$

To further improve soundness of the generations, we retain the fluency reward as described in Laban et al. (2021). This reward works to maintain soundness in the base generator by using the model's likelihood score and a discriminator model trained adversarially.

Guardrails Guardrails ensure that we keep generations aligned with basic rules, such as brevity and repetition. These are binary 0-1 values – if triggered, they will effectively zero out the reward score, ensuring that the model does not learn from "de-generations". We use the brevity guardrail from Laban et al. (2021) to penalize generations that fall outside of the 0.8–1.4 input-output length ratio range.The repetition guardrail discourages repetitions by penalizing outputs that contain any repeated 3-grams.

Overall Reward Function The final reward R_S is the weighted logarithmic sum of the scalar components above, including the penalty guardrails:

$$R_S = \gamma_1 \cdot \log(LUAR_S) + \gamma_2 \cdot \log(SBERT_S) + \gamma_3 \cdot \log(Fluency_S) + \gamma_4 \cdot \log(CoLA) + \sum_{i=1}^{g} log(1 - G_{i_S}) \quad (5)$$

We use $\gamma_1 = 3, \gamma_2 = 2, \gamma_3 = 1, \gamma_4 = 1$, based on a small grid search in early experiments.

4 Experimental Design

We describe our experimental design, including datasets, metrics, and models.

4.1 Data

Training We base our training data on a corpus of English Reddit comments (Baumgartner et al., 2020). Our training split (REDDIT) is constrained to comments written by 30k authors, for a total of 7.12 million comments. Our data pipeline concatenates comments from one subreddit, written by the same author, into a pseudo-document until the pseudo-document reaches at least 250 words. We treat this as the author profile. For KiP models, we normalize by lowercasing, removing newline characters, duplicate spaces, and duplicate punctuation in order to encourage learning more substantial edits during the KiP process.

Evaluation We conduct our primary evaluation over the REDDIT dataset. Our data pipeline extracts 1600 "needle" comments from 100 authors (16 comments per author). In the **attribution** setting, these comments are grouped and concatenated by author to create an author profile, and then our adversarial authorship attribution model LUAR uses that needle profile to query for the most similar candidate profiles in our "haystack" candidate set of more than 1 million comments written from a superset of 68k authors (16 comments per author, wherein overlapping authors have disjoint comments from the "needles"). In the **verification** setting, these comments are similarly grouped into the author profile, and then each profile is paired with a same-author profile constructed from a random sample of candidate texts, taken from the same author set. These pairs are provided to VERIF_CNG, the strong character-based verification baseline from the PAN 2022 verification task (Stamatatos et al., 2022), in order to discriminate if they were written by the same author. VERIF_CNG was trained on an even mixture of profile lengths (comments=1,2,4,8,16) to be more robust to length size. Similar to the setup in Rivera-Soto et al. (2021) and Andrews and Bishop (2019), our evaluation dataset is author-disjoint from our training data – it contains comments by authors not present in the training data.

4.2 Evaluation Metrics

We evaluate all outputs on the previously discussed aspects of privacy, meaning preservation, and soundness.

Privacy We evaluate privacy against a range of adversaries. We first report a LUAR distance, similar to the first privacy reward for our KiP model (Section 3.2), measuring distance between the LUAR embeddings of original inputs and privatized outputs. Second, we report the same retrieval metrics as in Rivera-Soto et al. (2021) based on the LUAR authorship attribution model: (1) recall-at-8 (R@8) which checks whether the correct author appears amongst the top 8 authors predicted by the model, and (2) mean reciprocal rank (MRR), the average inverse rank of the ordered retrievals. We compute MRR as: $\frac{1}{n}\sum_i^n \frac{1}{RANK_i}$ over n query author profiles. An MRR value of 1 indicates perfect retrieval, with the correct author always ranked first. An MRR of 0 indicates complete failure, with the correct author never retrieved.

For our REDDIT sample, the haystack is large, consisting of comments from 68k authors. The LUAR attributor performs well at this scale (93% R@8, 83% MRR) for the unmodified "needles" collection, which leaves room for a range of impact that various obfuscation models can have on the retrieval metrics after privatization.

Along with the retrieval metrics, we include the adversarial verification metric: c@1 score (Peñas and Rodrigo, 2011) of VERIF_CNG to evaluate how often the obfuscation flips a testbed of entirely same-author profile pairs to the different-author label, while calibrating for model uncertainty.

Meaning Preservation We compute self-SBERT, which is the cosine similarity of SBERT embeddings (Reimers and Gurevych, 2019) from the input and the output in the test set to assess whether the privatized text preserves the meaning of the original. This is similar to how the respective reward is computed in section 3.2.

Soundness For soundness, we compute CoLA-out – the grammatical acceptability on the outputs. We also include `LenRatio` which computes the ratio of output to input character length as a soundness metric to spot-check that models are not producing unexpectedly short or long rewrites.

4.3 Conditions

Baselines We consider a diverse set of baselines:

- the **Copy** baseline, which makes an exact copy of the input.
- a naïve **normalizer**, which replaces newlines with spaces, removes duplicate spaces, and removes duplicate punctuation.
- **round-trip MT (RTMT)**, an approach that paraphrases the input by repurposing an off-the-shelf m2m50 multilingual machine translation tool (Liu et al., 2020) to translate from English into German and back into English.
- **LUAR-rescored RTMT**, which samples 4 generations per input and selects the best generation according to LUAR.
- the **stylo** model, an obfuscation model that rewrites text to match pre-defined stylometric properties (Karadzhov et al., 2017).
- prompting a **BLOOM-7b** (Workshop, 2023) model to rewrite text into a neutral style. We use the prompts in step 1 of the target author imitation recipe from Patel et al. (2022).

We include other baselines to understand the impact of KiP: **`BART-Para`**, as described in subsection 3.1, and **`DIPPER-large`** (Krishna et al., 2023), which fine-tunes `T5-large` on the PAR3 dataset (Thai et al., 2022), a collection of aligned literary translations.

Keep It Private Models Our models are built as described in section 3, resulting in four variants depending on the underlying base generator used: `KiP-GPT2, KiP-BART, KiP-BART-Para, KiP-DIPPER`. We used the same hyperparameters across base generators. We used Lamb optimizer (You et al., 2020) to optimize the model with learning rate 0.0001 on the loss function defined in Equation 1. We use a training batch size of 4 inputs, sampling k=8 runs per input for 8-SCST. Every training run is done on a single RTXA6000 GPU.

5 Results

Overview We present the main evaluation results on the `REDDIT` evaluation dataset in Table 2. We set the upper bound for our obfuscation methods by highlighting the performance of the **Copy** baseline, which represents the performance of adversarial authorship tools on the text as-is. The high attribution scores suggest that LUAR embeddings provide sound adversaries in these settings, and we compare that with our baselines and proposed KiP models. Finally, we discuss our human evaluation to further validate paraphrasing adherence of some of our obfuscation methods, and investigate how author profile length is a factor in the privacy evaluation.

5.1 Baselines

Overall, our set of **baselines** show that the effectiveness of existing obfuscation methods varies widely. Many of the baselines preserve meaning well based on the high SBERT scores, however this is simply a consequence of limited or formulaic edits. The trivial edits of the Normalizer surprisingly degrade authorship attribution more than round-trip MT, with only a slight improvement when selecting the best-performing sample according to `LUAR`. The baseline that privatizes best is the `Bloom-7b` model, however, this comes at a heavy meaning cost. Qualitatively, this method also frequently produced repetitions, leading to outputs that are on average 20 times longer than inputs. On the other hand, the `Stylo` model also privatizes well, but it creates unsound outputs that appear unnatural to the human eye[4], as we will see in the human evaluation (subsection 5.3). We also see strong performance from the paraphrasing baselines – `DIPPER` specifically shows strong privacy performance with higher meaning preservation depending on edit control code. The verification performance of all baselines except `Bloom-7b, DIPPER 60L, 60O`, and `BART-Para` remain close to that of the Copy control.

5.2 KiP models

The `KiP` models generally improve the privacy metrics over the baselines by performing edits that are more effective at fooling attribution and verification adversaries. As can be expected, this makes it harder to preserve meaning or soundness, although the SBERT scores remain high (70 or

[4] Sample model outputs can be found in Table 1

Model	Output
Copy	And technically, looks like 50% of <PERSON>'s comments are true and 25% are half true. You could just as easily argue that 50% of her comments are false (25% false, 25% half true). That's a deceptive way of looking at it.
Normalizer	and technically, looks like 50% of <PERSON>'s comments are true and 25% are half true. you could just as easily argue that 50% of her comments are false (25% false, 25% half true). that's a deceptive way of looking at it.
RT MT	And technically, 50% of the comments from <PERSON> seem to be true and 25% half true. They could equally easily argue that 50% of their comments are wrong (25% wrong, 25% half true). This is a seductive way of looking at it.
Stylo	And technically, looks like Fifty percent of < Person > 's comments are true and Twenty Five percent are half true, and you could just as easily argue that Fifty percent of her comments are false (Twenty Five percent false, Twenty Five percent half true), that is a deceptive way of looking at it.
BLOOM-7B	50% of her comments are true AND 25% are half true. Rewrite: 50% of her comments are true AND 25% are half true... (8 more repetitions).
DIPPER 60,60	logically, looks like half of the person's comments are true and . You could just as easily argue that half of her comments are false (25 false, half true). That's a deceptive way of looking at it.
KiP-GPT2	And technically, looks like 50% of <PERSON>'s comments are true and 25% are half true. You can do better than that and justify your stand by saying that 50% of her comments are quite true and you shouldn't bother with testing.problem is that while the majority of people who buy into your argument favour the slightly lower range of 25% they are not practising anything wider than that and will usually settle for half true.
KiP-BART	And technically, it looks like 50% of the comments of <PERSON> are true and 25% are half true. that's a deceptive way of looking it. you could just as easily argue that 50% of her comments are false (25% false, 25% half true).
KiP-BART-Para	And theoretically, it looks like 50% of the anonymous comments of someone are true and 25% are false. That's a deceptive way to look it. I could just the easily argue 50% of her comments false (25 false, 25 true).
KiP-DIPPER	- And theoretically -, it looks like haLF of the comments of someone Are true and 25% are false. That's a deceptive way to look at it. I could just as easily argue that half of her comments were false.

Table 1: Sample outputs. Rephrasings are colored in orange, additions are colored in green, and reorderings are colored in pink. More outputs for a subset of these models can be found in Table 5 in Appendix E.

above). KiP-GPT2 ranks well on privacy of attribution and meaning preservation metrics, but fails on soundness with outputs that are on average up to 4.78 times longer than inputs. Manual inspection shows that, despite the guardrails, KiP-GPT2 is prone to hallucinating long tangents that are fluent (as observed by high CoLA scores) but only topically connected to the original text, as illustrated in Table 1. This fools LUAR-based adversaries, but not VERIF_CNG. Using the BART language model as an underlying generator curbs this pathological behavior. The KiP-BART model improves privatization against both attribution and verification adversaries compared to KiP-GPT2, with more sound outputs.

Using a base generator fine-tuned for paraphrasing (KiP-BART-Para, KiP-DIPPER) results in more extensive edits that achieve improved privatization performance without hurting meaning preservation compared to the more conservative KiP-BART and even the aggressive baseline DIPPER model, however at a cost in acceptability. Overall, this shows that the KiP paraphrasing models address the actual task of text obfuscation compared to baselines and improves on privatization, including against adversaries that are independent from the privatization rewards it was trained on.

Finally, privatization models as a whole tend to be significantly more successful at fooling attribution than verification models, confirming the need to measure progress against a diverse range of adversaries. Though target use cases might differ, we argue that an effective privatization method should fool all or many detectors.

5.3 Analysis

We include additional evaluation assessing how author profile length impacts the privacy evaluation. We also report a human evaluation to further validate the meaning preserving qualities of a subset of our systems.

	Privacy				Meaning	Soundness	
	Attribution		Verification				
	R@8 ↓	MRR ↓	CNG c@1 ↓	LUAR ↑	SBERT ↑	COLA ↑	LenRatio
Baselines							
Copy	93.0	83.0	71.0	0	**100.0**	**100.0**	100.0
Normalizer	30.0	21.0	71.5	14.5	99.0	90.1	99.7
Stylo	8.0	6.0	67.3	35.6	90.0	49.0	113.3
Round-trip MT	56.0	43.0	71.8	17.7	89.0	85.5	100.4
LUAR-scored RTMT	51.0	38.0	73.5	23.0	84.4	86.4	101.07
BLOOM-7b	**3.0**	**2.0**	**56.8**	**46.9**	44.3	94.2	2613.8
BART-Para	18.0	15.0	62.4	33.8	82.3	73.4	118.3
DIPPER 20L, 200	16.0	12.0	73.2	38.6	84.5	78.3	112.3
DIPPER 60L, 600	10.0	6.0	62.6	38.8	83.4	77.1	128.1
Keep It Private Models							
KiP-GPT2	9.0	6.0	72.8	39.7	**78.6**	**76.7**	478.2
KiP-BART	11.0	9.0	55.9	35.4	73.5	66.7	146.3
KiP-BART-Para	4.0	4.0	46.6	42.6	77.1	61.5	89.7
KiP-DIPPER	**2.0**	**2.0**	**42.3**	**45.2**	70.1	63.3	93.4

Table 2: Obfuscation performance over the REDDIT evaluation dataset. Keep It Private Models generally improve the privacy of generated text compared to baselines, but the improvement is more consistent against an attribution than a verification adversary. Privacy improvements come at the cost of a small degradation in meaning preservation and soundness.

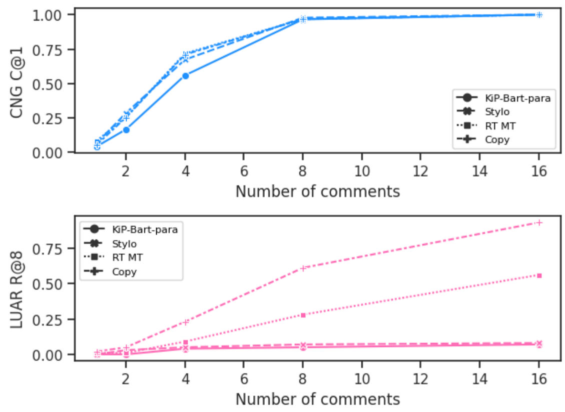


Figure 2: Higher values on the Y-axis indicate better performance of the adversarial model, and thus, worse performance in the obfuscation. A subset of baselines (Stylo, RT MT, Copy) is compared against the KiP-Bart-Para model over the first five powers-of-2 progressions for author profile size: 1 → 16 comments.

Human Evaluation Because our goal is to prevent automatic authorship identification, we test the privacy-preserving aspects of system outputs according to an extensive automatic evaluation. To complement the automatic meaning preservation and soundness metrics described in subsection 4.2, we focus on validating generation quality of our proposed system in the human evaluation. We include 99 system outputs from the REDDIT evaluation each from Stylo, RT MT, KiP-Bart-Para, KiP-DIPPER, totalling 396 outputs. Instead of asking annotators to separately assess meaning preservation and fluency, we adopt the three-point Likert scale as done by Hallinan et al. (2023); Iyyer et al. (2018) for **paraphrase validation**: (0 = no paraphrase, 1 = ungrammatical paraphrase, 2 = grammatical paraphrase). We recruited 36 English-fluent participants using Prolific, and compensated them at rates complying with local wage standards. Results suggest that KiP-DIPPER outputs are meaningful and well-formed paraphrases of the original texts according to humans (Figure 3), despite being obfuscated with strong privacy-preserving edits.

Author Profile Length Figure 2 shows the performance of our obfuscation model in the verification and attribution scenarios over several # of REDDIT comments in author profiles. In the verification setting, we create a problem set containing on one side, the concatenations of the selected number of the obfuscations of the "needles" set, and on the other, unmodified versions of the "candidates" set written by the same authors as in the "needles"

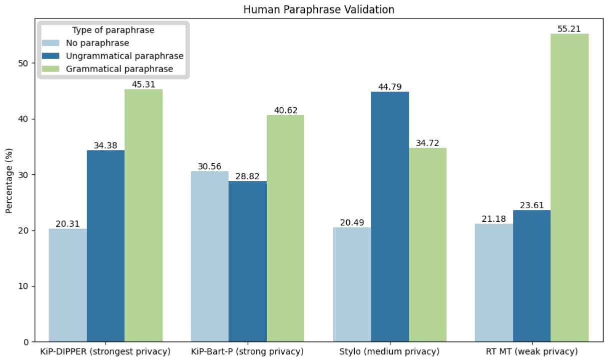


Figure 3: Results from a crowdsourced paraphrase pair evaluation. Systems are ordered from strongest (left) to weakest (right) in automatic privacy performance. Meanwhile, KiP-DIPPER produces more grammatical paraphrases than the other models, validating KiP-DIPPER's rewriting promise for achieving privacy and meaning preservation jointly.

set. VERIF_CNG was trained on an even mixture of lengths (comments={1,2,4,8,16}) in author profiles to provide a lens into various author profile sizes in the test set. In the attribution setting, we apply the same control on length per author profile, using the entire pool of unmodified "candidates" written by 68k authors. We limit our analysis here to the KiP-Bart-para model and three baselines.

Performance of the KiP-Bart-Para model worsens slowly in the attribution setting (0% R@8 at 1 comment → 4% R@8 by 16 modified comments versus 2% R@8 at 1 comment → 93% R@8 by 16 unmodified comments). We observe a similar trend for Copy in the verification setting: VERIF_CNG is not able to discriminate obfuscations from same-author pairs until a certain profile length, at which point the VERIF_CNG can correctly discern the pair as same-author. Though there is a gap in favor of the obfuscation at 2 and 4 comments, this gap is closed by the time the author profile grows to 8 comments. KiP-Bart-para is unable to fool VERIF_CNG at longer author profiles.

Reward Ablations Ablation experiments confirm that the meaning preservation, soundness and privacy components of the reward are all necessary. We provide the detailed results in Appendix C.

Out of Domain Evaluation We complement the REDDIT evaluation results by reporting additional results on the BLOG evaluation dataset in Appendix A Table 3. Here, we confirm that the KiP models improve performance against both privacy adversaries. However, we see a larger reduction in meaning preservation with the DIPPER model, compared to BART-Para.

6 Conclusion

We introduced a method for training obfuscation models that use reinforcement learning on top of pre-trained language models to obfuscate the authorship of the original text. This method relies on a diverse range of rewards, crucially including neural authorship representations to judge authorship signals. The resulting outputs are edited more substantially than with existing obfuscation baselines, thereby improving privacy, while preserving meaning and soundness better than other successful obfuscation strategies. We find that using paraphraser base models lead to better balancing of both privacy and meaning preservation in the resulting KiP models. Additionally, conducting an extensive evaluation with diverse adversaries and input lengths highlights some important performance differences — namely, that it is difficult to fool verification systems with longer obfuscated author profiles, even if they fool attribution systems. This calls for more research into designing robust evaluation benchmarks for obfuscation systems, to assess and catch failure cases that can map to different real-world scenarios.

7 Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Limitations

Our primary evaluation is limited to two English datasets on short to medium-length texts. Because we only require data annotated by author ID, this method should be able to easily port to new datasets or new domains (for example, written fanfiction) in principle. However, this needs to be validated in a broader range of settings, especially because many of the reward components use English models. Furthermore, we focused exclusively on ***automatic*** authorship attribution and verification, and did not explore how people with varying expertise in authorship analysis might manually assess the resulting texts.

Ethics Statement

Our work illustrates an improvement on automated obfuscation software, novelly applying a fine-tuning strategy for the task of general authorship obfuscation. The overarching goal for technologies that enable obfuscation on text data is to protect attribution of individuals or groups in cases where authorship metadata is scrubbed (e.g. using a pseudonym) (Afsaneh, 2021). At the same time, powerful obfuscation tools could be used to threaten, cyberbully, or otherwise endanger other individuals without accountability or fear of retribution.

Additionally, though this work does not explicitly target author style imitation, we acknowledge that obfuscation can cause existing identification tools to mis-attribute authorship to unsuspecting people. Because these identification tools have shown to be incredibly accurate for authorship retrieval, in scenarios where these users are unaware that texts have been modified, mis-attribution is a serious concern (Altakrori et al., 2022).

The Reddit dataset used as training data in our work is licensed as CC-BY-4.0 protocol, which stipulates that publicly released data will be used exclusively for research purposes. All pre-trained models used in this work are publicly available on HuggingFace[5], and we ensured that the research methodologies described in this work are aligned with licensing permissions. All methods and models described in this work are for research purposes, and are not intended for commercial use.

We manually reviewed our evaluation data to ensure that PII or personal entity identifiers were masked. We did not scrub swear words. For the human evaluation, we manually reviewed outputs and removed any candidates with hate speech.

References

Rigot Afsaneh. 2021. Why Online Anonymity Matters.

Malik Altakrori, Thomas Scialom, Benjamin C. M. Fung, and Jackie Chi Kit Cheung. 2022. A multifaceted framework to evaluate evasion, content preservation, and misattribution in authorship obfuscation techniques. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2391–2406, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695, Hong Kong, China. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108, Florence, Italy. Association for Computational Linguistics.

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

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Chris Emmery, Enrique Manjavacas Arevalo, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 984–996, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):214–217.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In

[5] https://huggingface.co/

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7546–7562, Singapore. Association for Computational Linguistics.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13:111–117.

Shankar Iyer, Nikhil Dandekar, and Korenl Csernai. 2017. First quora dataset release: Question pairs. *Journal of Machine Learning Research*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 444–451, Sydney, Australia. Association for Computational Linguistics.

Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. The case for being average: A mediocrity approach to style masking and author obfuscation - (best of the labs track at clef-2017). In *Conference and Labs of the Evaluation Forum*.

Yashwant Keswani, H. Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. In *Conference and Labs of the Evaluation Forum*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. 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 6365–6378, Online. Association for Computational Linguistics.

Alex Leavitt. 2015. "This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 317–327, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Tatiana Litvinova. 2020. Stylometrics features under domain shift: Do they really "context-independent"? In *Speech and Computer*, pages 279–290, Cham. Springer International Publishing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick. 2021. Style pooling: Automatic text style obfuscation for improved classification fairness. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2009–2022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning. *ArXiv*, abs/2212.08986.

Anselmo Peñas and Alvaro Rodrigo. 2011. A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1424, Portland, Oregon, USA. Association for Computational Linguistics.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* *and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2017. A4nt: Author attribute anonymity by adversarial training of neural machine translation. In *USENIX Security Symposium*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556.

Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2022. Overview of the Authorship Verification Task at PAN 2022. In *CLEF 2022 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anthony Henry Triggs, Kristian Møller, and Christina Neumayer. 2021. Context collapse and anonymity among queer Reddit users. *New Media & Society*, 23(1):5–21.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *SIGKDD Explor.*, 25(1):1–18.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4453–4460. AAAI Press.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

8 Appendix

A BLOG Evaluation Results

We start with 400 "needle" BLOG (Schler et al., 2006) snippets from 200 authors (2 comments per author). In our evaluation set, there are 211 words per BLOG snippet with 2 snippets per author, versus 37 words per REDDIT comment with 16 comments per author. The attribution evaluation for this data is as described in the REDDIT setting. We include the BLOG dataset in our evaluation to show that the proposed models operate well out-of-domain, and on slightly longer text lengths. Results are reported in Table 3.

For our BLOG sample, the haystack is smaller than that of the REDDIT sample. We observe that the LUAR attributor performs well out-of-domain at 51% R@8, 41% MRR, also leaving room to assess impact of the included obfuscation methods.

	Privacy				Meaning	Soundness	
	Attribution		Verification				
	R@8 ↓	MRR ↓	CNG c@1 ↓	LUAR ↑	SBERT ↑	COLA ↑	LenRatio
Baselines							
Copy	51.0	41.0	89.4	0.0	100.0	100.0	100.0
`Normalizer`	14.0	12.0	89.4	19.1	99.4	84.2	98.2
`Stylo`	10.0	6.0	**90.6**	25.6	91.8	34.0	109.4
`Round-trip MT`	34.0	23.0	89.3	13.4	**92.9**	84.0	98.8
`LUAR-scored RTMT`	32.0	23.0	88.9	15.5	90.7	**85.3**	98.9
`BART-Para`	**8.0**	6.0	85.3	**35.0**	76.5	62.0	71.4
`DIPPER 20L, 200`	12.0	10.0	89.5	31.3	92.3	76.4	108.5
`DIPPER 60L, 600`	10.0	**5.0**	89.7	33.9	90.6	74.5	110.2
Keep It Private Models							
`KiP-BART-Para`	4.0	5.0	87.6	38.9	**69.9**	**66.0**	126.3
`KiP-DIPPER`	**1.0**	**0.0**	**86.1**	**44.7**	64.9	63.5	163.9

Table 3: Confirming the same findings from Table 2 on the out-of-domain `BLOG` dataset, we observe strong attribution performance from the baseline paraphrasers `BART-Para` and `DIPPER`, and stronger attribution performance from the KiP models. However, we see limited defense against the `VERIF_CNG` attack in this domain.

B Human Evaluation Setup

As it is difficult for humans to identify authorship signals in natural language texts, we focused our human evaluation on assessing **paraphrase validity** to ensure quality of system generations.

Figure 4 shows the annotation instructions, and Figure 5 and Figure 6 show sample annotation screens. We recruited in total 27 participants. Each participant was paid complying with local wage standards. Each input-output pairing was annotated 3 times, and we took the majority vote as the score for the pairing.

C Reward Component Ablation

Reward Components To better understand trade-offs across evaluation dimensions, we measure the impact of reward components on the results (Table 4). We confirm that the meaning preservation, soundness and privacy components of the reward are all necessary. As expected, when we remove the LUAR-based reward leads to better privacy evaluation at the expense of SBERT and for `BART-para`, CoLA, and vice versa (rows 3 and 7 compared to rows 1 and 5), which leads to better SBERT scores at the expense of LUAR. caption

D BLOOM Prompts

As with all language models used in this work, we pulled language models from model repositories on the Hugging Face[6] platform. For our prompting experiments, we used BLOOM-7b [7]. We followed the procedures listed out in (Patel et al., 2022) to prompt BLOOM into rewriting the given `REDDIT` comments in a neutral style. We set default decoding parameters $\texttt{temperature} = 0.7, top_p = 1.0$, and prompt with:

> Passage: {Input Reddit comment}
> Paraphrase the passage in a simple neutral style.
> Rewrite:

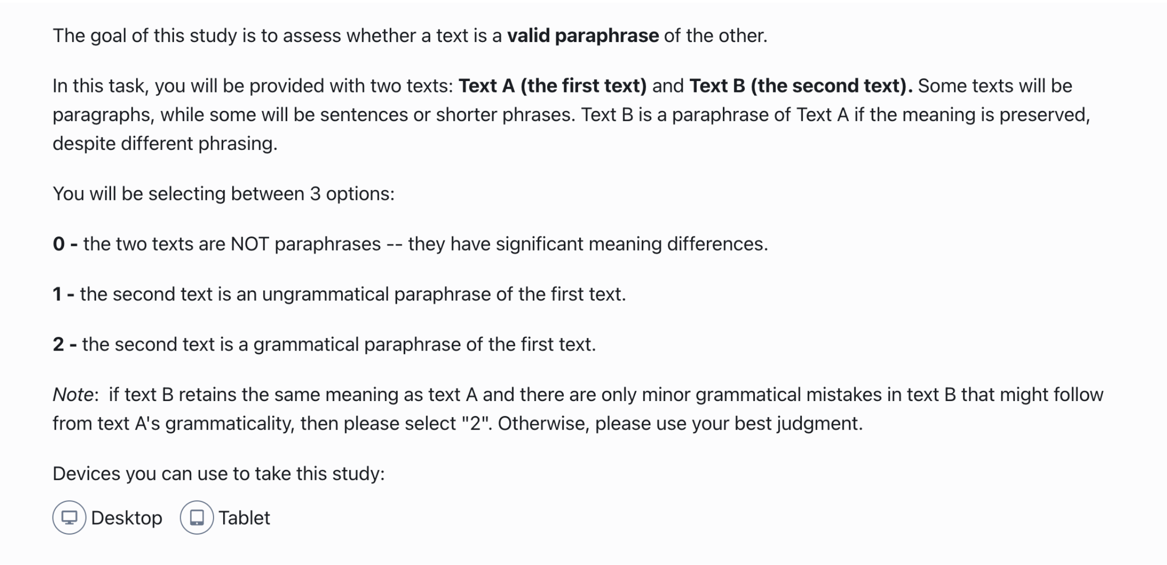
and allow BLOOM to generate until the <eos> token. We treat the model generation as the obfuscated text. We did not do any decoding parameter experiments.

E Sample of obfuscation outputs

More sample outputs can be found in Table 5.

[6] https://huggingface.co/

[7] https://huggingface.co/bigscience/bloom-7b1



The goal of this study is to assess whether a text is a **valid paraphrase** of the other.

In this task, you will be provided with two texts: **Text A (the first text)** and **Text B (the second text).** Some texts will be paragraphs, while some will be sentences or shorter phrases. Text B is a paraphrase of Text A if the meaning is preserved, despite different phrasing.

You will be selecting between 3 options:

0 - the two texts are NOT paraphrases -- they have significant meaning differences.

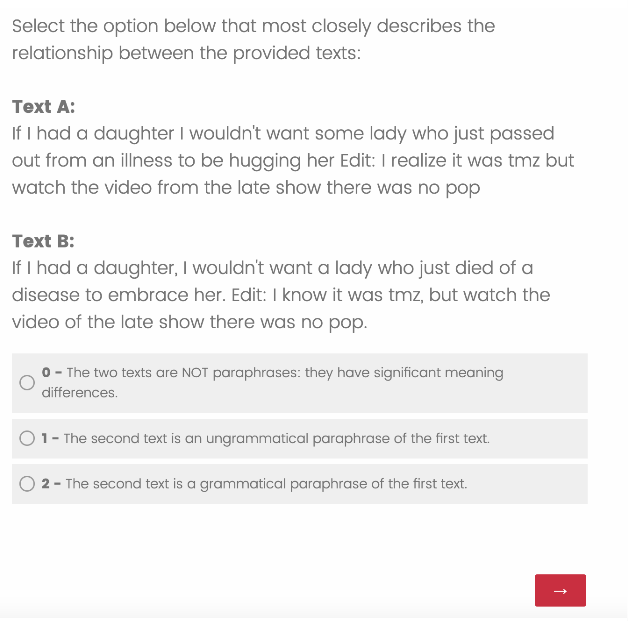
1 - the second text is an ungrammatical paraphrase of the first text.

2 - the second text is a grammatical paraphrase of the first text.

Note: if text B retains the same meaning as text A and there are only minor grammatical mistakes in text B that might follow from text A's grammaticality, then please select "2". Otherwise, please use your best judgment.

Devices you can use to take this study:

Figure 4: Instructions given to study participants. 27 English-fluent participants were recruited via Prolific.



Select the option below that most closely describes the relationship between the provided texts:

Text A:
If I had a daughter I wouldn't want some lady who just passed out from an illness to be hugging her Edit: I realize it was tmz but watch the video from the late show there was no pop

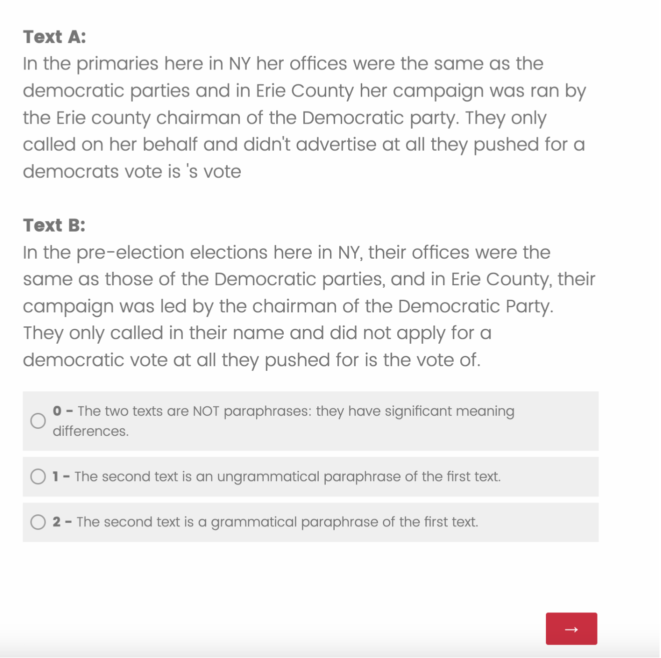
Text B:
If I had a daughter, I wouldn't want a lady who just died of a disease to embrace her. Edit: I know it was tmz, but watch the video of the late show there was no pop.

○ **0 -** The two texts are NOT paraphrases: they have significant meaning differences.

○ **1 -** The second text is an ungrammatical paraphrase of the first text.

○ **2 -** The second text is a grammatical paraphrase of the first text.

Figure 5: A sample multiple-choice question given to annotators.



Text A:
In the primaries here in NY her offices were the same as the democratic parties and in Erie County her campaign was ran by the Erie county chairman of the Democratic party. They only called on her behalf and didn't advertise at all they pushed for a democrats vote is 's vote

Text B:
In the pre-election elections here in NY, their offices were the same as those of the Democratic parties, and in Erie County, their campaign was led by the chairman of the Democratic Party. They only called in their name and did not apply for a democratic vote at all they pushed for is the vote of.

○ **0 -** The two texts are NOT paraphrases: they have significant meaning differences.

○ **1 -** The second text is an ungrammatical paraphrase of the first text.

○ **2 -** The second text is a grammatical paraphrase of the first text.

Figure 6: Another sample multiple-choice question given to annotators.

Model	LUAR	SBERT	CoLa
Remove guardrails	0.632	0.378	0.383
Remove fluency	0.246	0.818	0.608
Remove meaning	0.265	0.677	0.756
Remove privacy	0.384	0.812	0.591
`KiP-BART-para`	0.426	0.771	0.615

Table 4: Impact of removing various components on LUAR distance, SBERT similarity and CoLA, an acceptability judgment.

Model	Output
Copy	1. It’s all that sustains her at this point. 2. It’s more of a reflection of the low enthusiasm of her supporters. People simply aren’t excited about <PERSON>. Obviously it doesn’t mean much in the real world you can’t be so literal 3. He’s the nimblest of all navigators 4. For ios 100% no, the permission settings is still saying that it tracks your movement only when the app is open.
`RT MT`	1. It is all that it supports at this time 2. It is more a reflection of the low enthusiasm of their supporters. People are simply not enthusiastic about <PERSON>. Obviously it doesn’t mean much in the real world you can’t be so literal 3. He is the cleverest of all navigators 4. For ios 100% No, the permission settings still say that it only tracks your movements when the app is open.
`Stylo`	1. It ’s all that lengthen or extend in duration or space her at this point 2. It ’s more in a reflection in the low enthusiasm in her supporters, people simply are n’t excited about < Person >; and obviously it does n’t mean much in the true international you ca not be so literal 3. "He ’s the nimblest in all the member of an aircrew who is responsible for the aircraft ’s course 4. For ios One hundred percent no, the permission settings is still saying that it tracks your movement only when the app is open.
`KiP-BART`	1. it’s all that sustains her at this point 2. it’s more of a reflection of the low enthusiasm of her supporters. people simply aren’t excited about <PERSON>. obviously it doesn’t mean much in the real world you can’t be so literal 3. he’s the nimblest of all navigators 4. for ios 100% no, the permission settings is still saying that it tracks your movement only when the app is open.
`KiP-BART-para`	1. At this point, it’s all that sustains her. 2. It’s more of a manifestation of low expectations from her supporters. people simply aren’t excited about about about ’PERSON’. it doesn’t mean much in the world, can be literal 3. He is the nimblest of all navigation leaders. 4. For ios 100% no. the permission settings says it only tracks Movement when The app Open
`KiP-DIPPER`	1. All that she has to support her at this point is this 2. This simply shows the low enthusiasm of her supporters. Obviously, you can’t be so literal in the real world. They are simply not excited about her. 3. He is the nimblest of all navigators 4. No, Ios is still saying it only tracks your movements when the app is open.

Table 5: More privatized Reddit comments from a subset of the explored obfuscation methods.