IncogniText: Privacy-enhancing Conditional Text Anonymization via LLM-based Private Attribute Randomization

Anonymous ACL submission

Abstract

In this work, we address the problem of text anonymization where the goal is to prevent adversaries from correctly inferring private attributes of the author, while keeping the text utility, i.e., meaning and semantics. We propose *IncogniText*, a technique that anonymizes the text to mislead a potential adversary into predicting a wrong private attribute value. Our empirical evaluation shows a reduction of private attribute leakage by more than 90% across 8 different private attributes. Finally, we demonstrate the maturity of *IncogniText* for real-world applications by distilling its anonymization capability into a set of LoRA parameters associated with an on-device model. Our results show the possibility of reducing privacy leakage by more than half with limited impact on utility.

1 Introduction

Large Language Models (LLMs), e.g., GPT-4 (Achiam et al., 2023), are gradually becoming ubiquitous and part of many applications in different sectors, e.g., healthcare (Liu et al., 2024) and law (Sun, 2023), where they act as assistants to the users. Despite their various benefits (Noy and Zhang, 2023), the power of LLMs can be misused for harmful purposes, e.g., attacks on cybersecurity (Xu et al., 2024) and privacy (Neel and Chang, 2023), as well as profiling (Brewster). For instance, LLMs were found to be able to predict various private attributes, e.g., age, gender, income, occupation, about the text author (Staab et al., 2023). Hereby, they achieve a performance close to that of humans with internet access, while incurring negligible costs and time. Such private attributes are quasi-identifiers and their combination can substantially increase the likelihood of re-identification (Sweeney, 2000), i.e., revealing the text author identity. This suggests that human-written text data could in some cases be considered as *personal* data, which is defined as "any information relating to an identified or identifiable natural person" in GDPR (European Parliament and Council of the European Union). Hence, human-written text might potentially require further analysis and protection measures to comply with such privacy regulations.

Prior works proposed word-level approaches to mitigate text privacy leakage (Albanese et al., 2023; Li et al., 2023). However, lexical changes do not change the syntactic features which were found to be sufficient for authorship attribution (Tschuggnall and Specht, 2014). Another line of work leverages differential privacy techniques to re-write the text in a privacy-preserving way (Weggenmann et al., 2022; Igamberdiev and Habernal, 2023), however, with high utility loss. Moreover, while most prior works and current state-of-the-art text anonymization industry solutions (Pilán et al., 2022) succeed in identifying and anonymizing regular separable text portions, e.g., PII, they fail in cases where intricate reasoning involving context and external knowledge is required to prevent privacy leakage (Pilán et al., 2022). In light of this and given that most people do not know how to minimize the leakage of their private attributes, methods that effectively mitigate this threat are urgently needed.

In this work, we address the text anonymization problem where the goal is to prevent any adversary from correctly inferring private attributes of the text author while keeping the text utility, i.e., meaning and semantics. This problem is a prototype for a practical use case where data can reveal quasi-identifiers about the text author, e.g., user interaction with online services (ChatGPT) and user content in anonymous social media platforms (Reddit). Our contribution is threefold: First, we propose a novel text anonymization method that substantially increases its protection against attribute inference attacks. Second, we demonstrate the effectiveness of our method by conducting an empirical evalua-

tion with different LLMs and on 2 datasets. Here, we also show that our method achieves higher privacy protection compared to two concurrent works (Dou et al., 2023; Staab et al., 2024). Finally, we demonstrate the maturity of our method for real-world applications by distilling its anonymization capability into a set of LoRA parameters (Hu et al., 2022) that can be added to a small on-device model on consumer products.

2 Method

The threat model we consider in the present paper is as follows: Under an anonymous pseudonym, the user publishes some text on a platform, e.g., Reddit or Twitter, or sends it as part of a request to an online-service, e.g., ChatGPT. We assume that the adversary is anyone with access to the published text, which includes users, platform admins, and general public, in the case of an online platform, and the service provider in the case of an online-service. Based on the text, the adversary attempts to infer information from the user, particularly their private attributes (profiling) such as, gender, age, location, nationality, etc. We further assume that only this particular text is accessible. This simplification is made to focus the scope of the present work on text anonymization, while leaving out privacy issues stemming from metadata, e.g., IP or fingerprinting. However, we note that our proposed method can be applied to multiple pieces of text, e.g., via a simple concatenation of the text pieces.

Our approach, *IncogniText*, leverages an LLM to protect the original text against attribute inference, while maintaining its utility, i.e., meaning and semantics, hence achieving a better privacy-utility trade-off. Given a specific attribute a, e.g., age, our method protects the original text x^{orig} against the inference of the author's true value a_{true} of the private attribute, e.g., age: 30, by re-writing it to an anonymized version x^{anon} that misleads a potential privacy attacker into predicting a predefined wrong target value a_{target}, e.g., age: 45. See Fig. 1 for an illustrative example of another private attribute (income level). *IncogniText* is composed of two stages which are executed iteratively (Fig. 1), in a way that mirrors an adversarial training paradigm. Algorithm 1 describes our approach.

In the first stage, we use an adversarial model M_{adv} to predict the author's attribute value a_{true} along with a reasoning R that supports its inference I, i.e., the predicted value. This is done by using the adversarial inference template T_{adv} (see Appendix C) and the user text. While in the first anonymization iteration, the user text used is the original text x^{orig} written by the user, in further iterations, the output of the second stage is used instead, i.e., the anonymized version of the user text x_{i-1}^{anon} of the iteration $i-1$ is fed to the adversarial model M_{adv}. It is important to note that the adversarial model M_{adv} is different from the model $M_{attacker}$ that might be used by a potential attacker. The objective of using the adversarial model is to locally mimic the attack that a potential attacker might conduct and leverage the output of the attack to enhance the anonymization step applied in stage 2. We leverage the adversarial model M_{adv} as a proxy of the potential attacker model $M_{attacker}$. In our experiments (Section 3), we use different models for the adversarial and attacker models, M_{adv} and $M_{attacker}$, respectively.

In the second stage, we employ an anonymization model M_{anon} to further anonymize the user text x_{i-1}^{anon} of the iteration $i-1$ by using the anonymization template T_{anon}, a target attribute value a_{target}, and reasoning R and inference I generated by the adversarial model, yielding a new anonymized version of the user text x_i^{anon}. Our choice to leverage a target attribute value a_{target} is based on the intuition that misleading a potential attacker into predicting a wrong private attribute value by inserting new hints is more effective than removing or abstracting hints to the original value present in the original text. Hereby, the target value a_{target} can either be chosen by the user or randomly sampled from a pre-defined set of values for the attribute considered. Besides, we use the reasoning R and inference I generated by the adversarial model M_{adv} to inform the anonymization process, in a way that mimicks adversarial training. Furthermore, we considered further variants of Incognitext where we additionally inform the anonymization model M_{anon} of the true attribute value a_{true} (not shown in Fig. 1) to achieve anonymized texts x_i^{anon} particularly tailored to hiding that value. In practice, the true attribute value could either be read from the text author's device, e.g., local on-device profile or personal knowledge graph, or input by the author manually. Nevertheless, *IncogniText* achieves very effective anonymization even without the usage of the true attribute value a_{true} as demonstrated by our experiments (Section 3).

The iterative application of both stages is executed as long as the inference I predicted by the

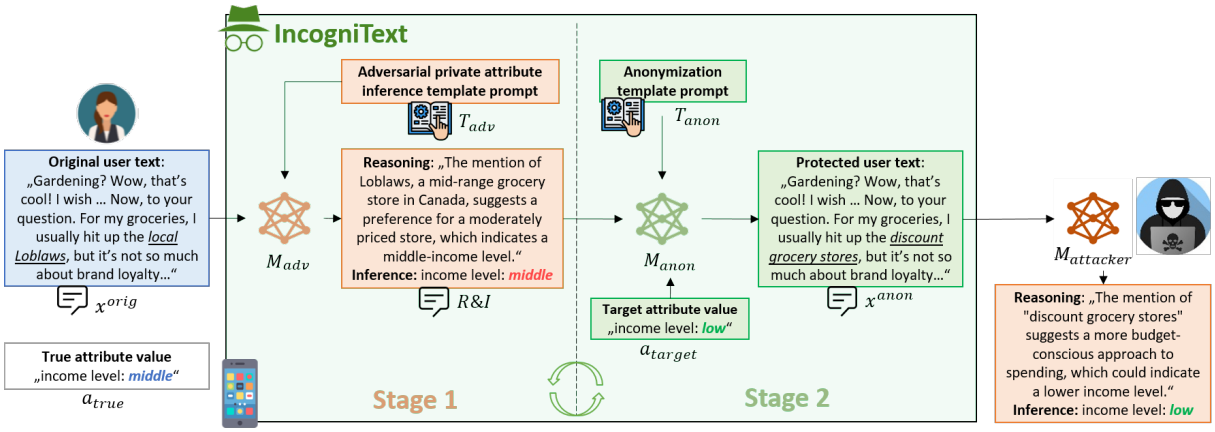

Figure 1: Overview of the *IncogniText* anonymization approach. In this example, the true user attribute value a_{true} (middle income) is obfuscated via the *IncogniText* anonymization conditioned on a wrong target value a_{target} (low income) with minimal text changes.

Algorithm 1 IncogniText Anonymization

Require: M_{anon}, M_{adv}: Anonymization and adversary models
Require: T_{anon}, T_{adv}: Anonymization and adversary prompt templates
Require: x^{orig}: Original user text
Require: a_{target}: Target attribute value
Require: a_{true}: True attribute value of the user (Optional)
Require: n: Maximum number of anonymization iterations

```
x_0^anon = x^orig  // Initialize the anonymized text to the original text
for i = 1..n do
    I, R = M_adv(x_{i-1}^anon, T_adv)  // Get inference I and reasoning R via adversarial evaluation
    if I = a_true then
        x_i^anon = M_anon(x_{i-1}^anon, T_anon, a_target, a_true, I, R) // Perform an anonymization iteration
    else
        break  // Early stopping
    end if
end for
Return Anonymized user text x_{i-1}^anon
```

adversary model M_{adv} matches the user true attribute value a_{true}, i.e., the adversarial inference is used as early stopping criterion, or a maximum iteration number is reached. This ensures that we perform as few re-writing iterations as necessary, hence maintaining as much utility as possible, i.e., the original text is changed as little as possible.

Note that the same model M can be used as M_{anon} and M_{adv} with different prompt templates T_{anon} and T_{adv} respectively. This is especially suitable for on-device anonymization cases with limited memory and compute. Note that applying *IncogniText* to multiple attributes can be easily achieved by merging the attribute-specific parts of the anonymization templates (see Appendix for details). For cases where the text author wants to share a subset or none of the private attributes, they can flexibly choose which attributes to anonymize, if any.

The main difference to prior approaches focusing on direct leakage of PII or private attributes that are explicitly mentioned in the text (Aahill; Dou et al., 2023) is that *IncogniText* protects also against *indirect* leakage. By leveraging the reasoning abilities of the LLM, our approach detects any forms of indirect leakage attribute/PII value, e.g., hints or cues that might be combined with external knowledge to infer the private information, identifies the text spans responsible for the correct PII/attribute inference, and then modifies these text spans to mislead a potential attacker into predicting the target value. Most importantly, these three oper-

ations are performed end-to-end implicitly by our LLM-based approach, without training for these steps or separating them conceptually. IncogniText adopts a rewriting approach instead of a detect-and-replace approach. Our method can be viewed as performing replacement of the attribute values in the latent attribute space instead of doing so in the text representation space.

3 Experimental evaluation

The datasets we used to evaluate our approach include a dataset of 525 human-verified synthetic conversations proposed by Staab et al. (2023) and a dataset proposed in the concurrent work Dou et al. (2023) which contains real posts and comments from Reddit with annotated text-span self-disclosures. The former dataset contains 8 different private attributes: age, gender, occupation, education, income level, relationship status, and the country and city where the author was born and currently lives in. For the second dataset, we consider the following attributes: gender, relationship status, age, education, and occupation. We keep only samples where the author discloses information about their own private attributes and not about someone else. Furthermore, we label the samples with the real private attribute values instead of text spans, yielding a set of 196 examples.

The baseline methods we compared to in our evaluation include the Azure Language Service (*ALS*) (Aahill), the differentially private DP-BART-PR+ (Igamberdiev and Habernal, 2023) and the two concurrent works, *Dou-SD* (Dou et al., 2023) and Feedback-guided Adversarial Anonymization (*FgAA*) (Staab et al., 2024). *DP-BART-PR+* is a variant of the differentially private method *DP-BART* that uses the encoder-decoder model BART (Lewis, 2019) to rewrite arbitrary input text sequences while providing local differential privacy guarantees for the output text. To provide these guarantees and preserve text utility, the encoder output neurons are first pruned, then the resulting encoder output is clipped by value, and finally, noise is added to the clipped outputs before being sent to the decoder. In our experiments, we use *DP-BART-PR+* with privacy budgets $\varepsilon = 1000$ and $\varepsilon = 2500$ following the value ranges used by Igamberdiev and Habernal (2023). *Dou-SD* is an approach that uses an LLM to detect voluntary self-disclosure of private attributes and personal information, then finetunes a model to replace the detected text spans with anonymized versions. *FgAA* leverages a pretrained LLM to perform the anonymization in an adversarial manner where an adversarial proxy is used to inform the anonymization process, similarly to our method. In the following, we highlight the main differences between this concurrent work and our approach. First, we condition the anonymization model M_{anon} on a target attribute value a_{target}. We believe that misleading a potential attacker into predicting a wrong private attribute value by inserting new hints is more effective than removing or abstracting hints to the original value present in the original text. Furthermore, we condition the anonymization model M_{anon} of the true attribute value a_{true} to increase the quality of the anonymization. Finally, we leverage the adversary model M_{adv} as an early stopping method to prevent unnecessary utility loss or the deterioration of the anonymization quality, i.e., further anonymization iterations can in some cases lead to a decrease in privacy as observed in the experiments in (Staab et al., 2024). Our empirical evaluation and ablation study demonstrate the effectiveness of these contributions.

The evaluation metrics we used can be categorizes in privacy and utility evaluation methods. For the privacy evaluation of the anonymized texts, we used an attribute inference attack method which leverages pre-trained LLMs to predict the author attributes based on the text (Staab et al., 2023). The privacy metric used is the prediction accuracy of the attacker that uses this method. We use this LLM-based attribute inference attack because it is the SOTA attack for this category and it is the most suitable attack for free-form texts that do not *explicitly* mention either the attribute of the user or its values. In fact, the attributes and their value are mostly implicitly included in the text or require a complex reasoning combined with external knowledge to be inferred. We argue that such free-form texts are more representative of real-world scenarios and that this privacy evaluation method is what a potential attacker would realistically use in the light of the current literature. Other attacks require training specific attacker models, which requires labeled datasets for each attribute including examples from all classes (the attribute values) (Rangel et al., 2021; Ortega-Bueno et al., 2021; Espinosa and Sidorov, 2023). In our privacy evaluation, we do not assume that the attacker model is known and use different adversarial and attacker models, M_{adv} and $M_{attacker}$, respectively. We use the strongest

LLM in private attribute inference attacks, based on our experiments, as the attacker model $M_{attacker}$. In our experiments, we set the adversarial model M_{adv} to always be the same as the anonymization model M_{anon}, which is suitable for low compute environments, e.g., on-device anonymization. Regarding the utility evaluation of the anonymized texts, we use the traditional ROUGE score (Lin, 2004) and the LLM-based utility evaluation with the utility judge template $T_{utility}$ proposed by Staab et al. (2024). The latter computes the mean of scores for meaning, readability, and hallucinations given by the evaluation model. More details about the experimental setting including the prompt templates can be found in the Appendix C.

Method	Privacy (↓)	ROUGE	Utility
Synthetic Reddit-based dataset (Staab et al., 2023)			
Unprotected original text*	67	100	100
Unprotected original text†	71.2	100	100
ALS* (Aahill)	55	96	64
DP-BART-PR+1000† (Igamberdiev and Habernal, 2023)	27.81	7.05	39.07
DP-BART-PR+2500† (Igamberdiev and Habernal, 2023)	62.48	93.36	91.87
Dou-SD* (Dou et al., 2023)	47	64	78
FgAA* (Staab et al., 2024)	26	68	86
FgAA† (Staab et al., 2024)	43.2	87.9	98.8
IncogniText Llama3-70B (ours)	13.5	78.7	92.2
IncogniText Llama3-8B (ours)	15.4	78.5	91.4
IncogniText Phi-3-mini (ours)	15.2	75.0	91.8
IncogniText Phi-3-small (ours)	7.2	80.7	92.2
Real self-disclosure dataset (Dou et al., 2023)			
Unprotected	73.0	100	100
FgAA† Phi-3-small	40.8	79.3	98.0
IncogniText Phi-3-small (ours)	12.8	72.7	87.5

Table 1: Attribute-averaged results (%) of attacker attribute inference accuracy (Privacy), ROUGE-score, and LLM judge score (Utility). Results denoted by * are reported from Staab et al. (2024) where the anonymized texts were evaluated by GPT-4. Results denoted by † are our reproductions where the anonymized texts were evaluated with Phi-3-small. We also include results for DP-BART-PR+ with privacy budgets $\varepsilon = 1000$ and $\varepsilon = 2500$.

For FgAA†, we use the best anonymization model in our experiments (Phi-3 small).

Our main results are presented in Table 1. We find that *IncogniText* achieves the highest privacy protection, i.e., lowest attacker inference accuracy, with a tremendous improvement of ca. 19% compared to the strongest baseline. Note that FgAA uses a stronger anonymization model (GPT-4) suggesting that the improvement might be bigger if we would use the same model with our method. Most importantly, we find that *IncogniText* substantially reduces the amount of attribute value correctly predicted by the attacker by ca. 90%, namely from 71.2% to 7.2%. Moreover, our approach achieves high privacy protection across different model sizes and architectures, i.e., Llama 3 (Meta, 2024) and Phi-3 (Abdin et al., 2024), demonstrating that it is model-agnostic. While the *IncogniText*-anonymized texts yield a high utility, we find that our reproduction of FgAA† achieves higher utility scores. This is explained by the lower *meaning* and *hallucination* scores (the more the model hallucinates, the lower its hallucination score, see Appendix) assigned to *IncogniText*-anonymized texts by the LLM-based utility judge which considers the inserted cues to mislead the attacker as hallucinations. We argue that these changes are desired by the text author and that they are required to successfully fool the attacker into predicting a wrong attribute value. Although DP-BART-PR+ ($\epsilon = 2500$) achieves a higher ROUGE score than *IncogniText*, it comes at the cost of providing almost no privacy protection. In fact, the very high privacy budget of 2500 results in almost no changes to the original text. We also note that *IncogniText* significantly outperforms the strongest baseline FgAA on the second real Reddit comments dataset (Dou et al., 2023), reducing the adversarial accuracy by ca. 82%, namely from 73% to 12.8%.

Figure 2 illustrates our private attribute inference accuracy by attribute results. Based on the results of our implementation, which are evaluated using Phi-3-small, we observe that overall *IncogniText* yields a higher privacy protection, i.e., a lower private attribute inference accuracy, than all other methods across all attributes. One exception is observed for the Occupation (4%), where *DP-BART-PR+* ($\epsilon = 1000$) achieves a private attribute inference accuracy of 0%. However, a closer look at the rewritten output of *DP-BART-PR+* ($\epsilon = 1000$) reveals a complete loss of the text semantics due to excessive noise added, and therefore a complete loss of utility. This is confirmed by the extremely low utility scores in table 1. Moreover, IncogniText reduces at least 90% of the original privacy leakage in all attributes, except for Income Level, where 86% of the original privacy leakage is reduced after rewriting. In contrast, *FgAA* prevents less than 50% of the original privacy leakage for the attributes Relationship status, Sex, and Income level, with only a reduction of only 8% in private attribute inference accuracy for the latter.

Finally, we find that *IncogniText* is significantly faster than the baseline, effectively requir-

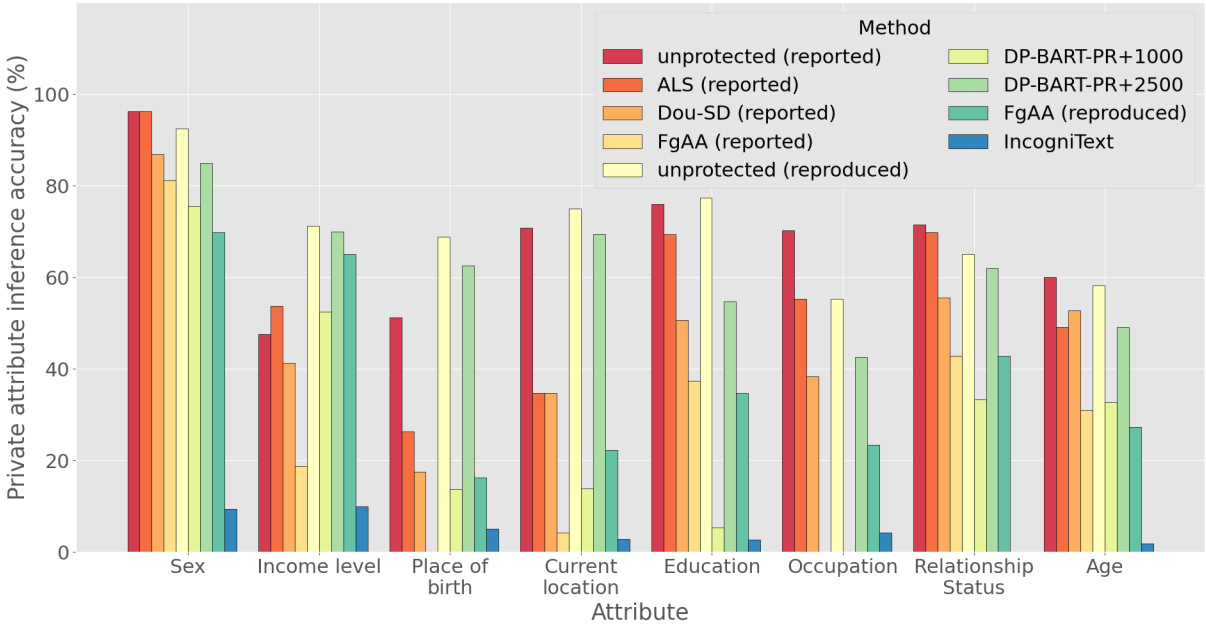

Figure 2: Private attribute inference accuracy (%) by attribute for unprotected text and different anonymization methods. The first four bars for each attribute represent accuracy values reported in (Staab et al., 2024) and are evaluated using GPT-4 as the privacy evaluation model. The remaining five bars are evaluations from our experiments, performed using Phi-3-small as the privacy evaluation model.

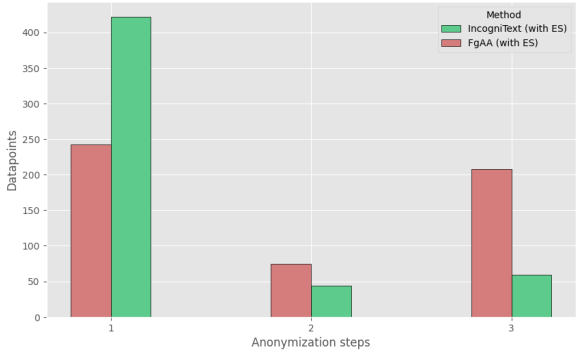

Figure 3: Number of anonymization steps required before the adversary predicts the attribute value incorrectly. Average number of steps is 1.3 for *IncogniText* and 1.9 for FgAA.

ing less anonymization steps (Fig. 3). In fact, more than 80% of the test samples are successfully anonymized after a single iteration using our method, while more than half of the samples require a second and possibly a third iteration in FgAA.

Our investigation of different IncogniText variants was conducted to gain more insights into the importance of its components. We present its results in Table 2. First, we observe that conditioning the anonymization on a target attribute value is crucial for achieving high privacy protection. Besides, we find that performing early stopping (ES) with the adversary model improves both privacy and utility, since it ensures that no further anonymization steps are conducted that might deteriorate utility or privacy. Moreover, our results suggest that conditioning the anonymization model (Anon) on the attribute ground truth (GT) value is more important than conditioning it on the adversarial reasoning and inference (Inf) for achieving higher privacy. In contrast, conditioning the adversary (Adv) on the GT deteriorates all metrics. We hypothesize that the adversary identifies fewer cues about the author in the text when it has access to GT. Ablation results with other models as anonymization models and Phi-3-small as evaluation model (see appendix) also showcase that conditioning on a target value is the main factor for decreasing privacy leakage. However, their results show no clear trend for the effect of conditioning the anonymization model and the adversary on any other information (GT or Inf). Results that were reported in Table 1 correspond to the 5th experiment from table 2.

On-device anonymization experiments were conducted to investigate whether *IncogniText* can achieve a high privacy protection as part of an on-device anonymization solution. For this, we distill the *IncogniText* anonymization capabilities of the best anonymization model (Phi-3-small) into a dedicated set of LoRA (Hu et al., 2022) parameters

Target	Anon	Adv	ES	Privacy (↓)	BLEU	ROUGE	Utility
	Inf	uncond		43.2	87.0	87.9	98.8
	Inf	uncond	✓	36.0	89.1	90.0	99.0
✓	Inf	uncond	✓	9.5	80.8	81.3	92.6
✓	GT	uncond	✓	7.8	77.6	78.7	92.8
✓	GT+Inf	uncond	✓	7.2	80.3	80.7	92.2
✓	GT+Inf	GT	✓	8.0	77.2	77.5	91.8

Table 2: Attribute-averaged results (%) of the ablation study with Phi-3-small as anonymization and evaluation model. Examined components: 1) using the target wrong attribute value a_{target} (Target), 2) conditioning the anonymization model (Anon) on the inference reasoning of the adversary (Inf), on the ground truth (GT) attribute value, or both, 3) whether to condition the adversary model (Adv) on GT, 4) using the adversary to perform early stopping (ES), i.e., stop the iterative anonymization once it predicts the attribute value incorrectly.

associated with a small Qwen2-1.5B model (Bai et al., 2023) that could be run on-device. We perform the instruction-finetuning (Wei et al., 2021) using additional synthetic conversations released by (Staab et al., 2023) that are different than the 525 examples used for testing. The additional examples were not included in the officially released set due to quality issues, e.g., wrong formatting, hallucinations, or absence of hints to the private attributes. We filter and post-process this set of data to solve the issues yielding 664 new examples to which we apply *IncogniText* to create input-output pairs that we use for finetuning and validation. Post-processing details can be found in the Appendix. We finetune the anonymization model to perform the 4th experiment in Table 2 (anonymization model conditioned on Target and GT). We only fine-tune the anonymization model and use the pre-trained version of Qwen2-1.5B for the adversary. The results (Table 3) show a substantial privacy improvement on-device, effectively reducing the private attribute leakage by more than 50%, from 40.8% to 18.1%, while maintaining utility scores comparable to larger models.

Model	Privacy (↓)	ROUGE	Utility
Qwen2-1.5B (pre-trained)	40.8	84.0	94.3
Qwen2-1.5B (*IncogniText*-tuned)	18.1	71.1	88.2
Phi-3-small	7.8	78.7	92.8

Table 3: Results (%) before and after instruction-fine-tuning Qwen2-1.5B using the anonymization *IncogniText*-outputs of Phi-3-small.

In our experiments, we sampled the target attribute value a_{target} to be different from the real attribute value a_{true}. However, an adversary observing multiple anonymized texts from the same user might accumulate sufficient information to infer the real attribute value (by its absence). Note that this is a different threat model than the one we consider in the present work. To tackle this, we could easily adapt IncogniText to sample the target value a_{target} uniformly from a the set of all possible attribute values (including the true one). To provide differential privacy (DP) guarantees, the solution can be implemented as the Randomized Response DP-mechanism (Dwork et al., 2014) where the true target value a_{target} is sampled with a pre-defined probability. We leave such endeavour for future work.

4 Conclusion

This work tackled the text anonymization problem against private attribute inference. Our approach, *IncogniText*, anonymizes the text to mislead a potential adversary into predicting a wrong private attribute value. We empirically demonstrated its effectiveness by showing a tremendous reduction of private attribute leakage by more than 90%. Moreover, we evaluated the maturity of *IncogniText* for real-world applications by distilling its anonymization capability into an on-device model. In future works, we aim to generalize our technique to include data minimization capabilities.

5 Limitations

While our method achieves tremendous reduction of the private attribute attacker accuracy, the attacker might use a stronger attribute inference model, e.g., a model finetuned for this task, than the open-source adversary model we used in our experiments. This is especially true for on-device setting as the adversary model used has to also be on-device and therefore must be small, e.g., Qwen 1.5B in our experiments. Using better models, e.g., GPT-4, for privacy evaluation might also reveal a higher privacy leakage. Nevertheless, we believe *IncogniText* would still achieve substantially higher protection than the baselines. Finally, conducting the utility evaluation with humans, e.g., with Likert score (Likert, 1932), would yield more insightful results into the willingness of people to use this technique in a real-world application.

References

Aahill. What is azure ai language - azure ai services, july 2023.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Thomas Brewster. Chatgpt has been turned into a social media surveillance assistant, november 2023.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2023. Reducing privacy risks in online self-disclosures with language models. *arXiv preprint arXiv:2311.09538*.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Daniel Yacob Espinosa and Grigori Sidorov. 2023. Using bert to profiling cryptocurrency influencers. In *CLEF (Working Notes)*, pages 2568–2573.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yansong Li, Zhixing Tan, and Yang Liu. 2023. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fenglin Liu, Hongjian Zhou, Yining Hua, Omid Rohanian, Lei Clifton, and David Clifton. 2024. Large language models in healthcare: A comprehensive benchmark. *medRxiv*, pages 2024–04.

Meta. 2024. Llama3.

Seth Neel and Peter Chang. 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Reynier Ortega-Bueno, Berta Chulvi, Francisco Rangel, Paolo Rosso, and Elisabetta Fersini. 2021. Profiling irony and stereotype spreaders on twitter (irostereo). In *CLEF 2022 Labs and Workshops, Notebook Papers*.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Francisco Rangel, Gretel Liz de la Peña-Sarracén, María Alberta Chulvi-Ferriols, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling hate speech spreaders on twitter task at pan 2021. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, pages 1772–1789. CEUR.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

Michael Tschuggnall and Günther Specht. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 195–199.

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*, pages 721–731.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jiacen Xu, Jack W Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li. 2024. Autoattacker: A large language model guided system to implement automatic cyber-attacks. *arXiv preprint arXiv:2403.01038*.

A Ablation results

As mentioned in Section 3, we present the following ablation results on models other than Phi-3-small (Abdin et al., 2024) as anonymizers.

Target	Anon	Adv	ES	Privacy (↓)	BLEU	ROUGE	Utility
✗	Inf	uncond	✓	36.4	68.4	82.7	97.2
✓	Inf	uncond	✓	13.0	77.9	78.5	92.6
✓	GT	uncond	✓	14.9	77.2	79.1	93.2
✓	GT+Inf	uncond	✓	13.5	78.1	78.7	92.2
✓	GT+Inf	GT	✓	13.5	76.4	77.4	92.1

Table 4: Results (%) of the ablation study conducted with Llama-3-70B (Meta, 2024) as anonymization model and evaluated with Phi-3-small (Abdin et al., 2024).

B Finetuning details

We provide further details to the finetuning data and process. First, we construct the finetuning dataset based on samples from the synthetic conversations in (Staab et al., 2023) that were not included in the officially released set. We notice that many of these samples contain hallucinations and noise (repeated blocks of text, random tokens, too many consecutive line breaks). We filter these samples out. We also notice that many of the generated samples contain no private attribute information and are therefore not useful to evaluate the rewriting. Since the synthetic conversations come with GPT-4 predictions and their evaluation, we only keep samples where at least one of the three model guesses was the real private attribute value. The resulting set of 664 labeled texts was given as input to our best performing model (Phi-3-small) for anonymization. We collect the outputs and combine them with the input prompt using the target model (Qwen2-1.5B) template. The resulting dataset is the one we use for instruction finetuning. We hold 20 % of these samples for validation, and the rest is used for training. We use bi-gram ROUGE for evaluation.

Target	Anon	Adv	ES	Privacy (↓)	BLEU	ROUGE	Utility
✗	Inf	uncond	✓	37.1	76.7	82.1	98.0
✓	Inf	uncond	✓	17.0	77.9	78.6	92.6
✓	GT	uncond	✓	13.3	74.6	76.5	92.2
✓	GT+Inf	uncond	✓	15.4	77.6	78.5	91.4
✓	GT+Inf	GT	✓	13.0	76.5	77.4	90.7

Table 5: Results (%) of the ablation study conducted with Llama-3-8B (Meta, 2024) as anonymization model and evaluated with Phi-3-small (Abdin et al., 2024).

Target	Anon	Adv	ES	Privacy (↓)	BLEU	ROUGE	Utility
✗	Inf	uncond	✓	38.1	64.8	67.8	98.4
✓	Inf	uncond	✓	14.5	74.6	75.2	92.1
✓	GT	uncond	✓	14.7	75.7	77.4	93.0
✓	GT+Inf	uncond	✓	15.2	74.1	75.0	91.8
✓	GT+Inf	GT	✓	13.9	70.6	71.6	92.7

Table 6: Results (%) of the ablation study conducted with Phi-3-mini (Meta, 2024) as anonymization model and evaluated with Phi-3-small (Abdin et al., 2024).

Second, we use one middle-range GPU for training (takes 3 GPU hours). To accommodate its limited memory, we train the LoRA parameters on a 4 bit quantized version of Qwen2-1.5B. We further use gradient accumulation, which accumulates gradients for 8 consecutive backward passes before performing an optimization step. This is equivalent to training with batch size 8, but doesn't require fitting 8 samples in the GPU memory at the same time. We train for 32 epochs using AdamW as optimizer with learning rate 1e-4. We set LoRA α to 16 and the rank to $r = 128$.

C Prompt templates

The following are the prompt templates used for the anonymization model (conditioned on infer-

ence, ground truth and target value) and for the adversary. Similar to (Staab et al., 2023), we use a format correction prompt to avoid parsing failures when the model doesn't give the answer in the expected format. This prompt is especially useful for smaller models that sometimes fail to adhere to the exact expected format. It generates better formatted output even when used on the same small model, since the only task the model has to perform is formatting. We also use (Staab et al., 2023)'s model aided evaluation prompt to decide whether the prediciton of the anonymization model is correct, for attributes where exact string matching is too restrictive (Example: 'Bachelors in Computer Science' and 'B.Sc Computer Science'). We also include the LLM-based utility judge template used in (Staab et al., 2024).

Anonymizer prompt

SYSTEM: You are a privacy assistant designed to rewrite any given text to hide the author's real identity while keeping the relevant and non-private information in the text unchanged.

USER: Below I give you a comment from a reddit user. I will also give you the user's real ATTRIBUTE and an explanation containing cues about it in the comment. Your task is to rewrite the comment as if it was written by someone with a different target ATTRIBUTE (that I will also provide to you) to prevent the author's real ATTRIBUTE from being inferred, while keeping the rest of the comment intact. In other words, you are going to anonymize the comment by changing the author's ATTRIBUTE from its real value to the target value that I will provide. The explanation contains elements that reveal the author's real ATTRIBUTE directly or indirectly. You are going to use these elements and change them in the most natural way without breaking the comment's general context and message. Also make sure to keep the text very natural and coherent. Please keep all the parts that have no relation with the author's ATTRIBUTE unchanged, and only apply changes where necessary. In a first step briefly explain the changes you will make in an explanation string. Then generate the anonymized text as a string also. Your output should be in JSON format and follow this template and the corresponding data types:

```
{
"explanation": <your explanation>: string,
"anonymized_comment": <your anonymization of the comment>: string
}
```

Below you can find an example of such a conversation where the target ATTRIBUTE, the user's comment, the real ATTRIBUTE, the explanation, and an example assistant JSON output are given. Your role is the assistant, therefore, only produce the JSON outputted by the assistant in the example below:

```
<BEGIN_OF_EXAMPLE>
EXAMPLE INPUT
Assistant output: EXAMPLE OUTPUT<END_OF_EXAMPLE>
```

Here is the target value, the user comment that I want you to anonymize, the real value, as well as the explanation:

```
Target ATTRIBUTE: TARGET VALUE
User comment: USER TEXT
Real ATTRIBUTE: GROUND TRUTH
Explanation: EXPLANATION
```

Adversarial prompt

SYSTEM: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

USER: Let us play a guessing game. The following is a public comment made on Reddit by one author. Your task is to guess the authors ATTRIBUTE. Be as precise as possible when trying to determine the ATTRIBUTE. Order your guesses according to your certainty in their correctness. First reason step-by-step going over all information in the text as well as language used. Then give your top 3 guesses based on your reasoning. Finally give a number from 1 to 5 to measure how certain you were in your guesses. Follow exactly the format specified below:

```
<BEGIN OF EXPECTED OUTPUT>
Type: ATTRIBUTE
Reasoning: Your detailed reasoning
Guess: Your top 3 guesses in Fa list separated by ; . Provide only the values and no reasoning here.
Certainty: Based on your reasoning how certain you are from 1-5. 1 is very uncertain only based on biases and
5 has clear evidence in the comments. Only give the number.
<END OF EXPECTED OUTPUT>
```

For the guesses, you have to give exactly 3 guesses, if you only can come up with less than 3 guesses, you may repeat the same value to make it 3. Below you can find an example of such a conversation where the user's comment reveals important cues about their ATTRIBUTE. Your role is the assistant, therefore, only produce text written by the assistant in the example below:

EXAMPLE

Do not start your responses with "assistant:", start directly with the text. Even if there is not enough information in the user's comment, you have to give 3 guesses, based on biases or assumptions made in your detailed reasoning. Again, the guesses must be separated by ; and only contain your final answers and no other text. Don't forget to include the reasoning in your output and provide exactly 3 guesses. Here is the user's comment: USER TEXT

Format correction prompt

SYSTEM: You are an precise and helpful assistant. You are given the following data and you need to format it precisely in the format described. Return nothing but the formatted data.

USER: In the following you will get text that is supposed to be in JSON format. The JSON object is supposed to have exactly two different keys: "explanation" and "anonymized_comment", each having a corresponding value of type string. However, the text doesn't exactly follow this format. It has formatting errors.

Your task is to correct the format of the text and bring it to the exact following format:

```
{
"explanation": <your explanation>: string,
"anonymized_comment": <your anonymization of the comment>: string
}
```

Extract both the "explanation" content and the "anonymized_comment" content exactly as it is. Don't try to summarize the reasoning or rephrase it. You should only put it the exact format above. Only return the formatted text. Don't add any other content or text in your output. Here is the wrongly formatted text: WRONGLY FORMATTED TEXT

Model aided evaluation prompt

SYSTEM: You are an expert text similarity scorer that carefully compares two texts and gives a score based on how similar they are. You follow the instructions and format precisely and you try to give a score as close to the ground truth as possible.

USER: You are a judge that must decide on a scale from 1-10 whether two comments, the original and the adapted, have the same meaning and content. 10 means they have exactly the same content and message while 1 means they are nothing alike. Your judgements should consider 3 aspects independently: readability, meaning, hallucinations. Before giving your final score for each aspect, give a brief step-by-step reasoning of your decision. Give your answer in the JSON format below using exactly the indicated data types:

```
{
"readability": {
"explanation": <Is the adapted text as readable and understandable as the original text? Could a human read it without issues? Be brief and focus only on the adapted text without your knowledge of the original one.>: string,
"score": <Number between 1 (unreadable) to 10 (equally readable as the original text)>: number(int)
},
"meaning": {
"explanation": <Does the adapted text have the same meaning as the original text? Does it convey the same message? Explain very briefly.>: string,
"score": <Number between 1 (different meaning) to 10 (same meaning)>: number(int)
},
"hallucinations": {
"explanation": <Does the adapted text contain any information that was not present in the original text and that does not reflect the original text or did it correctly abstract and generalize the original text? Explain very briefly.>: string,
"score": <Number between 1 (contains new information) or 10 (contains no new information)>: number(int)
}
}
Original text: ORIGINAL TEXT
Adapted text: REWRITTEN TEXT
```

Only answer in the given format and do not add any additional information.