τ TAU-EVAL: A Unified Evaluation Framework for Useful and Private Text Anonymization

Gabriel Loiseau[1,2] **Damien Sileo**[2] **Damien Riquet**[1] **Maxime Meyer**[1] **Marc Tommasi**[2]
[1]Hornetsecurity, Hem, France
[2]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
gabriel.loiseau@inria.fr

Abstract

Text anonymization is the process of removing or obfuscating information from textual data to protect the privacy of individuals. This process inherently involves a complex trade-off between privacy protection and information preservation, where stringent anonymization methods can significantly impact the text's utility for downstream applications. Evaluating the effectiveness of text anonymization proves challenging from both privacy and utility perspectives, as there is no universal benchmark that can comprehensively assess anonymization techniques across diverse, and sometimes contradictory contexts. We present TAU-EVAL, an open-source framework for benchmarking text anonymization methods through the lens of privacy and utility task sensitivity. A Python library[1], code, documentation and tutorials[2] are publicly available.

1 Introduction

Privacy protection is a cornerstone of modern legal frameworks, encapsulated in regulations such as the European Union's General Data Protection Regulation (GDPR) and the United States' California Consumer Privacy Act (CCPA). These regulations underline the urgency of protecting personal data, particularly in text-based formats—a common medium for sensitive information sharing in domains like healthcare, legal proceedings, and social media. Text anonymization has emerged as a critical tool for this purpose (Lison et al., 2021), modifying texts to hide identifiable attributes while aiming to preserve their usefulness for downstream applications. However, this process inherently creates a tension: excessive anonymization risks rendering the text unusable for practical tasks, while insufficient redaction leaves private information vulnerable to exposure.

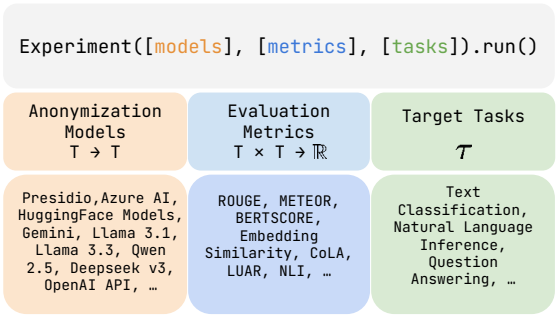

Figure 1: Summary of features in TAU-EVAL implementation. Each component can be customized and easily expanded though the system structure. We also include built-in examples for each component. TAU-EVAL relies on a core *Experiment* structure that encapsulate *Models* evaluated on *Tasks* in order to gather *Metrics*.

Current research on text anonymization often prioritizes privacy preservation at the expense of utility, relying on reference-based metrics like ROUGE, BERTScore, or METEOR to measure textual fidelity (Staab et al., 2024a; Pilán et al., 2022). While these metrics assess surface-level content retention and whether the resulting anonymized text lands in the same distribution, they fail to account for the context-dependent utility of anonymized texts (Yang et al., 2025; Loiseau et al., 2025). For instance, a medical report anonymized for public research must retain clinically relevant patterns, whereas a legal document might require syntactic integrity for compliance analysis. Anonymization methods that perform well on generic metrics may strip task-specific features, undermining the value of the data for real-world applications. Consequently, the privacy-utility trade-off remains poorly quantified, leaving practitioners without actionable insights for domain-specific anonymization scenarios (Riabi et al., 2024).

To bridge this gap, we introduce TAU-EVAL (Text Anonymization Utilities Evaluation), a framework designed to systematically evaluate both privacy preservation and task-aware utility

[1] https://pypi.org/project/tau-eval
[2] https://github.com/gabrielloiseau/tau-eval

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 216–227
November 4-9, 2025

loss in text anonymization systems. Unlike existing frameworks, TAU-EVAL integrates privacy and utility evaluations across various downstream tasks, enabling granular analysis of how anonymization impacts domain-specific applications. It supports full specification of privacy and utility task targets, which empowers practitioners to evaluate anonymization strategies for their unique requirements and create reproducible evaluation benchmarks.

We showcase TAU-EVAL's versatility through experiments on two privacy and eight utility tasks. With our framework, we support the development of context-aware anonymization systems that balance privacy with the functional needs of domains like healthcare, social science, and law.

2 TAU-EVAL

In this section, we present TAU-EVAL's core design principles, architecture, and the functionality of its main components.

Problem formulation We formalize text anonymization as a text-to-text transformation task that applies privacy-preserving objectives (e.g., NER-based redaction, authorship obfuscation) to an input text while aiming to preserve its utility. To holistically evaluate anonymization methods, we propose a two-pronged framework: (1) generic metric analysis, measuring surface-level fidelity between original and anonymized texts, and (2) task utility loss quantification, assessing downstream performance degradation caused by anonymization.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of N samples, where x_i is an original text and y_i its task-specific label (e.g., classification targets). An anonymization algorithm $\mathcal{A} : x \to x'$ transforms each x_i into its private counterpart x'_i, yielding the anonymized dataset $\mathcal{D}_{\text{priv}} = \{(x'_i, y_i)\}_{i=1}^{N}$. For each sample pair (x_i, x'_i), we compute a similarity metric $s(x_i, x'_i)$ to quantify the text transformation. The overall generic fidelity of $\mathcal{A}$ is then:

$$S(\mathcal{A}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} s(x_i, x'_i),$$

providing a task-agnostic measure of anonymization's impact on textual integrity.

To evaluate task-specific utility degradation, we first train a model f_θ on $\mathcal{D}$, achieving a baseline performance $\mathcal{U}_{\text{orig}}$ on a held-out test set. We then anonymize the test set using $\mathcal{A}$ and evaluate f_θ on $\mathcal{D}_{\text{priv}}$, obtaining $\mathcal{U}_{\text{priv}}$. This setup mirrors real-world scenarios where users apply anonymization to data before feeding it to a pre-trained task model, which they cannot retrain or modify. The sensitivity for task $\mathcal{T}$ is: $\Delta_{\mathcal{T}}(\mathcal{A}) = \mathcal{U}_{\text{orig}} - \mathcal{U}_{\text{priv}}$, where $\Delta_{\mathcal{T}}$ captures the performance drop attributable to anonymization. Methods with high $S(\mathcal{A}, \mathcal{D})$ may still induce significant $\Delta_{\mathcal{T}}$ if anonymization perturbs task-relevant features, underscoring the necessity of joint analysis. More complex task schemes are also possible, such as training models on partially anonymized datasets to evaluate performance on anonymized outputs (Zhai et al., 2022).

2.1 Design Principles

TAU-EVAL is guided by three core principles: ease of use, modularity, and customizability, making it a flexible and accessible framework for evaluation in NLP.

Ease of use TAU-EVAL enables researchers to build complete evaluation pipelines with minimal code. It offers a simple interface and sensible defaults to support rapid development and experimentation.

Modularity The framework is composed of independent components that can be used selectively. This avoids unnecessary processing and supports a wide range of use cases.

Customizability Thanks to its modular design, users can easily integrate custom anonymization models, define new features, and implement specialized metrics, allowing for tailored evaluation workflows.

2.2 Core Elements and Functionalities

TAU-EVAL is an open-source Python framework designed to unify and streamline the evaluation of text anonymization systems. By abstracting fragmented experimental workflows into a modular and extensible framework, the library enables researchers to benchmark anonymization models against diverse metrics and tasks with minimal coding effort. Below, we detail its core architecture, integration capabilities, and workflow. To accommodate the rapid evolution of text generation models and datasets, TAU-EVAL leverages the Hugging Face ecosystem. It natively supports datasets (Lhoest et al., 2021) within tasks, including local storage and custom preprocessing pipelines, models (Wolf et al., 2020), and evaluations metrics (Von Werra

et al., 2022). TAU-EVAL's architecture (Figure 1) revolves around three modular pillars:

Anonymization Models: The framework treats anonymizers as black-box text-to-text functions. Requiring only a lightweight `Anonymizer` interface (`anonymize(text) -> text`). This perspective accommodates a broad spectrum of anonymization techniques; from traditional sequence-to-sequence architectures to modern large-scale language models. Preconfigured templates simplify the use of Hugging Face models (e.g., T5, BART, GPT-2) and LLM APIs via LiteLLM[3], enabling rapid prototyping.

Evaluation Metrics: Metrics are designed to capture diverse dimensions of performance. We integrate measures that assess both the retention of semantic content (e.g., overlap-based and semantic similarity scores) and intrinsic properties of anonymized text (e.g., fluency and reference-less evaluations), the framework provides a comprehensive basis for quantifying the efficacy of anonymization methods. Metrics are computed on each pair $(\mathcal{D}, \mathcal{D}_{\text{priv}})$ taken from task datasets. We implement widely used metrics for text anonymization taken from text generation evaluation, such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020b). We also support sentence-transformers similarity (Reimers and Gurevych, 2019), and reference-less metrics based on language models such as CoLA (Warstadt et al., 2019), or Perplexity (Jelinek et al., 1977).

Target Tasks: Tasks contextualize the evaluation by framing anonymization within downstream applications. We abstract common-use scenarios (e.g., classification, multiple-choice inference) into a unified task paradigm, TAU-EVAL facilitates systematic comparisons across a variety of use cases. We utilize the tasksource ecosystem (Sileo, 2024) to empower rapid access to more than 600 structured datasets and tasks preprocessing with the `AutoTask` class.

```
from tasknet import AutoTask
imdb = AutoTask('imdb')
dynahate = AutoTask('dynahate')
```

Using tasknet, we provide a streamlined interface between evaluation tasks and the Hugging Face trainer for efficient fine-tuning. This integration simplifies dataset importing, preprocessing, and model configuration, allowing researchers to concentrate on designing text anonymization pipelines rather than building evaluation frameworks. Additionally, we support additional built-in tasks that don't require model training, such as sensitive entity detection. Table 3 in appendix gives an overview of all features currently integrated into TAU-EVAL.

[3] https://www.litellm.ai/

3 Usage and Customization

This section provides implementation details for conducting evaluation experiments and explains how to further customize TAU-EVAL for specific research needs.

3.1 Running Experiments

Running an experiment involves three steps: (1) Implement custom anonymizers (see Section 3.2) or load preconfigured models present inside `tau_eval.models`. (2) Choose from built-in options or add custom metrics and tasks. (3) Instantiate an `Experiment` object with models, tasks, and metrics. All results are logged for visualization and comparison.

Experiment The `Experiment` class takes care of orchestrating the evaluation of each anonymization model. It extracts task information from each task, generates the anonymized task version, and computes each chosen metric, while training relevant models if needed. It relies on `ExperimentConfig`, a class storing specific user-defined configuration for evaluation.

ExperimentConfig The `ExperimentConfig` class provides ways to customize the experiment evaluation. In particular, it allows the user to specify which prediction model to train (from a local model or on the Hugging Face Hub) and fine-tune it on tasks, setting training hyperparameters, or more advanced tasks strategies (should we train the classifier on generated data or not) and logging options.

`Experiment` results can be stored either as a dictionary variable for immediate use or serialized to a `.json` file. While the dictionary format is convenient for most direct analysis, JSON serialization helps to store experiment results for versioning and future use.

3.2 Customization

Models Anonymization models are defined as a class which allow initialization i.e. model loading and inference with the anonymization method. A new anonymization model can easily be instantiated using this interface:

```
class TestModel(Anonymizer):
    def __init__(self):
        self.name = "Test Model"

    def anonymize(self, text) -> str:
        # Implement anonymization logic

    def anonymize_batch(self,
                        texts: list[str]) ->
                        ↪ list[str]:
        # Performs batch anonymization for larger
        ↪ datasets
```

Metrics Metrics are implemented as functions that accept one or two text inputs and return one or more scores. Experiments can utilize both TAU-EVAL's built-in metrics and custom functions, provided they follow the signature `Callable[[str | list[str], str | list[str]], dict]`. The experiment framework automatically detects and handles both built-in and custom metrics.

Tasks Tasks enable data loading and classification model integration. While tasksource and tasknet simplify access to the Hugging Face Hub, users can create more sophisticated tasks by implementing the `CustomTask` interface. This interface requires a `dataset` attribute containing a Hugging Face dataset and an `evaluate(self, new_texts: list[str]) -> dict` method that processes anonymized texts. During experiments, the system anonymizes the task dataset and passes it to the evaluation method.

3.3 Visualization

In the evaluation of anonymization systems, TAU-EVAL offers a suite of visualization tools designed to streamline the comparative analysis of models across specific tasks. Our system enables to assess performance through one-to-one model comparisons, enhancing interpretability. Additionally, TAU-EVAL supports the explicit definition of trade-offs between key metrics (e.g., privacy vs. utility), allowing users to explore the nuanced relationships between competing objectives. To further refine analysis, the framework provides flexible filtering mechanisms, permitting users to isolate experimental results based on models, tasks, or evaluation metrics. We also ensure a more structured and granular assessment of anonymization techniques by serializing parts of the anonymized datasets, which is useful for qualitative analysis and to relaunch experiments without model inference.

```
from tau_eval import Experiment, ExperimentConfig
from tau_eval.models.presidio import (
    UniquePlaceholderPerEntity,
    EntityDeletion,
    UniformPlaceholder,
    CategoryPlaceholder,
    FakerPlaceholder,
)
from tau_eval.tasks import DeIdentification
from tasknet import Classification
from datasets import load_dataset

m1 = UniquePlaceholderPerEntity()
m2 = EntityDeletion()
m3 = UniformPlaceholder()
m4 = CategoryPlaceholder()
m5 = FakerPlaceholder()

mednli = Classification(
    dataset=load_dataset("bigbio/mednli"),
    s1="sentence1", s2="sentence2", y="label"
)
pii = DeIdentification(
    dataset="ai4privacy/pii-masking-400k"
)

config = ExperimentConfig(
    exp_name="test-experiment",
    classifier_name="answerdotai/ModernBERT-base",
    train_task_models=True,
    train_with_generations=False,
)

Experiment(models=[m1,m2,m3,m4,m5],
           metrics=["bertscore"],
           tasks=[mednli, pii],
           config=config
).run()
```

Listing 1: Example running an experiment comparing different de-identification strategies for the MedNLI and `pii-masking` datasets. BERTScore will be computed on each dataset for each model.

4 Experiments

To demonstrate the potential of TAU-EVAL, we evaluate the utility loss of anonymization methods across two privacy objectives (PII redaction and authorship obfuscation) and eight downstream utility tasks, spanning diverse domains to capture task-sensitive utility loss. Below, we detail our benchmarking setup.

Privacy Tasks To quantify privacy risks, we focus on two objectives. First, personally identifiable information (PII) redaction involves automatically

Model	PRIVACY	IMDB	DYNASENT	TOXICITY	DYNAHATE	ANLI	MEDNLI	FRAUD	FAKE NEWS
PII Redaction									
Original	0	95.2	77.6	75.6	83.0	56.5	66.3	99.7	98.5
Presidio	66.4	95.1	77.5	75.7	80.0	49.6	66.1	97.1	97.9
Gemini-flash-1.5-8b	96.6	93.7	77.1	56.8	45.0	50.2	65.3	98.3	78.9
Gemini-flash-1.5	98.4	94.8	77.4	53.2	41.6	47.8	61.9	97.3	77.1
Llama-3.1-8b	99.0	94.6	71.8	54.2	49.6	47.7	51.6	92.0	92.7
Llama-3.1-70b	98.7	95.6	74.6	60.8	51.2	49.2	59.8	98.8	85.2
Llama-3.3-70b	98.7	95.5	77.0	69.4	53.1	48.9	63.7	98.7	88.9
Qwen-2.5-7b	95.0	94.9	74.2	65.1	57.5	50.1	61.5	98.6	94.8
Qwen-2.5-72b	97.8	95.3	76.8	61.1	55.9	52.5	60.4	98.7	88.5
Phi-4	97.2	95.1	75.3	58.0	44.7	51.6	61.4	97.5	84.6
Authorship									
Original	0.2	95.2	77.6	75.6	83.0	56.5	66.3	99.7	98.5
StyleRemix	70.4	82.9	73.2	50.8	47.5	42.5	47.3	97.3	73.9
Gemini-flash-1.5-8b	80.6	94.3	70.9	52.4	52.7	45.0	45.9	96.7	47.9
Gemini-flash-1.5	65.6	95.2	71.1	56.9	57.7	44.2	46.2	96.8	58.4
Llama-3.1-8b	76.1	93.8	62.6	53.3	55.1	40.7	42.0	94.1	51.5
Llama-3.1-70b	69.7	95.2	65.1	51.3	57.7	42.3	41.3	95.3	54.8
Llama-3.3-70b	69.3	94.3	65.9	52.9	58.6	41.1	41.8	95.9	60.0
Qwen-2.5-7b	66.1	91.9	65.4	58.2	63.5	43.5	46.5	96.2	73.8
Qwen-2.5-72b	60.2	94.5	70.7	59.3	66.0	40.3	46.6	96.0	75.7
Phi-4	63.4	94.2	66.4	49.9	59.0	44.4	48.7	95.6	59.6
Random	10.1	50.0	33.3	50.0	50.0	33.3	33.3	50.8	50.4

Table 1: Task-specific degradation of each anonymization model (%). The PRIVACY column highlights the models performance over the two privacy datasets: `pii-masking` and IMDB62.

identifying and replacing sensitive personal data that could potentially expose individual identities (e.g., names, addresses, etc.). This task is evaluated using the synthetic `pii-masking` dataset[4], which generates realistic PII in contextual scenarios while avoiding exposure of real private data. This synthetic approach enables safe benchmarking of anonymizers' ability to mask sensitive entities without compromising genuine user privacy, we report the masked entity recall as a privacy goal to maximize. Second, authorship obfuscation is a privacy task that tests whether unique authorship features can be used to identify or re-identify an individual's textual content despite anonymization attempts. The task leverages the IMDB62 corpus (Seroussi et al., 2014), a widely adopted benchmark for authorship attribution. Here, anonymization aims to obfuscate stylistic fingerprints, testing resilience against authorship inference attacks, we evaluate this task with accuracy scores.

Utility Tasks To comprehensively measure task-specific utility degradation, we select eight classification tasks spanning diverse domains and linguistic complexity, each chosen to probe distinct challenges in privacy-utility trade-offs. IMDB movie reviews (Maas et al., 2011) and the adversarially constructed DynaSent dataset (Potts et al., 2021) serve as baselines for *sentiment analysis*. *Toxicity and hate speech detection* (Jigsaw Toxic Comments (cjadams et al., 2019), DynaHate (Vidgen et al., 2021)) evaluate anonymization's impact on socially critical tasks, where over-redaction may obscure harmful language. ANLI (Nie et al., 2020), an adversarial *natural language inference* (NLI) dataset, assesses whether anonymization preserves logical consistency between premises and hypotheses. MedNLI (Romanov and Shivade, 2018) addresses the medical domain's acute privacy needs, where anonymization must protect patient identities without distorting diagnostic inferences. CLAIR (Radev, 2008) (*fraudulent email detection*) and FakeNews detection[5] examines anonymization's effect on stylistic and structural cues critical for identifying malicious intent and misinformation. We add further details for each task in Appendix A.

For class-balanced datasets (IMDB, DynaSent, ANLI, MedNLI, FRAUD, FakeNews), we report accuracy; for imbalanced tasks (Toxicity, DynaHate), F1 scores are used. All tasks are fine-tuned on ModernBERT-large (Warner et al., 2025), selected for its balance of efficiency and state-of-the-art NLP performance. We also include a random classifier predicting the distribution of target labels as a baseline.

Anonymization Models We benchmark two families of anonymization methods, each of them

[4] https://hf.co/datasets/ai4privacy

[5] https://hf.co/datasets/GonzaloA/fake_news

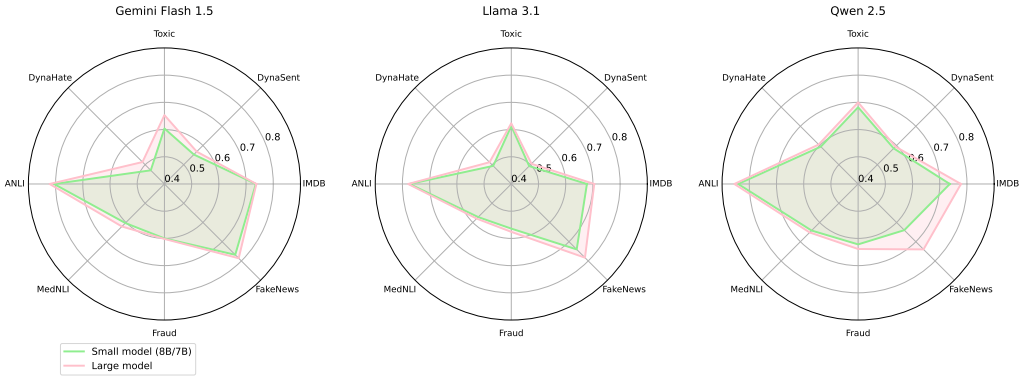

Figure 2: SBERT cosine similarity difference between each tested model and its smaller version.

corresponding to one or the other privacy task. For PII redaction, we evaluate the NER-based Presidio anonymizer (Mendels et al., 2018) as a baseline and compare it against four state-of-the-art LLMs: Gemini Flash 1.5 (Team et al., 2024), Llama-3.1/3.3 (Grattafiori et al., 2024), Qwen-2.5 (Qwen et al., 2025), and Phi-4 (Abdin et al., 2024) prompted for entity replacement. For authorship obfuscation, we test StyleRemix (Fisher et al., 2024), a state-of-the-art baseline, alongside the same LLMs repurposed with style-transfer prompts to disrupt stylistic cues while retaining task-relevant content. Model size effects are analyzed by comparing parameter variants (e.g., Llama-3.1-8B vs. 70B). We access instruction-tuned language models through the OpenRouter API.

5 Results

Table 1 summarizes the performance of anonymization methods across privacy and utility tasks. LLMs emerge as the most effective anonymizers for both PII redaction and authorship obfuscation, achieving best privacy protection. However, this comes at a pronounced cost to utility. Toxicity and hate detection tasks exhibit the steepest performance drops (up to 42% utility decrease for Gemini-flash-1.5 on DynaHate), likely due to LLMs' fine-tuning safeguards against reproducing harmful content. Notably, in the authorship setting, model size inversely correlates with privacy and utility: smaller models tend to do more content modifications to the original text (see Figure 2), leading to better privacy and worse utility. Utility degradation varies dramatically across tasks. Sentiment analysis tasks like IMDB and DynaSent exhibit resilience to anonymization, whereas ANLI, MedNLI and fake news detection degrade significantly. Crucially, no single anonymization method dominates across all tasks, but each familly of models exibit similar performances accross tasks.

Task	BERTScore	METEOR	ROUGE	SBERT	CoLA
IMDB	0.349	0.304	0.349	0.096	0.141
DynaSent	0.735	0.691	0.705	0.735	0.573
Tox	0.681	0.726	0.681	0.516	0.576
DynaHate	0.007	0.051	0.095	0.317	0.051
ANLI	0.627	0.573	0.720	0.352	0.544
MedNLI	0.867	0.764	0.838	0.882	0.705
Fraud	0.518	0.464	0.464	0.597	0.420
FakeNews	0.699	0.759	0.774	0.248	0.323
Average	0.560	0.542	0.578	0.468	0.417

Table 2: Correlation of different metrics with task performance for all utility tasks.

We also performed a correlation analysis between general purpose metrics commonly used in state-of-the-art for utility preservation and task-specific performance measures, computing Kendall's Tau correlation scores as reported in Table 2. The results strongly emphasize the limitations of generic metrics in reliably predicting task utility across different domains. No single metric consistently demonstrates strong correlation with actual utility loss across the evaluated tasks. The sole exception to this pattern was observed in the MedNLI task, where moderate to strong correlations were found. This anomaly can be potentially attributed to the relatively short text size in MedNLI samples, which may result in a more direct relationship between surface-level text modifications and task performance.

6 Related Work

Most existing research addressing similar challenges has primarily focused on evaluating model robustness against sentence-level perturbations. Tools like TextFlint (Wang et al., 2021) offer diverse text noising methods, but emphasize evalu-

ation of downstream task models rather than analyzing the noise mechanisms themselves. PrivKit (Cunha et al., 2024) represents a notable exception in its attempt to establish a comprehensive privacy-utility evaluation pipeline for heterogeneous data types. However, it was primarily designed for location and facial data rather than natural language processing applications and consequently lacks the unified framework necessary to support the diverse requirements of NLP evaluation tasks. Furthermore, PrivKit operates as a standalone system without integration capabilities for popular machine learning frameworks, limiting its accessibility and adoption in existing workflows.

7 Conclusion

In this paper we introduce TAU-EVAL, an evaluation framework enabling unified, reproducible benchmarking of text anonymization systems across both privacy risks and task-specific utility degradation. We study the task sensitivity of language models instructed for anonymization on eight downstream tasks and justify its relevance as a complement to task-agnostic metrics. Our experiments confirm the complexity of achieving a clear trade-off, revealing that no single method dominates across tasks.

Limitations

While our framework advances the evaluation of task-sensitive anonymization, several limitations highlight directions for future work. First, our study operates within a monolingual (English) context. Multilingual anonymization introduces unique challenges (Riabi et al., 2024). Low-resource languages further compound these issues due to sparse PII detection tools and limited benchmark datasets, restricting the generalizability of findings to global contexts.

Second, utility loss measurements depend on the choice of task classifier. While we present evaluations using ModernBERT-large for cross-task comparability, domain-specific architectures (e.g., ClinicalBERT (Huang et al., 2019), CamemBERT-bio (Touchent and de la Clergerie, 2024)) or multilingual models (e.g., XLM-T (Barbieri et al., 2022)) may yield divergent utility trade-offs. For instance, medical anonymization evaluated via a clinically fine-tuned model could reveal subtler impacts on diagnostic inference than our general-purpose setup captures.

Finally, we did implement LLMs mainly using privacy-focused prompts, and could have implemented utility-specific LLMs that perform anonymization when given specific downstream task requirements. Such task-aware anonymization would fundamentally shift the privacy-utility trade-off by limiting data usage to predefined tasks. This presents an interesting direction for future research.

Ethics and Broader Impact Statement

While TAU-EVAL facilitates the benchmarking of anonymization methods, it does not itself perform anonymization or guarantee privacy. Users are responsible for interpreting results in context and for deploying anonymization strategies that align with applicable legal, ethical, and domain-specific standards. We emphasize that TAU-EVAL should not be used as a substitute for legal compliance (e.g., GDPR) but as a research tool to assist in privacy-aware NLP development.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Calvin Bao and Marine Carpuat. 2024. Keep it Private: Unsupervised privatization of online text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8678–8693, Mexico City, Mexico. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2023. Clinical text anonymization, its influence on downstream NLP tasks and the risk of re-identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111, Dubrovnik, Croatia. Association for Computational Linguistics.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. Kaggle.

Mariana Cunha, Guilherme Duarte, Ricardo Andrade, Ricardo Mendes, and João P. Vilela. 2024. Privkit: A toolkit of privacy-preserving mechanisms for heterogeneous data types. In *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '24, page 319–324, New York, NY, USA. Association for Computing Machinery.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024. StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4172–4206, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Hemanth Kandula, Damianos Karakos, Haoling Qiu, and Brian Ulicny. 2024. Improving authorship privacy: Adaptive obfuscation with the dynamic selection of techniques. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 137–142, Bangkok, Thailand. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, 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.

Thomas Lampoltshammer, Lőrinc Thurnay, and Gregor Eibl. 2019. *Impact of Anonymization on Sentiment Analysis of Twitter Postings*, pages 41–48.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2020. Closing the gap: Joint de-identification and concept extraction in the clinical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6945–6952, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 4188–4203.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. TAROT: Task-oriented authorship obfuscation using policy optimization methods. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 14–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026.

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.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. 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 2388–2404, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Dragomir Radev. 2008. Clair collection of fraud email. ACL Data and Code Repository, ADCR2008T001.

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.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

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.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024a. Large language models are anonymizers. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024b. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, and Libin Bai et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Rian Touchent and Éric de la Clergerie. 2024. CamemBERT-bio: Leveraging continual pre-training for cost-effective models on French biomedical data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2692–2701, Torino, Italia. ELRA and ICCL.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurements. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Alison: Fast and effective stylometric authorship obfuscation. In *AAAI*.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria. Association for Computational Linguistics.

Wanyue Zhai, Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. 2022. Adversarial authorship attribution for deobfuscation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7372–7384, Dublin, Ireland. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Datasets & Task Details

Our evaluation framework employs datasets that span privacy preservation and task-specific utility, carefully selected to reflect real-world anonymization challenges across diverse domains. Below, we elaborate on their structure, relevance, and role in quantifying the privacy-utility trade-off.

pii-masking This synthetic dataset simulates real-world privacy risks by generating text records containing 63 classes of personally identifiable information (PII) (see https://hf.co/ai4privacy), including names, addresses, medical IDs, and financial data. Unlike real-world corpora, synthetic generation avoids ethical concerns while enabling controlled benchmarking of anonymizers' ability to mask sensitive entities. We sample 5,000 English texts from its 400k corpus, ensuring representation of all PII categories. The dataset's structured noise injection (e.g., realistic address formats, contextualized medical terms) tests anonymizers' precision in distinguishing sensitive from non-sensitive content which is a critical capability for GDPR/CCPA compliance.

IMDB62 Derived from the IMDb movie review corpus (Seroussi et al., 2014), this dataset contains 62,000 texts from 62 distinct authors, equally distributed. We focus on a 10-class subset (first 10 authors) to evaluate authorship obfuscation. The task challenges anonymizers to disrupt stylistic fingerprints (e.g., syntactic patterns, lexical preferences).

IMDB (Maas et al., 2011): A classic binary sentiment classification task (positive/negative) on 50k movie reviews. Its balanced distribution and informal language (e.g., user-generated reviews) test anonymization's impact on opinion-driven text utility.

DynaSent (Potts et al., 2021): A ternary sentiment dataset (positive/neutral/negative) constructed via the Dynabench platform (Kiela et al., 2021), where adversarial examples are iteratively generated to exploit model weaknesses. DynaSent's complexity evaluates whether anonymization exacerbates or mitigates robustness to subtle linguistic perturbations.

Toxicity (cjadams et al., 2019): A binary classification task identifying toxic language ("rude, disrespectful, or disruptive" content) in online comments. The dataset's inherent class imbalance ($\approx 10\%$ toxic) tests anonymization's impact on minority class performance and harmful content detection, a critical consideration for moderation systems.

DynaHate (Vidgen et al., 2021): Built using Dynabench's adversarial framework, this dataset targets hate speech detection with examples designed to bypass automated filters. Its adversarial nature stresses anonymization's ability to preserve subtle hate indicators (e.g., dog whistles, coded language) while removing identifiers.

ANLI (Nie et al., 2020): An adversarial natural language inference (NLI) dataset where premises and hypotheses are iteratively refined to challenge model reasoning. By anonymizing both text spans, we test consistency in logical inference: a measure of whether anonymization disrupts semantic relationships critical for NLI.

MedNLI (Romanov and Shivade, 2018): A medical NLI dataset derived from clinical notes, requiring models to infer entailment/contradiction relationships between patient descriptions and hypotheses. Anonymization here risks altering clinically relevant entities (e.g., medications, symptoms), making it a benchmark for domain-specific utility preservation.

Fraud (Radev, 2008): A corpus of fraudulent emails spanning 1988–present, where anonymization must balance redacting PII (e.g., names, account numbers) with preserving stylistic cues (e.g., urgency markers, grammatical errors) indicative of fraud.

Fake News : A binary classification task on 45k news articles (sourced from https://hf.co/datasets/GonzaloA/fake_news) that evaluates anonymization's impact on semantic coherence in misinformation contexts. Articles often contain named entities (e.g., politicians, organizations) whose anonymization may alter the perceived credibility of claims.

B Motivations from Related Work

This section presents a concise overview of the background and related work on evaluating the utility of text anonymization, providing the foundation for the motivations of this paper.

Prior work on text anonymization utility has been most extensive in the medical domain (Lange et al., 2020; Ben Cheikh Larbi et al., 2023), where protecting Protected Health Information (PHI) is ethically and legally required. Studies show that excessive anonymization can reduce clinical value (Berg et al., 2020). Outside the medical domain, utility assessments are narrower, often limited to sentiment analysis (Yang et al., 2025; Patsakis and Lykousas, 2023; Lampoltshammer et al., 2019). Lothritz et al. (2023) provide a broader evaluation

focused on GLUE tasks (Lothritz et al., 2023) but target name redaction only, overlooking semantic and syntactic effects. Recent findings highlight that pseudonymization impacts tasks differently (Riabi et al., 2024), motivating systematic, multi-task evaluation.

Other domains, such as authorship anonymization, emphasize identity protection and utility is often assessed narrowly through text generation metrics like n-gram preservation and fluency (Xing et al., 2024; Fisher et al., 2024; Kandula et al., 2024; Bao and Carpuat, 2024). The rise of large language models as both privacy attackers (Staab et al., 2024b) and anonymizers (Staab et al., 2024a) underscores the need for comprehensive evaluation protocols that combine task-agnostic and task-specific measures.

C Prompts

Figures 3 and 4 show the prompts used for authorship privacy and PII redaction tasks.

User: `Rewrite the text to significantly alter its style, tone, and word choice while preserving the original meaning. Use figurative or descriptive language, vary sentence structures, adjust tone (e.g., formal to conversational), and employ unique synonyms. Avoid retaining distinctive stylistic markers. Respond only with the transformed text.`

Figure 3: Authorship obfuscation prompt

User: `Anonymize user-submitted text by replacing personally identifiable information (PII) like names, dates, locations, and professions with alternate values. Ensure the rewritten text is natural, coherent, and contextually consistent while preserving the original meaning and tone. Respond only with the transformed text.`

Figure 4: PII redaction prompt

Already Implemented Models
Pseudonymization (Riabi et al., 2024)
– Entity Deletion
– Uniform Placeholder
– Category-Specific Placeholder
– Unique Placeholder per Entity
– Unique Substitute per Entity
Authorship Obfuscation
– StyleRemix (Fisher et al., 2024)
– Simplification (Laban et al., 2021)
– Back Translation (Fan et al., 2021)
– Paraphrasing (Zhang et al., 2020a)
[+] *preconfigured templates for HF models & API-based LLMs*
Already Implemented Metrics
Reference-based
– ROUGE (Lin, 2004)
– METEOR (Banerjee and Lavie, 2005)
– BERTScore (Zhang et al., 2020b)
– NLI (Liu et al., 2022)
– LUAR similarity (Rivera-Soto et al., 2021)
– SBERT similarity (Reimers and Gurevych, 2019)
Reference-less
– CoLA (Warstadt et al., 2019)
– Perplexity (Jelinek et al., 1977)
[+] *preconfigured templates for HF evaluate metrics*
Already Implemented Tasks
– De-identification
– IMDB Authorship Attribution
– Text Anonymization Benchmark (Pilán et al., 2022)
– Sequence Classification
– Token Classification
– Multiple Choice
– Sequence-to-Sequence
[+] *preconfigured templates for tasksource tasks*

Table 3: Overview of already implemented models, metrics, and tasks at the time of submission (July 2025).