TAE: Text Anonymization Evaluator

Benet Manzanares-Salor[a], David Sánchez[a], Pierre Lison[b], Ildikó Pilán[b]

[a] *Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, ComSCIAM-Center for Computational Science and Applied Mathematics, Tarragona, 43007, Spain.*
Email:{benet.manzanares,david.sanchez}@urv.cat
[b] *Norwegian Computing Center, Postboks 114 Blindern, Oslo, NO-0314, Norway.*
Email:{plison,pilan}@nr.no

Abstract

Text anonymization is essential for privacy-compliant data sharing, yet its evaluation remains challenging. Existing evaluation metrics are fragmented, and many rely on manual annotations, which hamper applicability and introduce subjectivity. We present the *Text Anonymization Evaluator* (TAE), a Python-based package that consolidates both traditional and state-of-the-art metrics for assessing privacy and utility in anonymized text. By integrating these metrics into a unified and extensible toolkit, TAE enables systematic, scalable, and objective evaluation and comparison of text anonymization methods.

Keywords: text anonymization, evaluation, privacy, utility

Metadata

1. Motivation and significance

The increasing availability of textual data has fueled major advances in data mining and machine learning across most scientific domains. However, text frequently contains personal information whose disclosure can harm individuals and violate privacy regulations such as the European *General Data Protection Regulation* (GDPR) [9].
To share or release personal data without conflicting with legal frameworks, data controllers primarily rely on *anonymization*. Anonymization seeks to suppress identifiers (*e.g.*, full name or passport), which directly reveal subjects' identity, and mask quasi-identifiers (*e.g.*, birth date, ZIP code, occupation), which may appear innocuous in isolation but can be combined to uncover someone's identity. However, because suppression and masking result in information loss that can affect the analytical utility of data, anonymization

Nr.	Code metadata description	Metadata
C1	Current code version	1.2.0
C2	Permanent link to code/repository used for this code version	`https://github.com/BenetManzanaresSalor/TAE`
C3	Permanent link to Reproducible Capsule	None
C4	Legal Code License	MIT License
C5	Code versioning system used	Git
C6	Software code languages, tools, and services used	Python
C7	Compilation requirements, operating environments & dependencies	Python 3.9 and above, dependencies listed in pyproject.toml
C8	If available Link to developer documentation/manual	`https://github.com/BenetManzanaresSalor/TAE/blob/main/README.md`
C9	Support email for questions	benet.manzanares@urv.cat

Table 1: Code metadata

methods should carefully balance the trade-off between privacy protection and preservation of data utility.

Anonymization of structured data has reached a mature stage [6], benefiting from well-established privacy models such as k-anonymity [32] or differential privacy [8]. Nonetheless, text anonymization remains an open and complex problem [16, 4]. The absence of explicit attribute boundaries in plain text makes the detection of (quasi-)identifying terms challenging, a problem that is often addressed through manual annotation [14, 10, 15, 1]. To mitigate this tedious manual effort, some automated mechanisms have been proposed. Practical approaches rely primarily on *named entity recognition* (NER) tools to detect entity types assumed to be disclosive, such as person names and dates [13, 37, 11, 39, 12, 2]. More recently, some authors have started leveraging *large language models* (LLMs) for this purpose [34, 2, 27, 35, 38].

A major challenge in text anonymization is how to precisely assess the quality of the anonymization process. Most works in the literature rely on comparisons against a manually annotated ground truth and use recall and precision metrics to measure, respectively, the coverage of masked identifying terms (as a proxy for privacy protection assessment) and the proportion of correctly masked terms (as a proxy for utility preservation assessment). Nevertheless, the use of standard precision and recall metrics presents several challenges

in text anonymization. Indeed, those metrics are based on subjective human judgments and are therefore prone to human-induced errors and biases. As they put uniform weights on all personal identifiers, they also ignore the varying contribution of individual terms to the re-identification risk and information loss [16, 22, 19, 23, 20]. To address some of those challenges, [28] presented an extension of the standard precision and recall measures specifically tailored to text anonymization that, crucially, allow evaluation with multiple ground-truth answers, reflecting the fact that several masking strategies can be equally effective at preventing re-identification. However, those measures still depend on imperfect human judgments and fail to consider the impact of the masking strategy (*e.g.* suppression, replacement) on utility preservation [29].

These challenges also apply to the availability of evaluation tools. Although frameworks exist for assessing anonymized structured data, such as ARX [30], sdcMicro [36], μ-ANT [33] and the MobiDataLab Anonymizer, there is a lack of comparable tools for text. The closest is τ-Eval [17], but its scope is limited to evaluating NER-based anonymization methods. For privacy evaluation, it uses benchmarks for personal identifiable information (PII) detection and authorship attribution, while for utility, it relies on standard text quality and similarity metrics. While this approach allows for a general ranking of anonymization methods, τ-Eval is not designed to assess the anonymization quality of a user-defined *corpus*. This is at odds with practitioners' needs, as they usually require tools to evaluate whether an anonymization technique works effectively on their specific corpus. Furthermore, the privacy assessment process depends on manual annotations, which are both costly and subjective.

Therefore, a comprehensive tool for automatically and systematically evaluating privacy protection and utility preservation for user-defined corpora remains unavailable. In the absence of such a tool, making meaningful comparisons between anonymization approaches and conducting grounded evaluations prior to the publication of specific data is severely constrained.

In this respect, novel evaluation metrics have recently been proposed [22, 19, 24, 3, 25, 23, 29], which leverage the capabilities of language models (LMs) to provide a fully automated, highly configurable and masking-aware assessment without relying on manual annotations. Rather than being proxies for privacy and utility, these metrics directly quantify the re-identification risk and information loss of anonymized texts by adhering to the same well-established principles used in structured data anonymization.

The *text anonymization evaluator* (TAE) software package introduced in this paper provides a unified and user-friendly framework for both traditional

measures and state-of-the-art evaluation metrics based on LMs. TAE is particularly well suited for experimental setups where multiple anonymization outputs (whether from free-text documents, web pages, or any other textual formats) must be systematically and objectively compared. Users can supply an original text corpus along with several anonymized versions, and the software will generate a comprehensive comparison in terms of privacy protection and utility preservation. Thus, TAE enables researchers to benchmark new anonymization methods and supports data holders in selecting the most suitable anonymization method for their specific scenario.

The rest of the paper is organized as follows: Section 2 describes the implemented software, detailing its architecture and functionalities. Section 3 provides a usage example with empirical results. Section 4 discusses the current and future impact of the package. Finally, Section 5 concludes and outlines future work directions.

2. Software description

This section outlines TAE's design and core functionalities, details its architecture and usage options, and describes the implemented metrics and their settings.

2.1. Software architecture

TAE is Python-based and its architecture processes input data, conducts the assessment, and generates a corresponding evaluation results file, as shown in Figure 1. Its components are detailed below.

TAE requires the following inputs, with the formatting details provided in the documentation referenced in Table 1:

- **Original corpus**: Documents to anonymize, which constitute the ground truth for some of the metrics. Each document has a unique ID that corresponds to the individual to which it refers (one per document).
- **Anonymizations**: Anonymized versions of the original corpus to be subjected to evaluation. Each anonymization specifies, for every document in the original corpus (referenced by ID), a list of masked terms. Each mask is represented as a tuple specifying the positions of the first and last character indices. Optionally, the tuple can include a third element that states a replacement used to mask the original term

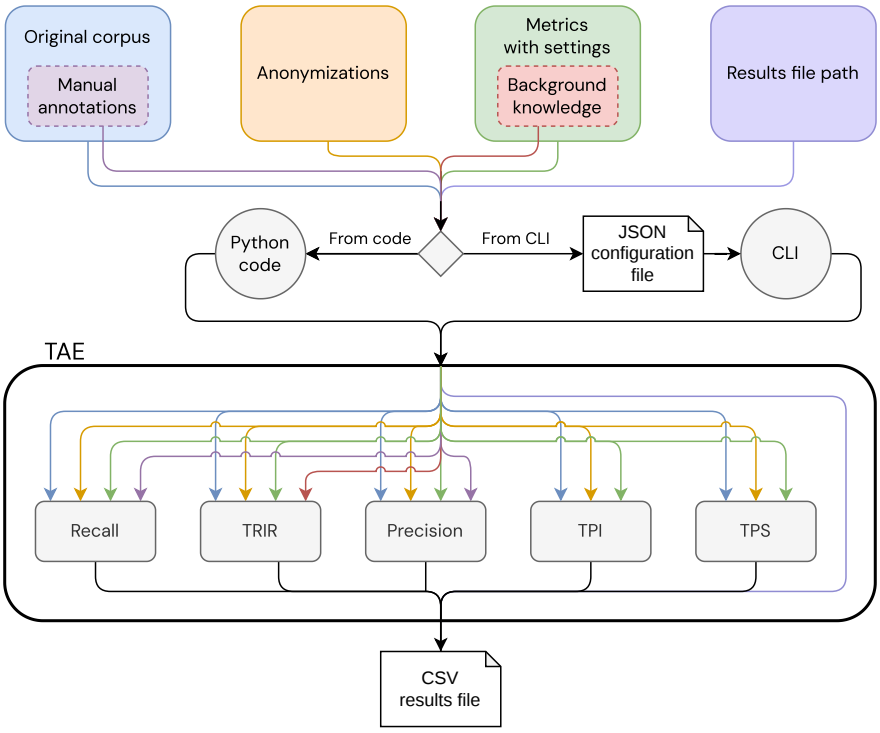


Figure 1: Pipeline diagram. Each input is color-coded, with colored arrows connecting it to the metrics that use it. Dashed lines indicate inputs required only for specific metrics.

(*e.g.*, "European city" in place of "Oslo"); if no replacement is provided, suppression-based masking is assumed.

- **Metrics with settings**: List of evaluation metrics to assess, along with their respective settings (see Section 2.2). Default values are assigned for non-mandatory settings.

- **Results file path**: The location where the evaluation results will be stored.

Certain metrics require additional inputs. For precision and recall, manual annotations must be included in the original corpus (following the format outlined in [28]), and TRIR requires the background knowledge (BK) for the set of referred individuals to be included in the settings. Details on these inputs are provided in the respective metric definitions of Section 2.2.

TAE supports two ways to provide these inputs. The first one is a *command line interface* (CLI) through the command `python -m tae config.json`.

This runs the software with a configuration JSON file specifying all required inputs: paths to the original corpus and anonymization JSON files, metrics with settings, and the file path for the results.
Alternatively, TAE can be executed directly from Python code. This requires passing the input parameters to the `TAE` class constructor and the `evaluate` function. The parameters can mirror those in the JSON configuration file, but the procedural definition of the corpus and anonymizations is also allowed.
In both cases, the results will be reported through execution logs and stored in the specified *results file path* in CSV format: one row per metric and one column per input anonymization.

The package has been designed with extensibility and modularity by design. To incorporate new metrics, users only need to implement a class inheriting from the "MetricABC" abstract base class. Comprehensive guidance on this process is provided in the documentation referenced in Table 1. This approach not only simplifies the addition of new or custom metrics but also promotes code maintainability and reusability. Furthermore, LM-based metrics support model swapping by simply changing the model name in the settings, with no need to alter the source code. This significantly reduces the technical overhead associated with model experimentation and enables users to utilize domain-specific models customized for their particular corpus, or the latest advances in the area. Crucially, not only the quality of the results, but also the operational scalability, heavily depends on these backbone models. This adaptability has profound practical benefits, enabling resource-constrained data holders to execute evaluations locally. Such local execution is not only convenient for benchmarking but also essential for handling sensitive data, as it avoids reliance on external resources, including remote servers hosting proprietary, closed-weight models.

2.2. Software functionalities

This section provides an overview of the privacy and utility metrics implemented in TAE. Default presets are provided for all optional settings, including backbone LMs and metric-specific hyperparameters. Advanced users can refer to the documentation referenced in Table 1 for comprehensive details on these settings. A comparison between all metrics is provided in Table 2.

2.2.1. Privacy metrics

- **Recall**: The traditional metric for privacy protection assessment [16, 19, 29]. It compares an anonymization against manual annotations on the original corpus, which are assumed to be the ground truth. Recall quantifies the percentage of annotated terms ($\#TruePositives + \#FalseNegatives$) successfully masked by the input anonymization ($\#TruePositives$):

$$Recall = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \tag{1}$$

  The contribution of each term to the metric is binary, and it is determined solely by whether it matches the reference or not, neglecting replacement-based masking. Manual annotations must be included in the original corpus for this metric to be applicable.

  Our implementation incorporates the extensions to recall proposed in [28]. In the first, recall is calculated separately for identifiers and quasi-identifiers, leveraging annotations may include this information. In the second, recall is computed independently for each annotated entity type (*e.g.*, PERSON or LOCATION), if available. In the third, users can choose between token-level or mention-level recall; while token-level recall is the standard approach, mention-level recall requires all occurrences of an entity (*e.g.*, a location name) to be masked for it to be considered valid. When several alternative ground-truth annotations are available, the recall measure is micro-averaged across the annotations. Even though all of these extensions still rely on manual annotations, they enable a more fine-grained privacy analysis.

- **Text Re-Identification Risk (TRIR)** [22]: This metric empirically assesses re-identification risk by applying a *text re-identification attack (TRIA)* to the input anonymization. TRIA adapts the principles of the record linkage attack [5] (the principled way to measure re-identification risk in structured data) to textual data, inferring the individual corresponding to each document. TRIR is defined as the success rate of the attack, computed from the ratio of correctly linked documents to the total number of documents:

$$\text{TRIR} = \frac{\#CorrectlyLinkedDocuments}{\#Documents} \tag{2}$$

  In contrast to recall, this metric evaluates the residual privacy risk without relying on manual annotations. The assessment involves fine-tuning an LM with background knowledge (BK) consisting of a collection of publicly available, identified texts about the individuals to be

protected. For instance, if the (anonymized) corpus consists of medical reports from a hospital, the BK may include social media posts from people living in the city where the hospital is located. This BK is a mandatory setting for this metric, and it is structured as a dictionary where each key is an individual's ID and the value is their corresponding public text.

2.2.2. Utility metrics

- **Precision**: The traditional metric for utility preservation assessment [16, 19, 29]. It measures the proportion of terms masked by the method ($\#TruePositives + \#FalsePositives$) that match the manually annotated ones ($\#TruePositives$):

$$Precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \tag{3}$$

  As for recall, precision only considers whether the masked terms align with the manual annotations (which are mandatory), disregarding any potential replacements.

  Our implementation also incorporates the extensions to precision proposed in [28]. For documents with multiple annotations, the micro-average across all annotators is computed. In addition, precision can be calculated at either token-level or mention-level. Whereas token-level precision considers individual words as true positives, mention-level precision requires the entire annotated mention to be masked for it to qualify as a true positive. Masked spans can also be weighted by their information content (IC) using an LM pre-trained on the masked language modeling task. The IC is estimated by assessing a language model's predictability for the original terms based on the masked ones: the less likely the LM is to correctly predict the original term, the higher its IC. Thanks to this weighting scheme, the contribution of each (masked) term to utility preservation can be more accurately measured.

- **Text Preserved Information (TPI)** [23]: This metric measures the ratio of the IC retained in the anonymized text relative to the original text, denoted as $TIC(A)$ and $TIC(O)$ in Eq. 4:

$$TPI(O, A) = \frac{TIC(A)}{TIC(O)} \tag{4}$$

  TPI seeks therefore to directly assess the information loss rather than compare the anonymization outputs to manual annotations. IC calculation is based on the LM-based methodology introduced in [28].

The total IC of the original document is computed by summing the IC of all its terms and, for its anonymized version, the IC associated with the terms masked is subtracted from that total (*i.e.*, assuming a suppression-based masking strategy).

- **Text Preserved Similarity (TPS)** [29]: This metric extends TPI to account for replacement-based masking. Instead of assuming that the IC of masked terms is completely lost (as for suppression-based masking), it weights the original IC by the semantic similarity between the original term t_i and its replacement t_i^*:

$$TPS(O, A) = \frac{\sum_{(t_i, t_i^*) \in (O,A)} IC(t_i) * Similarity(t_i, t_i^*)}{TIC(O)} \tag{5}$$

The more similar the replacements are, the lower the information loss. Similarity is assessed with an LM-based embedding model employing the SentenceTransformers [31] library, with scores ranging from 0 to 1. In essence, TPS is a similarity-weighted variant of TPI, which takes into account that replacements in the anonymized document will retain a subset of the information of the original terms they mask.

3. Illustrative example

We demonstrate the capabilities of our software through the *text anonymization benchmark* (TAB) [28], a well-established evaluation dataset consisting of procedural sections of 1,268 court rulings. Each document in the corpus is annotated for anonymization by multiple human experts following privacy-oriented guidelines.

For evaluation, we use the TAB test set and apply all the implemented metrics with their default configurations. Precision and recall are calculated using the aforementioned annotations. For TRIR's BK, we define a realistic worst-case scenario where the attacker has access to the most similar (yet non-overlapping) documents, namely the non-procedural sections of the court rulings. We also include two metric variants: i) recall restricted to identifiers (excluding quasi-identifiers), and ii) precision weighted by the IC of terms.

Our assessment covers a wide range of text anonymization methods, including several based on NER, three approaches that prompt an LLM[1] to detect

[1]GPT-4o-2024-09-03: `https://platform.openai.com/docs/models/gpt-4o`

Table 2: Comparison of privacy and utility metrics implemented in TAE by objectiveness (*i.e.*, independence from manual annotations), support for replacement-based masking, required resources, and per-term contribution.

Privacy		*Utility*		
Recall [28]	TRIR [22]	Precision [28]	TPI [23]	TPS [29]
Can be applied without manual annotations?				
✗	✓	✗	✓	✓
Supports the assessment of replacement-based masks?				
✗	✓	✗	✗	✓
What are the required resources?				
Manual annotations	Background knowledge and an LM	Manual annotations and, optionally, an LM	An LM	LMs for IC estimation and similarity assessment
How is the per-term contribution computed?				
Recall w.r.t. annotations	Depending on the attack model	Precision w.r.t. annotations, optionally IC-weighted	Information Content (IC)	IC and similarity to the original term

disclosive terms, human annotations, and two replacement-based masking techniques. For the latter, replacements were generated specifically for the manually annotated terms.

The results for each combination of metric and anonymization method, as reported by TAE, are presented in Table 3.

In terms of privacy, the results show notable divergences between metrics and methods. For methods such as St.NER3, standard recall is approximately half of the identifiers-only recall, depicting the difficulty of detecting quasi-identifiers for NER-based methods. This contrasts with LLM-based approaches like Sparks, which result in virtually identical figures for both recall versions. Furthermore, risk assessment via the TRIR metric does not (inversely) correlate well with recall, highlighting the fact that the manual annotations on which recall relies yield unreasonably high risks. This aligns with prior studies [21, 22, 19, 23, 20], where human-based references pro-

[2] `https://spacy.io/api/entityrecognizer`

[3] `https://github.com/microsoft/presidio/releases/tag/2.2.354`

Table 3: Evaluation results on the TAB test set for each metric and anonymization method, the latter grouped by type from top to bottom: NER-based, LLM-based, manual annotations, and replacement-based masking.

Method	*Privacy*			*Utility*			
	Recall↑	Recall (identif.)↑	TRIR↓	Precision↑	Precision (IC)↑	TPI↑	TPS↑
St.NER3 [18]	33.6%	62.3%	52.8%	40.2%	49.7%	88.2%	88.2%
St.NER4 [18]	36.6%	66.9%	47.2%	33.0%	41.4%	83.7%	83.7%
St.NER7 [18]	66.5%	68.5%	47.2%	53.2%	58.6%	79.4%	79.4%
spaCy NER[2]	91.8%	91.9%	58.3%	43.6%	51.0%	73.2%	73.2%
Presidio[3]	74.4%	69.2%	59.1%	77.0%	80.0%	86.1%	86.1%
Student [34]	46.1%	61.8%	58.3%	56.1%	61.0%	88.1%	88.1%
Sparks [2]	87.7%	86.7%	51.2%	51.5%	56.7%	75.6%	75.6%
MvM [27]	86.5%	94.5%	44.9%	47.0%	54.1%	69.8%	69.8%
Manual (supres.) [28]	100.0%	100.0%	57.5%	69.7%	73.3%	77.7%	77.7%
Dou et al. [7]	100.0%	100.0%	57.5%	69.7%	73.3%	77.7%	88.4%
INTACT [29]	100.0%	100.0%	57.5%	69.7%	73.3%	77.7%	88.5%

duced risks significantly exceeding the random guess. Finally, replacement-based masking methods achieve suppression-based TRIR levels, suggesting that chosen replacements are not inherently disclosive. Utility results also expose a misalignment between human-based and automatic LM-based evaluations. Precision, even when accounting for terms' IC, consistently yield lower values than the utility measured by TPI. In addition, TPS underscores the substantial improvement in utility achieved through replacement-based masking, which is neglected by the other utility metrics. Overall, the evaluation figures provided by TAE enable a nuanced analysis of the privacy–utility trade-off. While anonymization methods providing stronger protection generally preserve less information, exceptions exist. For example, spaCy NER achieves the second-highest TRIR and second-lowest TPI, while St.NER4 attains the second-lowest TRIR and fourth-highest TPI, illustrating both worse and better trade-offs.

Regarding runtime, the whole evaluation took 21 minutes on a mid-tier workstation equipped with an NVIDIA RTX 4080 GPU with 16GB of VRAM, an Intel i7-12700 CPU, and 32GB of DDR4 RAM. The most computationally demanding step was the fine-tuning of the TRIR's model, which took 13 minutes and required $\leq$6GB of GPU memory. This efficiency stems from the default choice of compact, open-weight language models, which enable local

execution on consumer-grade hardware.
We also assessed the software's scalability using the Wikipedia-based corpus from [26], constituted by 553 randomly sampled biographies. This dataset features an average document length of 96 words, which is around 10 times shorter than the 983 words per document in TAB. However, the mean runtime per anonymized document for this new corpus was 228 milliseconds, representing a decrease of only 4 times compared to the 902 milliseconds recorded for TAB. This discrepancy suggests that text length plays a secondary role in performance, with execution time being heavily dominated by other factors such as model and sample loading overhead.

4. Impact

TAE facilitates research in text anonymization by providing a unified and ready-to-use evaluation framework. Beyond traditional precision and recall metrics (and their extensions), TAE integrates state-of-the-art LM-based metrics for the first time. The latter metrics enable automatic assessment of privacy protection and utility preservation, eliminating the need for subjective and labor-intensive manual annotations.
For researchers, TAE facilitates systematic and scientifically grounded benchmarking of anonymization methods. For data holders, it enables corpus-specific selection of the best performing anonymization method, in addition to the evaluation of manual anonymization efforts, which are often the backbone of real world data releases [14, 10, 15].
At the time of writing, the papers presenting the state-of-the-art metrics implemented in TAE [22, 21, 23, 29] have garnered a total of 44 citations. This reflects significant research engagement, particularly given that these papers are less than two years old on average. Additionally, the paper introducing the extensions for precision and recall metrics [28] also implemented in TAE has accumulated 140 citations since its publication four years ago.

5. Conclusions and future work

The Text Anonymization Evaluator (TAE) responds to the fragmented landscape of text anonymization evaluation by providing a unified framework that integrates traditional and state-of-the-art metrics for privacy protection and utility preservation. The package empowers researchers and data holders to conduct objective and automated assessments, and to make informed decisions on the relative performance of text anonymization methods. Furthermore, even though it integrates the use of language models, TAE is efficient

enough to allow local execution on consumer-grade hardware, ensuring data sovereignty remains with the data holders.
As future work, we intend to expand the framework's capabilities by introducing support for diverse text file formats beyond standard plain text, such as PDF. Additionally, we plan to integrate optimized default configurations tailored for various LMs. This includes both domain-specific models (*e.g.*, for the medical field) and multilingual support featuring automatic language detection, moving past the current English-only assumption. Finally, we will continue to implement novel metrics as they emerge in the literature and address user-reported limitations to enhance the overall robustness, utility, and applicability of our framework.

Acknowledgements

We acknowledge support from the Norwegian Research Council (CLEANUP project `http://cleanup.nr.no/`, grant nr. 308904), the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez) and MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2024-157271NB-I00 "CLEARING-IT". The first author is also supported by the Spanish Government under an FPU grant (ref. FPU23/01785).

References

[1] Eric Bier, Richard Chow, Philippe Golle, Tracy Holloway King, and Jessica Staddon. The rules of redaction: Identify, protect, review (and repeat). *IEEE Security & Privacy*, 7(6):46–53, 2009.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023.

[3] Lucas Georges Gabriel Charpentier and Pierre Lison. Re-identification of de-identified documents with autoregressive infilling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1209, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.60.

[4] Gergely Márk Csányi, Dániel Nagy, Renátó Vági, János Pál Vadász, and Tamás Orosz. Challenges and open problems of legal document anonymization. *Symmetry*, 13(8):1490, 2021.

[5] Josep Domingo-Ferrer and Vicenc Torra. Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13(4):343–354, 2003. ISSN 1573-1375.

[6] Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, volume 8. Synthesis Lectures on Information Security, Privacy, and Trust, Online, 2016.

[7] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages and Programming*, volume 4052, pages 1–12. Springer, 2006.

[9] European Commission. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[10] Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Marti, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools. *PhysioNet*, November 2023.

[11] Fadi Hassan, Josep Domingo-Ferrer, and Jordi Soria-Comas. Anonymization of unstructured data via named-entity recognition. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 296–305, Mallorca, Spain, 2018. Springer.

[12] Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. Texthide: Tackling data privacy in language understanding tasks. In

Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2020, pages 1368–1382, Punta Cana, Dominican Republic, 2020. Association for Computational Linguistics.

[13] Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the Association for Computing Machinery Conference on Health, Inference, and Learning*, pages 214–221, Toronto, Ontario, Canada, 2020. Association for Computing Machinery.

[14] Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Jan Rodríguez-Miret, Santiago Frid, Xavier Pastor, Xavier Borrat, and Martin Krallinger. A textual dataset of de-identified health records in spanish and catalan for medical entity recognition and anonymization. *Scientific Data*, 12(1):1088, Jul 2025.

[15] S Lima-López, E Farré-Maduell, L Gasco Sánchez, A López Rueda, S Frid, A Conesa, X Pastor, and M Krallinger. Carmen-i: Anonymization protocol for clinical reports in english. Technical report, Zenodo, 2023.

[16] Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 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. Association for Computational Linguistics, 2021.

[17] Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. Tau-eval: A unified evaluation framework for useful and private text anonymization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 216–227, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0.

[18] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

[19] Benet Manzanares-Salor and David Sánchez. Enhancing text anonymization via re-identification risk-based explainability. *Knowledge-Based Systems*, 310:112945, 2025. ISSN 0950-7051.

[20] Benet Manzanares-Salor and David Sánchez. A comparative analysis, enhancement and evaluation of text anonymization with pre-trained large language models. *Expert Systems with Applications*, 297:129474, 2026. ISSN 0957-4174.

[21] Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Automatic evaluation of disclosure risks of text anonymization methods. In Josep Domingo-Ferrer and Maryline Laurent, editors, *Privacy in Statistical Databases - International Conference, PSD 2022, Paris, France, September 21-23, 2022, Proceedings*, volume 13463 of *Lecture Notes in Computer Science*, pages 157–171. Springer, 2022.

[22] Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. *Data Mining and Knowledge Discovery*, 38(6):4040–4075, 2024.

[23] Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Unsupervised utility evaluation of text anonymization methods via neural language models. *Neural Networks*, 202:109079, 2026. ISSN 0893-6080.

[24] John X Morris, Thomas R Campion, Sri Laasya Nutheti, Yifan Peng, Akhil Raj, Ramin Zabih, and Curtis L Cole. DIRI: Adversarial patient reidentification with large language models for evaluating clinical text anonymization. *AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science*, 2025:355–364, 2025.

[25] Sebastian Ochs and Ivan Habernal. The Conundrum of Trustworthy Research on Attacking Personally Identifiable Information Removal Techniques. *Computational Linguistics*, pages 1–33, 2026. doi: 10.1162/COLI.a.615.

[26] Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France, June 2022. European Language Resources Association.

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

[28] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(8):1053–1101, 2022. ISSN 0891-2017.

[29] Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Truthful text sanitization guided by inference attacks. *Applied Soft Computing*, 185:114013, 2025. ISSN 1568-4946.

[30] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. Flexible data anonymization using ARX - current status and challenges ahead. *Software: Practice and Experience*, 50(7):1277–1304, 2020.

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019.

[32] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6): 1010–1027, 2001.

[33] David Sánchez, Sergio Martínez, Josep Domingo-Ferrer, Jordi Soria-Comas, and Montserrat Batet. μ-ant: semantic microaggregation-based anonymization tool. *Bioinformatics*, 36(5):1652–1653, 2020.

[34] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S Baker. De-identifying student personally identifying information with gpt-4. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 559–565, 2024.

[35] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Language Models are Advanced Anonymizers. In *The Thirteenth International Conference on Learning Representations*, 2024.

[36] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36, 2015. doi: 10.18637/jss.v067.i04.

[37] Hui Yang and Jonathan M Garibaldi. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38, 2015. ISSN 1532-0464.

[38] Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1404.

[39] Vithya Yogarajan, Michael Mayo, and Bernhard Pfahringer. A survey of automatic de-identification of longitudinal clinical narratives. *arXiv:1810.06765*, 2018.