PrivBench: A Holistic and Modular Benchmarking Platform for Evaluating Text-to-Text Privatization

Stephen Meisenbacher[1,2,3,4], **Andreea-Elena Bodea**[2], **Ahmet Bilal Akın**[1], **Alexandra Klymenko**[1], **Jana Diesner**[1,2,3,4], and **Florian Matthes**[1]
[1]Technical University of Munich, School of Computation, Information and Technology
[2]Technical University of Munich, School of Social Sciences and Technology
[3]Munich Center for Machine Learning, Munich, Germany
[4]Munich Data Science Institute, Munich, Germany
Correspondence: stephen.meisenbacher@tum.de

Abstract

Natural Language Processing methods have enabled novel solutions and advances in the field of privacy, particularly in the sub-domain of *text-to-text privatization*, where the goal is to transform a sensitive input text into a privatized output by ideally masking (in)directly identifiable or otherwise private information. The evaluation of text-to-text privatization, however, is not straightforward, and the extant literature has utilized a myriad of techniques and metrics to quantify the privacy-preserving capabilities of privatization methods. Seeking to unify the evaluation of text-to-text privatization, we introduce PrivBench, a holistic and modular benchmarking platform for researchers and practitioners working on text privatization. PrivBench is holistic in that it evaluates privatization on a series of defined desiderata, which are structured into *modules*. PrivBench is not only modular but also extensible, allowing for future updates and benchmark versions. PrivBench is user-centered and promotes competition via real-time evaluation and a live public leaderboard. The platform is free to use and openly accessible at `https://privbench.com/`.

1 Introduction

The field of privacy-preserving Natural Language Processing (PPNLP) has grown in recent years (Habernal et al., 2023; Sousa and Kern, 2023), in line with the increase in NLP and LLM applications. Under the umbrella of PPNLP fall a variety of text privatization methods, including classic anonymization, Named Entity Recognition-based techniques, and modern methods leveraging LLMs as text privatizers (Deußer et al., 2025). Of particular interest are *text-to-text* privatization methods, where an input text is transformed, rewritten, or changed in a manner that preserves privacy, while also ideally maintaining the original usefulness and meaning of the text (Mattern et al., 2022b).

While the design of effective text-to-text privatization methods is challenging in its own right, the *evaluation* of these methods presents an additional hurdle (Ren et al., 2025): beyond the general difficulty with accurately incorporating societal notions of "privacy" into evaluation procedures, the particular field of text privatization must not only measure privacy protection, but also balance privacy with utility preservation, often referred to as the *privacy-utility trade-off* (Loiseau et al., 2025). In this confluence of privacy and utility, numerous evaluation approaches have been designed for text privatization, with little to no uniformity (Ren et al., 2025; Loiseau et al., 2025).

This lack of standardization is coupled with a lack of accessible mechanisms to comparatively analyze the benefits and limitations of text-to-text privatization methods across a wide variety of disciplines, such as anonymization (Pilán et al., 2022), Differential Privacy (Klymenko et al., 2022), or synthetic data (Yue et al., 2023). This differs from other sub-fields in NLP that have long-standing and competitive benchmark frameworks, such as GLUE and MMLU for Natural Language Understanding (Wang et al., 2018; Hendrycks et al., 2021), or BIG-Bench (Srivastava et al., 2023) and HELM (Liang et al., 2022) for more comprehensive Language Model benchmarking.

To address these shortcomings, we introduce PrivBench, a fully operational benchmarking platform that centers on the holistic evaluation of text-to-text privatization methods across a range of privacy and utility evaluation modules. The platform hosts and runs these benchmarks in real time upon user submission. PrivBench features the ability to create and manage user accounts, and each user can choose to feature their submissions on a public leaderboard, allowing for competition and comparability. Because the platform is modular, it is also extensible, allowing for the addition, editing, or removal of evaluation modules in future updates,

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 391–413
October 24-29, 2026 ©2026 Association for Computational Linguistics

tracked by a versioning system. Thus, the mission of PrivBench is not only to provide a unified benchmarking resource, but also to foster community engagement and growth. Our contributions are:

1. We extend previous work with a comprehensive survey of evaluation techniques and metrics for text-to-text privatization.
2. We leverage these theoretical foundations to design and implement a fully operational and deployed benchmarking platform: PrivBench.
3. We make the platform openly accessible, free, and available for immediate use, found at `https://privbench.com/`.
4. We test our holistic benchmark in a comparative case study of 14 privatization methods.

2 Related Work

Although the core of prior work in text privatization focuses on the evaluation of the proposed methods, some recent papers either systematize evaluation in general, or propose new benchmarks for text privatization. In a recent survey, Ren et al. (2025) review 47 related publications to investigate how privacy in text is evaluated, and categorize privacy metrics into six categories. Their survey sheds light on the multi-faceted nature of text privatization evaluation. Other surveys also highlight the importance of evaluation in PPNLP (Klymenko et al., 2022; Igamberdiev et al., 2022; Sousa and Kern, 2023).

Benchmark frameworks have been published in other PPNLP sub-domains. Focusing on privacy in language models, PrivLM-Bench (Li et al., 2024a), PrivacyLens (Shao et al., 2024b), and PrivAuditor (Zhu et al., 2024a) have been proposed. However, these benchmarks do not address text-to-text privatization. Other papers propose benchmarks for specific frameworks like Differential Privacy (Igamberdiev et al., 2022; Meisenbacher et al., 2024d; Sun et al., 2025b), which lack generalizability to the wider spectrum of text-to-text privatization.

More recently, Loiseau et al. (2025) proposed TauEval, a framework for the unified evaluation of text-to-text privatization approaches. Their framework evaluates privatization methods across eight different tasks, which focus on different aspects of *utility* given privatized texts. As with Ren et al. (2025), the importance of multi-dimensional privacy evaluation is supported. However, Loiseau et al. (2025) emphasize utility-based evaluation over privacy metrics, and the choice of metrics is not grounded in a review of literature. Furthermore,

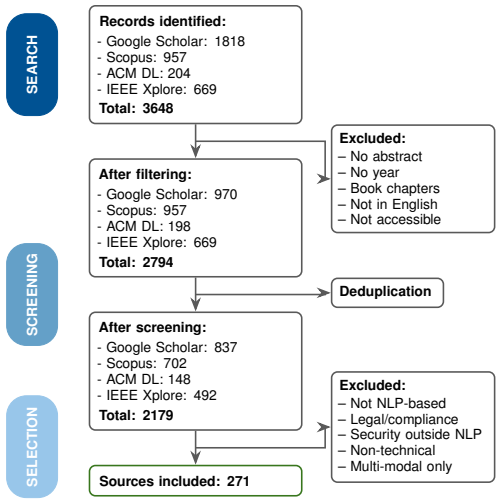

Figure 1: Systematic literature review selection process.

TauEval is released as a Python software package without a public-facing platform, as is present in benchmarks for many other NLP fields.

We build on these prior contributions by proposing a comprehensive and holistic text-to-text privatization benchmark that is (1) modular to reflect the many dimensions of privacy evaluation, (2) publicly accessible, no-code, and interactive to promote competition and to connect researchers, and (3) extensible to allow for improvements and version updates as text privacy evaluation evolves.

3 Evaluating Text Privatization

Literature review. To ground our platform in a holistic evaluation of text-to-text privatization, we conducted a systematic literature review of 271 papers. From this, we aggregate evaluation strategies to inform the modular approach of PrivBench.

We conducted our literature review according to the systematic approach outlined by Kitchenham et al. (2015). The first step was the design of a search string to probe relevant academic literature databases. This was done collaboratively among our research team, where the focus was to include key terms that indicate NLP *methods* and privacy. The general focus was scoped to ensure a comprehensive selection of publications. The final search string was:

("Natural language processing" OR "NLP" OR "Language models" OR "Language" OR "Text" OR "Document" OR "Information extraction" OR "Named entity recognition" OR "Part-of-speech tagging" OR "Sentiment analysis" OR "Machine translation" OR "Information retrieval") AND ("Privacy" OR "Private" OR "Privatization" OR "Data privacy" OR "Privacy-preserving" OR "Privacy-enhancing technologies" OR "PETs" OR "Privacy concerns" OR "Data protection" OR "Anonymization" OR "Rewriting" OR "Sanitization" OR "Obfuscation" OR "Scrubbing")

Using this string, we queried four prominent academic databases for technical papers: *Google*

Category	Module	Description	#	Representative Papers	Dataset(s)
Privacy	Attribute Inference (ATT)	Measures resistance to authorship inference attacks. A pretrained dataset-specific DeBERTa authorship identification classifier[1] is run on both the original and privatized texts, and the micro-F1 scores are compared. The score reflects how much the classifier's accuracy degrades after privatization; a higher score means better privacy protection.	31	(Macko et al., 2024; Meisenbacher et al., 2025d)	Reddit, Yelp
	Masked Token Inference (MTI)	Measures resistance to token-level inference attacks. Each token in the privatized text is masked one at a time, and a masked language model attempts to predict the masked token. The predicted tokens are then compared against the corresponding original text tokens.	5	(Tong et al., 2025b; Meisenbacher et al., 2025d; Chen et al., 2023; Yue et al., 2021)	WikiText, TAB
	Nearest Neighbor Privacy (NNP)	Measures resistance to record-linkage attacks by testing whether original texts can be matched back to their privatized counterparts through semantic similarity. Texts are embedded using a sentence transformer, and for each original text, its nearest neighbor rank among the privatized texts is computed. A higher score indicates that linking original to privatized texts is difficult.	3	(Meisenbacher et al., 2025d; Carvalho et al., 2023; Xu et al., 2021)	WikiText, Yelp
	Private Entity Masking (PEM)	Measures the effectiveness of named entity removal. A spaCy named entity recognition model is applied to the original texts to identify all named entities. Each detected entity is then checked for presence in the corresponding privatized text using case-insensitive matching. The score represents the proportion of original entities that were successfully removed or obfuscated.	11	(Vinod et al., 2025; Mancera et al., 2025; Vats et al., 2024; Smith et al., 2024; Qu et al., 2021)	WikiText, TAB
	Length Robustness (LRO)	Measures the consistency of a module across different text lengths. Texts are grouped into bins by word count, and the module is evaluated separately on each bin. The final score combines the average performance across bins with a consistency penalty that penalizes large variation.	-	Meta-metric: builds on top of another module (in v1.0.0: Private Entity Masking).	Yelp
Utility	Text Coherence (COH)	Measures the naturalness and fluency of privatized texts using perplexity of a GPT-2 model. Perplexity is computed for both the original and privatized texts, and the score reflects the ratio between them.	37	(Awon et al., 2025; Flemings et al., 2024; Vats et al., 2024; Meisenbacher and Matthes, 2024b; Li et al., 2023a)	IMDb, WikiText
	MAUVE (MAU)	Measures the distributional similarity between the original and privatized text collections as a whole. Rather than comparing individual text pairs, MAUVE evaluates whether the overall distribution of privatized texts resembles the distribution of originals in the embedding space.	12	(Tong et al., 2025b; Vinod et al., 2025; Carranza et al., 2024; Xie et al., 2024; Uchendu et al., 2023)	WikiText, Yelp
	Similarity (SIM)	Measures how well the privatized texts preserve the meaning and content of the originals. Two complementary metrics are combined: BLEU score captures lexical overlap at the n-gram level, while cosine similarity between sentence embeddings captures semantic preservation. The final score is the average of both metrics.	49	(Tong et al., 2025b; Higashi and Nakai, 2025; Awon et al., 2025; Huang et al., 2024b; Meisenbacher et al., 2024c; Yu et al., 2022)	PubMedQA, Yelp
	Task Utility (TUT)	Measures the degree to which downstream task performance can be preserved after privatization. A fine-tuned model for sentiment analysis[2] (IMDb) or question answering[3] is run on both the original and privatized texts, and the micro-F1 scores are compared. The score reflects how much the classifier's accuracy degrades after privatization.	147[4]	The large majority of works perform some manner of downstream task utility.	PubMedQA, IMDb
	Length Variation (LVA)	Measures whether privatized texts maintain appropriate length compared to the originals. The module computes the ratio of privatized to original text length for each pair and evaluates both the mean ratio and its variance. A mean ratio close to one indicates that privatized texts are neither dramatically shorter nor longer than the originals.	-	Meta-metric: newly proposed, motivated by the discussions of Mattern et al. (2022b), Meisenbacher et al. (2024c), and others.	WikiText, Yelp

Table 1: Overview of the 10 evaluation modules of PrivBench v1.0.0. Modules are balanced between *privacy* and *utility* metrics. # indicates the number of papers in which this metric was utilized, according to our literature review.

Scholar, *Scopus*, *ACM Digital Library*, and *IEEEXplore*. Note that the ACL Anthology is indexed by Google Scholar. We queried for title only, in order to maintain a manageable number of results. The final search was performed in June 2026.

From the four databases, the results (3648 in total) were filtered for sources without a year or abstract, published as a book, not written in English, not publicly accessible, or otherwise clearly lacking quality. The resulting set (2794) were combined and deduplicated based on title. This resulted in a filtered set of 2179 papers. These papers were all screened by three researchers, who read the title and abstract of each paper and applied inclusion/exclusion criteria to vote on candidacy for inclusion. The primary inclusion criteria was sources researching text-to-text privatization, whereas exclusion criteria are outlined in Figure 1.

Papers receiving a unanimous three votes were automatically included. Any paper receiving one or two votes was reviewed again by the primary researcher, who made an executive decision based on team discussion. This process resulted in a final set of 271 candidates for full-text reading.

All 271 papers were read by one researcher (as the qualitative coding required in-depth domain knowledge) (Bernard et al., 2016), who screened specifically for evaluation approaches undertaken in the publication. These approaches were tabulated, enumerated, and later aggregated. The datasets used for evaluation were also noted. 40 papers were double-coded by the primary researcher to ensure agreement. All extracted evaluation methods can be found in Table 3 of the Appendix; there are 35 in total, and they are divided into six categories. We note that only evaluation approaches appearing in >2 papers were included, in order to manage a large number of singleton approaches.

From this, we hand-selected 10 evaluation approaches to be operationalized in the first version of PrivBench, justified and briefly introduced next.

A holistic set of text-to-text privatization evaluation techniques. The 10 distinct approaches represent a diverse selection across privacy and utility metrics appropriate for text-to-text privatization. These were selected in favor of approaches that have been more widely adopted (and thus validated by researcher adoption), and to balance the evaluation of privacy protection *and* utility preservation.

From the selected metrics, we sought to find a minimal set of datasets that have been used for the respective metrics in the literature. This resulted in six total datasets. All 10 approaches are introduced in Table 1 with representative sources and corresponding datasets. These form the foundation of the evaluation *modules* of PrivBench.

4 System Design

PrivBench is implemented as a full-stack web application, utilizing a JavaScript (React) frontend, a Python (Flask) backend, a PostgreSQL database, and a Celery task management system with Redis.

[1] https://huggingface.co/sjmeis/yelp_authorship_small and https://huggingface.co/sjmeis/reddit-mental-health_authorship/

[2] https://huggingface.co/sjmeis/imdb_sentiment_small

[3] https://huggingface.co/sjmeis/pubmedqa_base/

[4] Based on the *Accuracy* metric alone.

System architecture. The web application is set up as a traditional full-stack application with a frontend, backend, and persistent database. The application is running on a local server with a 32-core CPU, 32 GB of RAM, and a single Nvidia RTX 5060 TI 16GB GPU, as well as 2TB of storage. Future plans include to migrate the system to a setup that can facilitate more concurrent requests.

The web application is dockerized for easy management and deployment. Each module is spawned as its own separate Docker container, allowing for task queuing and parallel benchmarking. Celery is used as a distributed task queue, where a user submission is distributed to all modules for evaluation. Redis is utilized as a message broker between the application and Celery workers.

PrivBench is made publicly accessible via nginx, which is mapped to the custom `https://privbench.com` domain. The site's SSL certificate is checked or renewed daily via certbot.

User personas and user journey. For PrivBench, we envision two primary user personas: the researcher/practitioner (henceforth, researcher) and the administrator (admin). Their user journeys are narrated in the following, and these descriptions are mirrored by the snapshots of the PrivBench experience, provided in Figures 2 and 3.

Researcher. The (privacy) researcher is the primary target audience of PrivBench, where the goal is to enable a standardized evaluation of text privatization methods in development, which in turn enables comparability to existing methods.

The researcher starts the user journey by registering an account on the site, which must be verified via a unique code sent to the provided email address. Password resets are also possible. Once the user is verified, the account profile can be accessed, where a profile picture can be uploaded, and a short bio can be provided. In the account area, all user submissions (discussed next) can also be viewed.

The researcher sees an overview page of the current modules by clicking on *How does it work?* on the navigation bar. Then, to start a benchmarking submission, the *Submission* tab can be clicked. This page outlines a clear five-step submission process: (1) informational overview, (2) dataset download (i.e., the original text datasets to be privatized), (3) metadata input for a text-to-text privatization method (including from a saved template), (4) privatized dataset upload, and (5) evaluation progress overview (i.e., after submission). The user is informed if placed in the queue for any module (implying other user(s) are currently ahead in line).

Following the successful completion of a benchmarking run, an overview of the results is presented. The researcher is prompted to head to the account area to view the result, where there exists the option to make the result public, thereby placing it on the public leaderboard. After this, the researcher can then navigate to the *Rankings* tab, which houses the leaderboard. There, users can search and filter the table, and as an advanced feature, the scores can be modified via user-defined weights for each of the modules. Each public result can also be explored in a "detailed view", which gives an in-depth overview of the submission and scores.

Finally, navigating to the footer, the researcher can view PrivBench's version history, which outlines all published changes. There is also the ability to contact the admin team via a simple form, which sends a message to the PrivBench email inbox.

Admin. For the beta, admins serve as maintainers of PrivBench[1]. Admins manage all current benchmarking modules, which includes the ability to add, edit, or remove modules. When adding or editing modules, the admin is responsible for providing the source code (Python logic) and requirements files, as well as selecting appropriate dataset mappings. Whenever any changes are made to the collection of modules, the changes are not enforced until the admin publishes the updates. Publishing requires an admin to update the PrivBench version, following standard versioning convention (e.g., v1.1.2). Here, an admin must also provide version notes, which are published on the version history page. Upon publishing updates, users are informed in their submissions panel (as well as via email), and they are prompted to update their submissions to reflect the newest PrivBench version.

Admins are also able to manage datasets, which allows for future updates in which new or updated datasets are used for module execution. Likewise, admins can manage all users of the platform, with the primary capabilities being the deletion of users (for example, spam or stale users), as well as updating daily submission allowances. Finally, the admin panel features direct bridges to the underlying VM, where the admin can monitor, restart, or delete Docker containers and images (i.e., those running the modules). A simple hardware utilization dashboard enables the admin to track VM load.

[1] This will be revisited if PrivBench scales beyond the current beta version.

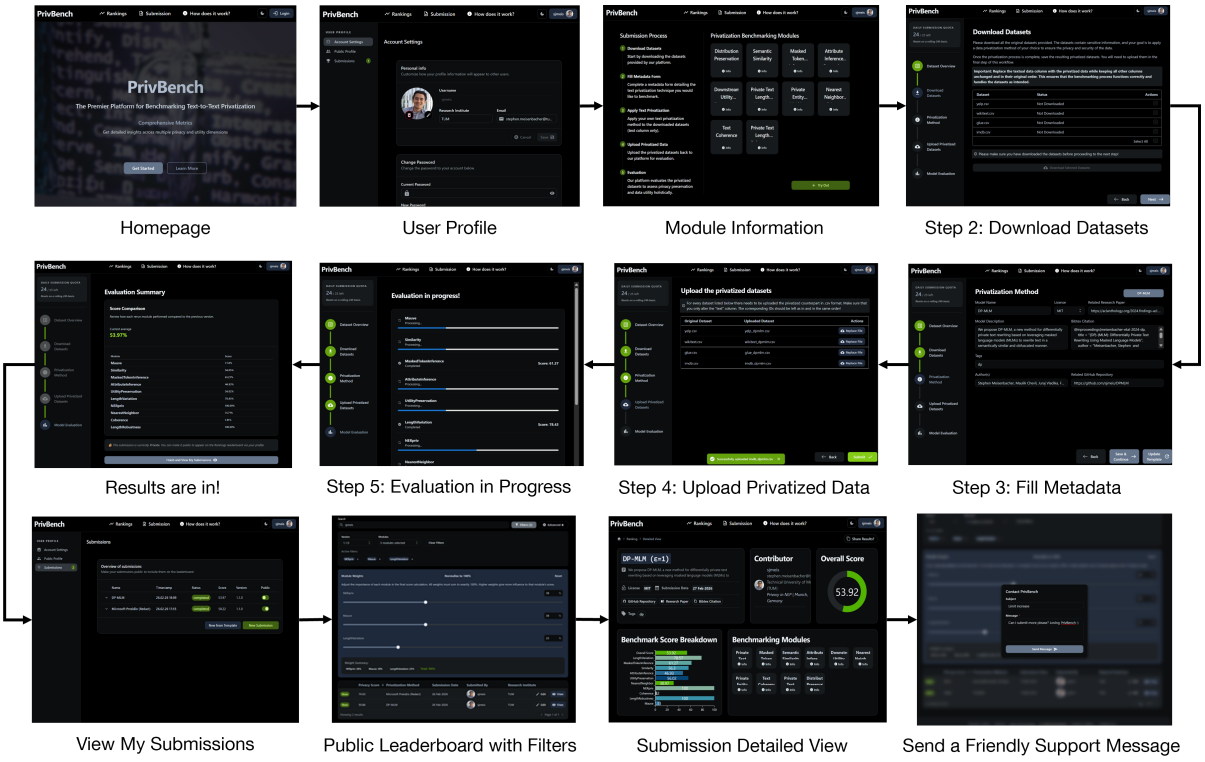

Figure 2: The *researcher* user journey.

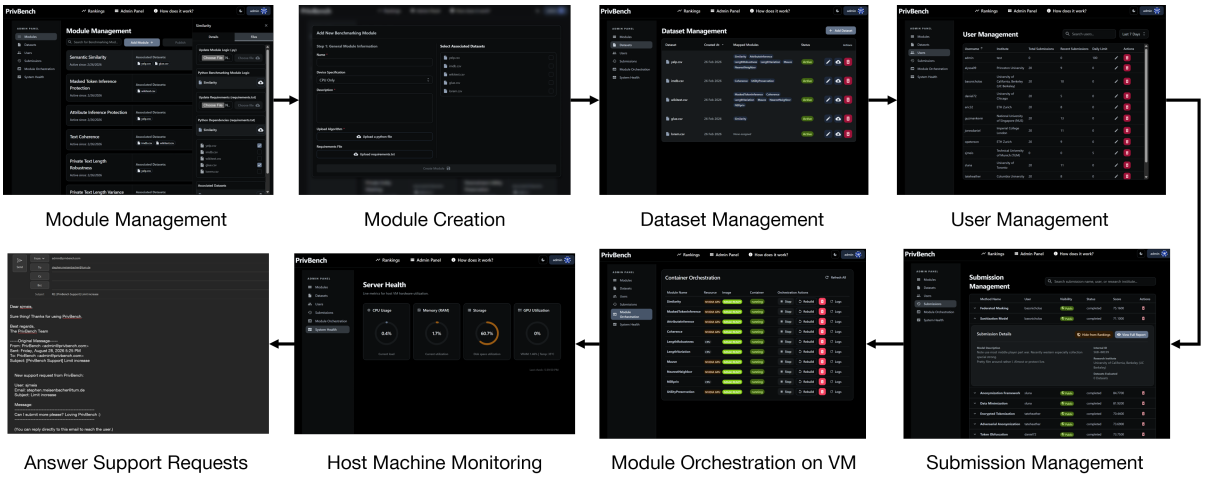

Figure 3: The *admin* user journey.

Versioning. As introduced in the admin journey, we have constructed PrivBench to evolve with the inevitable changes that will come in the understanding of how to best evaluate text privatization. As such, PrivBench features a git-style versioning system, which allows for the addition, removal, or updating of component evaluation modules. Once a change (or batch of changes) are published by an admin, PrivBench is upgraded to a new version (following the `Major.Minor.Patch` convention). Older versions are persisted, including the scores of all submissions under this version. Rollbacks are also possible via the admin panel. A complete account of the Version History is publicly viewable.

All platform users are notified via email when a new version of PrivBench is published. In the user's submissions panel, two actions are possible. In the case of a module *deletion*, the user can simply select to update the submission to the newest version (necessitating only a re-calculation of the overall score). However, if a module is updated or added (including dataset changes), the user will be prompted to re-run their submission, or if necessary, provided corresponding newly privatized

datasets. This system allows for the continual evolution of PrivBench to adapt with state-of-the-art research, but also for flexibility in the user experience to update submissions if and when desired.

License of use. We release the entire PrivBench codebase publicly, which includes the code for both the web application and the underlying benchmarking modules, as well as runner scripts for all tested privatization methods (see next Section 5). The code repository is released under a GPLv3 license. The repository can be found at: https://github.com/sjmeis/PrivBench

We are also committed to initially providing PrivBench as a free-to-use resource for researchers and practitioners in the privacy community. Thus, the beta release of PrivBench is provided under a permissive, yet non-commercial CC BY-NC-ND license. Although we impose a daily submission limit (default: 5) to enable fair resource sharing and to manage server load, we are open to requests for higher limits based on need. We make our Privacy Policy and Terms of Service clear.

Finally, we seek to gradually add admins and code contributors to the platform in order to create a diverse and rounded administrative team to foster the growth and sustainability of PrivBench. Members of the community are welcome to contact us if interested at admin@privbench.com.

5 A Comparative Case Study on Holistic Text Privatization Evaluation

To showcase the benefits of PrivBench, more specifically its purpose as an enabler of comparability between text-to-text privatization methods, we conduct a case study that evaluates 14 privatization methods across a diverse range of sub-disciplines.

Methods. We select 14 text-to-text privatization methods (29 configurations), which primarily stem from those screened during our literature review.

We test methods focusing on entity-based anonymization, including Presidio[4], Philter (Norgeot et al., 2020), Textwash (Kleinberg et al., 2022), and GLiNER (Zaratiana et al., 2024). These are run with all defaults of the public implementations.

We include several Differential Privacy-based methods, including the word-level 1-Diffractor (Meisenbacher et al., 2024a), the token-replacement methods of SanText(+) (Yue et al., 2021), CusText(+) (Chen et al., 2023), DP-Paraphrase (Mattern et al., 2022b), DP-Prompt (Utpala et al., 2023), DP-MLM (Meisenbacher et al., 2024c), the document-level ADePT (Krishna et al., 2021), and DP-BART (Igamberdiev and Habernal, 2023a). To ensure a fair evaluation, we utilize these methods with a document-level privacy parameter (ε) equal to the average word length per document per dataset (i.e., $\varepsilon = mean(\text{doc_length})$), as well as $2 \cdot \varepsilon$.

We also include RUPTA (Yang et al., 2025), a method that privatizes texts in a multi-LLM setup, and Privacy Filter[5], a new model from OpenAI.

With entity-based methods (Presidio, GLiNER, and Privacy Filter), we test two modes: (1) *redact*, where private entities are deleted, and (2) *replace*, where these are placeholdered, e.g., "[NAME]".

Datasets. Each PrivBench module is mapped to one or more datasets (Table 1). The current datasets used are IMDb (Maas et al., 2011), WikiText (Merity et al., 2017), Yelp (Zhang et al., 2015; Utpala et al., 2023), PubMedQA (Jin et al., 2019), TAB (Pilán et al., 2022), and Reddit Mental Health (Meisenbacher et al., 2025a). We take a 1000-row random sample from each dataset. In the case of multiple datasets being mapped to a module, the resulting scores are averaged over all datasets.

Results. The case study results are presented in Table 2, showing how our holistic approach enables a comprehensive as well as nuanced view of privatization performance. While DP-based methods are strong in ensuring privacy protections, they are not in utility metrics. Conversely, entity-based methods excel in utility but are poor in achieving meaningful privacy protection. The utility-preserving capabilities, however, dominate the privacy-utility balance, leading to higher overall scores.

These results highlight the great difficulty text privatization approaches have in preserving utility alongside strong privacy. Indeed, the two sides of Table 2 confirm that few methods can achieve consistently similar results in both. Thus, a crucial research gap is highlighted, namely the design of methods that empirically protect privacy, but not at the cost of extreme utility degradation. We propose PrivBench as a facilitator in tackling this challenge.

From case study to PrivBench v1.1.0. To internally verify the robustness and fairness of the PrivBench evaluation modules, we checked for

[4] https://github.com/data-privacy-stack/presidio

[5] https://openai.com/index/introducing-openai-privacy-filter/

	ATT	MTI	NNP	PEM	LRO	COH	MAU	SIM	TUT	LVA	μ 1.0	μ 1.1
Presidio (Redact)	0.0	58.9	0.1	52.0	29.0	74.3	87.0	**100**	**100**	72.2	57.4	47.8
Presidio (Replace)	0.4	57.3	0.1	51.5	28.8	**100**	87.7	**100**	**100**	72.1	59.8	51.2
Philter	0.4	57.2	0.2	29.0	9.2	99.6	93.8	**100**	99.9	71.8	56.1	48.3
Textwash	0.0	68.9	0.3	87.0	54.7	**100**	65.2	**100**	99.0	70.0	**64.5**	56.1
GLiNER (Redact)	0.0	59.8	0.2	23.0	19.7	75.6	82.3	**100**	**100**	73.2	53.4	46.9
GLiNER (Replace)	0.2	58.2	0.2	23.0	19.8	**100**	91.2	**100**	**100**	73.2	56.6	50.2
1-Diffractor (ε)	6.5	60.1	0.1	49.5	26.0	34.0	62.3	94.6	98.8	70.0	50.2	42.2
1-Diffractor ($2 \cdot \varepsilon$)	0.4	60.1	0.1	59.5	20.5	43.4	18.5	98.6	99.5	71.3	47.2	42.2
SanText (ε)	32.1	68.5	31.5	97.0	29.0	1.0	5.3	65.2	77.2	69.3	47.6	44.1
SanText ($2 \cdot \varepsilon$)	**37.9**	67.0	20.3	89.0	20.2	1.7	14.6	72.4	83.4	69.3	47.6	42.8
SanText+ (ε)	16.3	60.8	0.1	47.5	35.3	7.7	21.4	88.8	95.3	69.3	44.2	49.3
SanText+ ($2 \cdot \varepsilon$)	13.0	59.7	0.2	46.0	25.5	9.8	24.4	90.2	96.3	69.2	43.4	39.1
CusText (ε)	23.5	62.0	0.8	93.0	77.8	1.8	6.7	78.4	84.7	69.3	49.8	45.7
CusText ($2 \cdot \varepsilon$)	23.8	64.8	0.6	92.0	78.0	1.9	5.9	79.4	85.1	69.3	50.1	46.2
CusText+ (ε)	28.7	66.1	0.3	93.5	86.2	4.8	10.6	81.3	92.4	69.3	53.3	49.7
CusText+ ($2 \cdot \varepsilon$)	30.1	66.7	0.2	93.0	83.7	5.4	10.4	82.4	93.1	69.3	53.4	49.8
DP-Paraphrase (ε)	32.3	70.3	9.1	97.0	88.8	21.8	7.4	65.7	80.4	76.4	54.9	54.2
DP-Paraphrase ($2 \cdot \varepsilon$)	35.9	67.8	9.5	96.0	88.8	36.0	10.9	64.1	78.5	**78.1**	56.6	57.8
DP-Prompt (ε)	31.6	69.2	21.3	92.0	65.3	2.0	15.2	78.1	86.8	59.0	52.0	47.9
DP-Prompt ($2 \cdot \varepsilon$)	29.6	69.9	4.5	83.0	33.5	16.7	25.9	83.2	91.5	50.4	48.8	42.3
DP-MLM (ε)	23.5	64.7	34.8	**100**	**100**	1.8	4.8	53.9	78.0	71.2	53.3	53.4
DP-MLM ($2 \cdot \varepsilon$)	21.7	64.2	31.5	**100**	**100**	1.9	4.3	54.3	76.6	71.5	52.6	52.7
ADePT (ε)	31.1	61.1	42.1	98.0	**100**	**100**	2.7	58.1	74.9	28.8	59.7	**62.6**
ADePT ($2 \cdot \varepsilon$)	30.1	74.3	39.6	97.5	56.8	**100**	2.3	61.4	74.9	33.2	57.0	58.4
DP-BART (ε)	28.2	**75.6**	**47.9**	**100**	**100**	0.1	5.0	52.3	77.5	64.2	55.1	56.2
DP-BART ($2 \cdot \varepsilon$)	28.7	68.6	46.9	**100**	**100**	0.1	5.3	52.2	76.2	64.1	54.2	55.0
RUPTA	2.4	57.6	0.2	49.0	16.8	**100**	69.4	**100**	99.9	73.9	56.9	50.1
Privacy Filter (Redact)	0.0	58.2	0.1	10.0	18.8	96.1	**99.9**	**100**	**100**	72.4	55.6	49.4
Privacy Filter (Replace)	0.0	57.9	0.1	10.0	19.0	**100**	99.8	**100**	**100**	72.2	55.9	49.9

Table 2: Case study results. All scores are out of 100 (↑), and the final score (μ) is the mean of all 10 module scores (v1.0.0), or the remaining seven modules in v1.1.0 (i.e., excluding PEM, MAU, and SIM, see Section 5). **Bold** scores denote the best per column. Module abbreviations are from Table 1.

inter-module redundancies, using the results of our case study as data points. With the scores of Table 2, we calculated the Spearman's rank correlation coefficient (r_s) between all modules. In particular, we calculated correlations separately between the five utility and five privacy modules. This analysis revealed five highly correlated modules ($r_s > 0.8, p < 0.05$), namely: SIM ↔ TUT (0.96), MAU ↔ TUT (0.94), MAU ↔ SIM (0.93), PEM ↔ LRO (0.90), and NNP ↔ PEM (0.85).

To address these redundancies and mitigate the risk of multiply rewarding the same evaluation signal, we propose an updated PrivBench v1.1.0 that removes SIM, MAU, and PEM, thus leaving three utility modules and four privacy modules. Note: in v1.1.0, PEM is still used as the basis for LRO.

The updated scores for v1.1.0 are also provided in Table 2. As shown there, the module validation (i.e., redundancy removal) shifts the landscape in terms of prevailing scores. We find that DP-based methods are able to reach more favorable balances, whereas the difficulties of non-DP methods in providing empirical privacy protections are now more prominently emphasized. Nevertheless, the overall score distributions still lend concrete evidence to the fact that text-to-text privatization has a significant challenge ahead in optimizing the delicate balance between privacy and utility.

6 Conclusion

We introduce PrivBench, a fully deployed, public, and free-to-use benchmarking platform motivated by the holistic and modular evaluation of text-to-text privatization methods. The base platform is released with 10 benchmarking modules in v1.0.0 and a reduced set of seven in v1.1.0. PrivBench creates a space for researchers and practitioners working in text privacy to collaborate in a competitive yet fair environment, and concurrently, for motivated administrators to ensure PrivBench's upkeep and to guide its future. Our case study showcases the opportunity of PrivBench to promote better comparability and higher transparency of text privatization effectiveness. In this spirit, we hope that PrivBench helps to push the boundaries of text privatization, and that it continues to evolve as the understanding of the evaluation thereof advances.

Limitations

As with the general evaluation of *text privatization*, PrivBench is limited to the current understanding of how to best evaluate "good" technical approaches to privacy protection in textual data. We mitigate this limitation via our modular approach, which holistically captures different aspects of text privatization evaluation. As such, PrivBench can be modified and updated as more evaluation strategies are developed (or suggested) by the community. We remain committed to facilitating continuous improvements to the modular benchmark, to build on the basis we establish with this first iteration.

We decided to strike a balance between *comprehensive* and *efficient* evaluation. As such, evaluation datasets (those associated with modules) are capped at 1000 rows/texts per module. While this eases the burden and turnaround time for users of the platform, we acknowledge that larger datasets would lead to more robust scoring. This remains a point for future work, especially if PrivBench is migrated to a host with higher resource capacities.

In the weighting of the different evaluation modules, we opted for a simple equal weighting of all the modules for the calculation of the overall score per submission. Future work may improve this simpler scheme by assigning relative weights, which could be obtained either via automatic means or through the modeling of human preferences.

We also acknowledge that due to the nature of the platform design, we cannot evaluate the computational efficiency of proposed text privatization methods, which would also be a key evaluation criterion, particularly for practical applicability.

Although we conducted our case study to validate the function and capabilities of PrivBench, we did not run user studies to validate its user-friendliness and general UX. This would represent a logical next step in the development of PrivBench, particularly as users begin to adopt the platform.

Ethical Considerations

We confirm that all underlying datasets that comprise the current PrivBench are used in accordance with their original (open-source) licenses, namely:

- IMDb: CC BY 4.0
- WikiText: CC BY-SA 4.0
- Yelp: non-commercial
- PubMedQA: MIT
- TAB: MIT
- Reddit: non-commercial

In solely using public research datasets, we mitigate and discourage adversarial or malicious use of PrivBench. As a result of this structure, we do not elicit or ask for the submission of proprietary or private data to participate in the benchmark. Proper use of the website involves only submitting privatized versions of the provided, publicly available datasets. We explicitly advise users against uploading personally identifiable information, proprietary content, or sensitive private data in lieu of the provided datasets. Any upload of such data is strictly at the user's sole discretion and risk.

We take further measures to prevent malicious use of our website, including strong access controls, bot prevention, and following standard secure web development practices. In addition, we invite and appreciate any reports of issues or potential security vulnerabilities. The Code of Conduct and contribution guidelines are found in our repository.

Finally, we make our Privacy Policy and Terms of Service transparent and clear, and we will work to keep these updated as PrivBench evolves.

Acknowledgments

We extend a heartfelt thank you to the following people for making PrivBench possible. Firstly, to Han My Do, Simeon Ivanov, Souraya Kharfi, and Tobias Unterhauser for the foundation. To James Flemings, Timour Igamberdiev, and Ibrahim Baroud for the support in the initial literature review and brainstorming. And to Muhammad Jibran for the systems support and the rest of sebis for the general support. We appreciate you all very much!

References

Maryam Abbasalizadeh and Sashank Narain. 2025. Privacy-aware detection for large language models using a hybrid bilstm-hmm approach. *IEEE Access*, 13:121880–121901.

David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2020. Privacy Guarantees for De-Identifying Text Transformations. In *Interspeech 2020*, pages 4666–4670.

David Ifeoluwa Adelani, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2021. Preventing author profiling through zero-shot multilingual back-translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8687–8695, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harshvardhan Aditya, Siddansh Chawla, Gunika Dhingra, Parijat Rai, Saumil Sood, Tanmay Singh, Zeba Mohsin Wase, Arshdeep Bahga, and Vijay K Madisetti. 2024. Evaluating privacy leakage and memorization attacks on large language models (llms) in generative ai applications. *Journal of Software Engineering and Applications*, 17(5):421–447.

Abdulrahman Alabdulkareem, Christian M Arnold, Yerim Lee, Pieter M Feenstra, Boris Katz, and Andrei Barbu. 2024. Securellm: Using compositionality to build provably secure language models for private, sensitive, and secret data. *arXiv preprint arXiv:2405.09805*.

Mohammed Alawad, Hong-Jun Yoon, Shang Gao, Brent Mumphrey, Xiao-Cheng Wu, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Linda Coyle, Lynne Penberthy, and Georgia Tourassi. 2021. Privacy-preserving deep learning nlp models for cancer registries. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1219–1230.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2022. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023a. Driving context into text-to-text privatization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 15–25, Toronto, Canada. Association for Computational Linguistics.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023b. Guiding text-to-text privatization by syntax. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 151–162, Toronto, Canada. Association for Computational Linguistics.

Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K. Goudos, Ioannis D. Moscholios, Konstantinos E. Psannis, and Panagiotis Sarigiannidis. 2024. Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches. In *2024 13th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–6.

Ahmed Musa Awon, Yun Lu, Shera Potka, and Alex Thomo. 2025. CluSanT: Differentially private and semantically coherent text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3676–3693, Albuquerque, New Mexico. Association for Computational Linguistics.

Ahmad Al Badawi, Louie Hoang, Chan Fook Mun, Kim Laine, and Khin Mi Mi Aung. 2020. Privft: Private and fast text classification with homomorphic encryption. *IEEE Access*, 8:226544–226556.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2023. Privacy adhering machine un-learning in NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 268–277, Nusa Dua, Bali. Association for Computational Linguistics.

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, and Aurélie Névéol. 2022. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130:104073.

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.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. Privacy preserving text representation learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 275–276.

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.

H Russell Bernard, Amber Wutich, and Gery W Ryan. 2016. *Analyzing qualitative data: Systematic approaches*. SAGE publications.

David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. Anonymization of german financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, 13(2):151–161.

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

Gerardo Canfora, Andrea Di Sorbo, Enrico Emanuele, Sara Forootani, and Corrado A Visaggio. 2018. A nlp-based solution to prevent from privacy leaks in social network posts. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pages 1–6.

Aldo Carranza, Rezsa Farahani, Natalia Ponomareva, Alexey Kurakin, Matthew Jagielski, and Milad Nasr. 2024. Synthetic query generation for privacy-preserving deep retrieval systems using differentially private language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3920–3930, Mexico City, Mexico. Association for Computational Linguistics.

Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 883–890. SIAM.

Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852.

Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Anima Anandkumar, and Fei Wang. 2024. Combating security and privacy issues in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 8–18, Mexico City, Mexico. Association for Computational Linguistics.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Tianyu Chen, Hangbo Bao, Shaohan Huang, Li Dong, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. 2022. THE-X: Privacy-preserving transformer inference with homomorphic encryption. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3510–3520, Dublin, Ireland. Association for Computational Linguistics.

Richard Chow, Ian Oberst, and Jessica Staddon. 2009. Sanitization's slippery slope: the design and study of a text revision assistant. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, New York, NY, USA. Association for Computing Machinery.

Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2024. pypantera: A python package for natural language obfuscation enforcing privacy & anonymization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5348–5353, New York, NY, USA. Association for Computing Machinery.

Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025a. A comparative study of large language models and traditional privacy measures to evaluate query obfuscation approaches. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2711–2716.

Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025b. Dp-comet: A differential privacy contextual obfuscation mechanism for texts in natural language processing. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 4700–4705, New York, NY, USA. Association for Computing Machinery.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. 2025. A survey on current trends and recent advances in text anonymization. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9.

Minxin Du, Xiang Yue, Sherman S. M. Chow, and Huan Sun. 2023. Sanitizing sentence embeddings (and labels) for local differential privacy. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2349–2359, New York, NY, USA. Association for Computing Machinery.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. 2024. On the privacy risk of in-context learning. *arXiv preprint arXiv:2411.10512*.

Qi Feng, Debiao He, Zhe Liu, Huaqun Wang, and Kim-Kwang Raymond Choo. 2020. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 178–186, New York, NY, USA. Association for Computing Machinery.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219.

Oluwaseyi Feyisetan and Shiva Kasiviswanathan. 2021. Private release of text embedding vectors. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 15–27, Online. Association for Computational Linguistics.

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024. JAMDEC: Unsupervised authorship obfuscation using constrained decoding over small language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1552–1581, Mexico City, Mexico. Association for Computational Linguistics.

James Flemings and Murali Annavaram. 2024. Differentially private knowledge distillation via synthetic text generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12957–12968, Bangkok, Thailand. Association for Computational Linguistics.

James Flemings, Meisam Razaviyayn, and Murali Annavaram. 2024. Differentially private next-token prediction of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4390–4404, Mexico City, Mexico. Association for Computational Linguistics.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, page 2490–2501, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. 2024. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 64–75, St. Julian's, Malta. Association for Computational Linguistics.

Antonio Ginart, Laurens van der Maaten, James Zou, and Chuan Guo. 2022. Submix: Practical private prediction for large-scale language models. *arXiv preprint arXiv:2201.00971*.

Kang Gu. 2024. *Towards Trustworthy LLMs: Understanding the Security and Privacy Risks of Large Language Models*. Ph.D. thesis, Dartmouth College.

Ivan Habernal, Fatemehsadat Mireshghallah, Patricia Thaine, Sepideh Ghanavati, and Oluwaseyi Feyisetan. 2023. Privacy-preserving natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–30, Dubrovnik, Croatia. Association for Computational Linguistics.

Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. Iron: private inference on transformers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fadi Hassan, David Sánchez, and Josep Domingo-Ferrer. 2023. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):1058–1071.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

Takuya Higashi and Tsunato Nakai. 2025. Is differential privacy-enhanced parameter-efficient fine-tuning effective for large language models? In *Proceedings of the Workshop on Privacy in Large Language Models (LLM) and Natural Language Processing (NLP) 2025*, LM-SHIELD '25, page 1–12, New York, NY, USA. Association for Computing Machinery.

Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shilong Hou, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. 2025a. A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation. *Preprint*, arXiv:2502.15233.

Yihao Hou, Christoph Bert, Ahmed Gomaa, Godehard Lahmer, Daniel Höfler, Thomas Weissmann, Raphaela Voigt, Philipp Schubert, Charlotte Schmitter, Alina Depardon, and 1 others. 2025b. Fine-tuning a local llama-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *Frontiers in Artificial Intelligence*, 7:1493716.

Yuzheng Hu, Ryan McKenna, Da Yu, Shanshan Wu, Han Zhao, Zheng Xu, and Peter Kairouz. 2026. ACTG-ARL: Differentially private conditional text generation with RL-boosted control. In *Forty-third International Conference on Machine Learning*.

Chao-Long Huang, Babam Rianto, Jun-Teng Sun, Zheng-Xin Fu, and Chung-Hong Lee. 2024a. Privacy protection and standardization of electronic medical records using large language model. In *International Workshop on Deidentification of Electronic Medical Record Notes*, pages 60–71. Springer.

Shuo Huang, William Maclean, Xiaoxi Kang, Qiongkai Xu, Zhuang Li, Xingliang Yuan, Gholamreza Haffari, and Lizhen Qu. 2025a. NAP2: A benchmark for naturalness and privacy-preserving text rewriting by learning from human. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8954–8970, Suzhou, China. Association for Computational Linguistics.

Shuo Huang, Xingliang Yuan, Gholamreza Haffari, and Lizhen Qu. 2025b. Zero-shot privacy-aware text rewriting via iterative tree search. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9175–9190, Suzhou, China. Association for Computational Linguistics.

Tianhao Huang, Tao Yang, Ivan Habernal, Lijie Hu, and Di Wang. 2024b. Private language models via truncated laplacian mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3993, Miami, Florida, USA. Association for Computational Linguistics.

Tianhao Huang, Tao Yang, Ivan Habernal, Lijie Hu, and Di Wang. 2024c. Private language models via truncated laplacian mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3993, Miami, Florida, USA. Association for Computational Linguistics.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2022. Privacy-preserving graph convolutional networks for text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 338–350, Marseille, France. European Language Resources Association.

Timour Igamberdiev and Ivan Habernal. 2023a. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023b. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Timour Igamberdiev, Doan Nam Long Vu, Felix Kuennecke, Zhuo Yu, Jannik Holmer, and Ivan Habernal. 2024. DP-NMT: Scalable differentially private machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 94–105, St. Julians, Malta. Association for Computational Linguistics.

Lubana Isaoglu and Zeynep Orman. 2026. A unified evaluation framework for utility and privacy risks of llm-generated synthetic text data. *Engineering Perspective*, 6(3):456–466.

Celestine Iwendi, Syed Atif Moqurrab, Adeel Anjum, Sangeen Khan, Senthilkumar Mohan, and Gautam Srivastava. 2020. N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. *Computer Communications*, 161:160–171.

Wei Jiang, Mummoorthy Murugesan, Chris Clifton, and Luo Si. 2009. t-plausibility: Semantic preserving text sanitization. In *2009 International Conference on Computational Science and Engineering*, volume 3, pages 68–75.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Rodrigo Juez-Hernandez, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. Agora: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. *Applied Soft Computing*, 145:110540.

Sandeep Kalari, Sahithi Padidela, Vikas Ashok, and Ravi Mukkamala. 2026. Revisiting data anonymization: Limitations and challenges in the era of large language models. In *2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*, page 1166–1173.

Supriya Kamoji, Rohan Mathew, Justin Abreo, Jaden Dsa, Heenakausar Pendhari, and Unik Lokhande. 2024. A privacy preserving sentiment analysis using federated learning. In *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, pages 510–515.

Ioannis Karamitsos, Nikolaos Roufas, Khalil Al-Hussaeni, and Andreas Kanavos. 2025. Legner: a domain-adapted transformer for legal named entity recognition and text anonymization. *Frontiers in Artificial Intelligence*, 8.

Nikos Karousos, George Vorvilas, Despoina Pantazi, and Vassilios S. Verykios. 2025. Evaluating ner approaches to support qualitative educational data anonymization in the greek language: A comparative study. In *2025 16th International Conference on Information, Intelligence, Systems & Applications (IISA)*, page 1–8.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

Umair Khadam, Muhammad Munwar Iqbal, Meshrif Alruily, Mohammed A Al Ghamdi, Muhammad Ramzan, and Sultan H Almotiri. 2020. Text data security and privacy in the internet of things: threats, challenges, and future directions. *Wireless Communications and Mobile Computing*, 2020(1):7105625.

Umair Khadam, Muhammad Munwar Iqbal, Saqib Saeed, Saadat Hanif Dar, Awais Ahmad, and Mudassar Ahmad. 2021. Advanced security and privacy technique for digital text in smart grid communications. *Computers & Electrical Engineering*, 93:107205.

Mandar Khoje. 2024. Navigating data privacy and analytics: The role of large language models in masking conversational data in data platforms. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, pages 1–5.

Jongwook Kim. 2025. Privrewrite: Differentially private text rewriting under black-box access with refined sensitivity guarantees. *Applied Sciences*, 15(22):11930.

Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash – automated open-source text anonymisation. *Preprint*, arXiv:2208.13081.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing: The story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Martin Kuo, Jingyang Zhang, Jianyi Zhang, Minxue Tang, Louis DiValentin, Aolin Ding, Jingwei Sun, William Chen, Amin Hass, Tianlong Chen, Yiran Chen, and Hai Li. 2025. Proactive privacy amnesia for large language models: Safeguarding PII with negligible impact on model utility. In *The Thirteenth International Conference on Learning Representations*.

Phung Lai, NhatHai Phan, Tong Sun, Rajiv Jain, Franck Dernoncourt, Jiuxiang Gu, and Nikolaos Barmpalios. 2022. User-entity differential privacy in learning natural language models. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1465–1474.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024a. PrivLM-bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 54–73, Bangkok, Thailand. Association for Computational Linguistics.

Jianwei Li, Sheng Liu, and Qi Lei. 2023a. Beyond gradient and priors in privacy attacks: Leveraging pooler layer inputs of language models in federated learning. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Marvin Li, Jason Wang, Jeffrey Wang, and Seth Neel. 2023b. MoPe: Model perturbation based privacy attacks on language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13647–13660, Singapore. Association for Computational Linguistics.

Mingchen Li, Heng Fan, Song Fu, Junhua Ding, and Yunhe Feng. 2025. DP-GTR: Differentially private prompt protection via group text rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1573–1585, Suzhou, China. Association for Computational Linguistics.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He,

and Dawn Song. 2024b. Llm-pbe: Assessing data privacy in large language models. *Proc. VLDB Endow.*, 17(11):3201–3214.

Xianzhi Li, Ran Zmigrod, Zhiqiang Ma, Xiaomo Liu, and Xiaodan Zhu. 2024c. Fine-tuning language models with differential privacy through adaptive noise allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8368–8375, Miami, Florida, USA. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Zi Liang, Pinghui Wang, Ruofei Zhang, Nuo Xu, Shuo Zhang, Lifeng Xing, Haitao Bai, and Ziyang Zhou. 2024. Merge: Fast private text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19884–19892.

Shaobo Liu, Guiran Liu, Binrong Zhu, Yuanshuai Luo, Linxiao Wu, and Rui Wang. 2024. Balancing innovation and privacy: Data security strategies in natural language processing applications. In *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, pages 609–613. IEEE.

Yanxin Liu and Qianqian Su. 2024. Pptif: Privacy-preserving transformer inference framework for language translation. *IEEE Access*, 12:48881–48897.

Zhe Liu and Ozlem Kalinli. 2024. Forgetting private textual sequences in language models via leave-one-out ensemble. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10261–10265.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. 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. Association for Computational Linguistics.

Jinglong Luo, Yehong Zhang, Zhuo Zhang, Jiaqi Zhang, Xin Mu, Hui Wang, Yue Yu, and Zenglin Xu. 2024. SecFormer: Fast and accurate privacy-preserving inference for transformer models via SMPC. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13333–13348, Bangkok, Thailand. Association for Computational Linguistics.

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.

Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason S Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship obfuscation in multilingual machine-generated text detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6348–6368, Miami, Florida, USA. Association for Computational Linguistics.

Pulkit Madaan, Krithika Ramesh, Lisa Bauer, Charith Peris, and Anjalie Field. 2026. Multi-token completion for text anonymization. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 5894–5908, Rabat, Morocco. Association for Computational Linguistics.

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, and Aurélien Bellet. 2022. Fair NLP models with differentially private text encoders. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6913–6930, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. 2024. Split-and-denoise: protect large language model inference with local differential privacy. *arXiv preprint arXiv:2310.09130*.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621*.

Gonzalo Mancera, Aythami Morales, Julian Fierrez, Ruben Tolosana, Alejandro Peña, Miguel Lopez-Duran, Francisco Jurado, and Alvaro Ortigosa. 2025. PBa-LLM: Privacy-and bias-aware nlp using named-entity recognition (ner). In *International Conference on Document Analysis and Recognition*, pages 3–20. Springer.

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

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schölkopf, and Mrinmaya Sachan. 2022a. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022b. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Cleo Matzken, Steffen Eger, and Ivan Habernal. 2023. Trade-offs between fairness and privacy in language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6948–6969, Toronto, Canada. Association for Computational Linguistics.

Andrew J McMurry, Britt Fitch, Guergana Savova, Isaac S Kohane, and Ben Y Reis. 2013. Improved de-identification of physician notes through integrative modeling of both public and private medical text. *BMC medical informatics and decision making*, 13(1):112.

Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. Sentence-level privacy for document embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3367–3380, Dublin, Ireland. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024a. 1-diffractor: Efficient and utility-preserving text obfuscation leveraging word-level metric differential privacy. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, IWSPA '24, page 23–33, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024b. A collocation-based method for addressing challenges in word-level metric differential privacy. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 39–51, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025a. Leveraging semantic triples for private document generation with local differential privacy guarantees. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8976–8992, Suzhou, China. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025b. On the impact of noise in differentially private text rewriting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 514–532, Albuquerque, New Mexico. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024c. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9314–9328, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Meisenbacher, Angelo Kleinert, and Florian Matthes. 2026. A systematic exploration of text decomposition and budget distribution in differentially private text obfuscation. In *Proceedings of the Seventh Workshop on Privacy in Natural Language Processing*, pages 118–139, San Diego, California. Association for Computational Linguistics.

Stephen Meisenbacher, Alexandra Klymenko, Andreea-Elena Bodea, and Florian Matthes. 2025c. The double-edged sword of llm-based data reconstruction: Understanding and mitigating contextual vulnerability in word-level differential privacy text sanitization. In *Proceedings of the 24th Workshop on Privacy in the Electronic Society*, WPES '25, page 162–176, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher, Chaeeun Joy Lee, and Florian Matthes. 2025d. Spend your budget wisely: Towards an intelligent distribution of the privacy budget in differentially private text rewriting. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '25, page 84–95, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher and Florian Matthes. 2024a. Just rewrite it again: A post-processing method for enhanced semantic similarity and privacy preservation of differentially private rewritten text. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher and Florian Matthes. 2024b. Thinking outside of the differential privacy box: A case study in text privatization with language model prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5656–5665, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024d. A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 174–185, Torino, Italia. ELRA and ICCL.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Tomoaki Mimoto, Shinsaku Kiyomoto, Koji Kitamura, and Atsuko Miyaji. 2020. A practical privacy-preserving algorithm for document data. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1376–1383.

Abhijit Mishra, Mingda Li, and Soham Deo. 2024. Sentinellms: encrypted input adaptation and fine-tuning of language models for private and secure inference. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Kushagra Mishra, Harsh Pagare, and Kanhaiya Sharma. 2025. A hybrid rule-based nlp and machine learning approach for pii detection and anonymization in financial documents. *Scientific Reports*, 15(1):22729.

Amirhossein Ebrahimi Moghaddam, Buvana Ganesh, and Paolo Palmieri. 2024. Privacy-preserving sentiment analysis using homomorphic encryption and attention mechanisms. In *International Conference on Applied Cryptography and Network Security*, pages 84–100. Springer.

Gabriele Morabito, Valeria Lukaj, Armando Ruggeri, Maria Fazio, Maria Annunziata Astone, and Massimo Villari. 2023. Docflow: Supervised multi-method document anonymization engine. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6.

Nina Mouhammad, Johannes Daxenberger, Benjamin Schiller, and Ivan Habernal. 2023. Crowdsourcing on sensitive data with privacy-preserving text rewriting. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 73–84, Toronto, Canada. Association for Computational Linguistics.

Balázs Nagy, István Hegedundefineds, Noémi Sándor, Balázs Egedi, Haaris Mehmood, Karthikeyan Saravanan, Gábor Lóki, and Ákos Kiss. 2023. Privacy-preserving federated learning and its application to natural language processing. *Know.-Based Syst.*, 268(C).

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, and 1 others. 2020. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57.

Sebastian Ochs and Ivan Habernal. 2025. Private synthetic text generation with diffusion models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 10612–10626, Albuquerque, New Mexico. Association for Computational Linguistics.

Vladimir Panov, Mikhail Kovalchuk, Anastasiia Filatova, and Sergey Teryoshkin. 2022. Mucaat: Multilingual contextualized authorship anonymization of texts from social networks. *Procedia Computer Science*, 212:322–329.

Anthi Papadopoulou, Pierre Lison, Mark Anderson, Lilja Øvrelid, and Ildikó Pilán. 2026. Neural text sanitization with privacy risk indicators: An empirical analysis. *Language Resources and Evaluation*, 60(2):32.

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.

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

Richard Plant, Valerio Giuffrida, and Dimitra Gkatzia. 2025. You are what you write: Author re-identification privacy attacks in the era of pre-trained language models. *Computer Speech & Language*, 90:101746.

Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. CAPE: Context-aware private embeddings for private language learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7970–7978, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. 2022. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193, Dublin, Ireland. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Evgenia Psarra and Kyriakos Stefanidis. 2026. On the applicability of llms and slms for privacy-preserving named entity recognition in financial applications. *Applied Sciences*, 16(7):3332.

Ian Pépin, Furkan Alaca, and Farhana Zulkernine. 2024. Privacy-preserving multi-party keyword-based classification of unstructured text data. In *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pages 203–210.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural

language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1488–1497, New York, NY, USA. Association for Computing Machinery.

David Raffetseder, Carl Weilguny, Patrick Haidinger, Hans-Peter Pichler, and Wolfgang Narzt. 2025. Support ticket anonymization: advancing data privacy with transformer-based named entity recognition. *Proceedings of the 58th Hawaii International Conference on System Sciences*.

Yaxuan Ren, Krithika Ramesh, Yaxing Yao, and Anjalie Field. 2025. How do we measure privacy in text? a survey of text anonymization metrics. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1532–1544, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Amanda Resende, Davis Railsback, Rafael Dowsley, Anderson C. A. Nascimento, and Diego F. Aranha. 2022. Fast privacy-preserving text classification based on secure multiparty computation. *IEEE Transactions on Information Forensics and Security*, 17:428–442.

Alireza Salemi and Hamed Zamani. 2025. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 286–296.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David Sánchez and Montserrat Batet. 2017. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34.

David Sánchez, Montserrat Batet, and Alexandre Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics*, 52:189–198.

Taisho Sasada, Masataka Kawai, Yuzo Taenaka, Doudou Fall, and Youki Kadobayashi. 2021. Differentially-private text generation via text preprocessing to reduce utility loss. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 042–047.

Manuel Senge, Timour Igamberdiev, and Ivan Habernal. 2022. One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7340–7353, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sakib Shahriar, Rozita Dara, and Fattane Zarrinkalam. 2025. Mitigating authorship attribution in social media texts: A privacy by translation approach. *IEEE Access*, 13:185709–185725.

Hanyin Shao, Jie Huang, Shen Zheng, and Kevin Chang. 2024a. Quantifying association capabilities of large language models and its implications on privacy leakage. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 814–825, St. Julian's, Malta. Association for Computational Linguistics.

Wei Shao, Lemao Liu, Yinqiao Li, Guoping Huang, Shuming Shi, and Linqi Song. 2026. Towards privacy-preserving machine translation at the inference stage: A new task and benchmark. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–15.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024b. PrivacyLens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022a. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. 2022b. Just fine-tune twice: Selective differential privacy for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6340, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhan Shi, Yefeng Yuan, Liang Cheng, and Yuhong Liu. 2025. Reinforcement learning-guided large language model fine-tuning for privacy-preserving text rewriting. In *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing*, SEC '25, page 1–7, New York, NY, USA. Association for Computing Machinery.

Mohammad Shokri, Sarah Ita Levitan, and Rivka Levitan. 2025. Personalized author obfuscation with large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1153–1162, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A N Ramya Shree and P Kiran. 2020. Sensitivity context aware privacy preserving text document summarization. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1517–1523.

Shridhar Singh. 2024. Enhancing privacy and security in large-language models: a zero-knowledge proof approach. In *International Conference on Cyber Warfare and Security*, pages 574–582. Academic Conferences International Limited.

Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2024. Identifying and mitigating privacy risks stemming from language models: A survey. *Preprint*, arXiv:2310.01424.

Chen Song, Ruwei Huang, and Sai Hu. 2024. Private-preserving language model inference based on secure multi-party computation. *Neurocomputing*, 592:127794.

Samuel Sousa and Roman Kern. 2023. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, 56(2):1427–1492.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Dimitri Staufer, Frank Pallas, and Bettina Berendt. 2024. Silencing the risk, not the whistle: A semi-automated text sanitization tool for mitigating the risk of whistleblower re-identification. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 733–745, New York, NY, USA. Association for Computing Machinery.

Jinglin Sun, Basem Suleiman, and Imdad Ullah. 2024. Effectiveness of privacy-preserving algorithms for large language models: A benchmark analysis. In *2024 21st Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–8.

Yidan Sun, Viktor Schlegel, Srinivasan Kolumam Nandakumar, Iqra Zahid, Yuping Wu, Warren Del-Pinto, Goran Nenadic, Lam Siew Kei, Jie Zhang, and Anil Bharath. 2025a. Evaluating differentially private generation of domain-specific text. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 5273–5278, New York, NY, USA. Association for Computing Machinery.

Yidan Sun, Viktor Schlegel, Srinivasan Nandakumar, Iqra Zahid, Yuping Wu, Yulong Wu, Hao Li, Jie Zhang, Warren Del-Pinto, Goran Nenadic, Siew Kei Lam, and Anil Anthony Bharath. 2025b. SynBench: A benchmark for differentially private text generation. *Preprint*, arXiv:2509.14594.

Yifan Sun and Weibin Guo. 2026. Dpmasker: A context-aware reversible differential privacy perturbation algorithm for medical text. In *2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, page 2374–2380.

Zhonghao Sun, Zhiliang Tian, Yiping Song, Yuyi Si, Juhua Zhang, Minlie Huang, Kai Lu, Zeyu Xiong, Xinwang Liu, and Dongsheng Li. 2025c. Dpga-textsyn: Differentially private genetic algorithm for synthetic text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 16159–16179, Vienna, Austria. Association for Computational Linguistics.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*.

Yakini Tchouka, Jean-François Couchot, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2024. Differentially private de-identifying textual medical document is compliant with challenging nlp analyses: Example of privacy-preserving icd-10 code association. *Intelligent Systems with Applications*, 23:200416.

Fengwei Tian, Payel Bhattacharjee, Heidi Hanson, Geoffrey D Rubin, Joseph Y. Lo, and Ravi Tandon. 2026. Stamp: Selective task-aware mechanism for text privacy. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1319–1333, Rabat, Morocco. Association for Computational Linguistics.

Zhiliang Tian, Yingxiu Zhao, Ziyue Huang, Yu-Xiang Wang, Nevin L Zhang, and He He. 2022. Seqpate: Differentially private text generation via knowledge distillation. *Advances in Neural Information Processing Systems*, 35:11117–11130.

Giovanni Paolo Tobia, Stefano Patarnello, Carlotta Masciocchi, Camilla Nero, Marco Carlo Passarotti, Giovanni Moretti, Antonio Marchetti, Giovanni Arcuri, and Livia Lilli. 2025. Privacy in italian clinical reports: A nlp-based anonymization approach. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, page 630–635. ISSN: 2575-2634.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. 2025a. On the vulnerability of text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), page 5150–5164, Albuquerque, New Mexico. Association for Computational Linguistics.

Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, and Zhikun Zhang. 2025b. InferDPT: Privacy-preserving inference for closed-box large language models. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4625–4640.

Tugtekin Turan, Dietrich Klakow, Emmanuel Vincent, and Denis Jouvet. 2022. Adapting language models when training on privacy-transformed data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4367–4373, Marseille, France. European Language Resources Association.

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

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Arpita Vats, Zhe Liu, Peng Su, Debjyoti Paul, Yingyi Ma, Yutong Pang, Zeeshan Ahmed, and Ozlem Kalinli. 2024. Recovering from privacy-preserving masking with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10771–10775.

Dinusha Vatsalan, Raghav Bhaskar, Aris Gkoulalas-Divanis, and Dimitrios Karapiperis. 2021. Privacy preserving text data encoding and topic modelling. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1308–1316.

Vishnu Vinod, Krishna Pillutla, and Abhradeep Guha Thakurta. 2025. InvisibleInk: High-utility and low-cost text generation with differential privacy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. 2024. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 507–527, Miami, Florida, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

WenHao Wang, Xiaoyu Liang, Rui Ye, Jingyi Chai, Siheng Chen, and Yanfeng Wang. 2024. KnowledgeSG: Privacy-preserving synthetic text generation with knowledge distillation from server. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7677–7695, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 305–314.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Isabella C. Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P. Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. 2025. Deidentifying medical documents with local, privacy-preserving large language models: The llm-anonymizer. *NEJM AI*, 2(4).

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K Meyer, Radhika Juglan, Zunamys I Carrero, Daniel Paech, Jens Kleesiek, Matthias P Ebert, and 1 others. 2024. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digital Medicine*, 7(1):257.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. 2024. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

Dominik Wunderlich, Daniel Bernau, Francesco Aldà, Javier Parra-Arnau, and Thorsten Strufe. 2022. On the privacy–utility trade-off in differentially private hierarchical text classification. *Applied Sciences*, 12(21):11177.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. 2024. Differentially private synthetic data

via foundation model APIs 2: text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Siqi Xie and Xiuxia Tian. 2023. Fused syntactic and semantic parsing for measuring privacy level in unstructured text. In *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using regularized mahalanobis metric. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 7–17, Online. Association for Computational Linguistics.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a utilitarian approach to privacy preserving text generation. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 11–20, Online. Association for Computational Linguistics.

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.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6137–6147, New York, NY, USA. Association for Computing Machinery.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Ying Yin and Ivan Habernal. 2022. Privacy-preserving models for legal natural language processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 172–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hong-Jun Yoon, Christopher Stanley, J Blair Christian, Hilda B Klasky, Andrew E Blanchard, Eric B Durbin, Xiao-Cheng Wu, Antoinette Stroup, Jennifer Doherty, Stephen M Schwartz, and 1 others. 2022. Optimal vocabulary selection approaches for privacy-preserving deep nlp model training for information extraction and cancer epidemiology. *Cancer Biomarkers*, 33(2):185–198.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, and 1 others. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Junwei Yu, Jieyu Zhou, Yepeng Ding, Lingfeng Zhang, Yuheng Guo, and Hiroyuki Sato. 2024. Textual differential privacy for context-aware reasoning with large language model. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 988–997.

Xiaojie Yu, Xiaojun Chen, and Jinqiao Shi. 2017. Vector based privacy-preserving document similarity with lsa. In *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, pages 1383–1387.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Sheng Yun, Zakirul Alam Bhuiyan, Md. Taufiq Al Hasib Sadi, and Shen Su. 2023. Privacy-preserving federated learning through clustered sampling on fine-tuning distributed non-iid large language models. In *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 531–538.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. 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 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Tymon Lesław Żarski and Artur Janicki. 2025. Enhancing privacy while preserving context in text transformations by large language models. *Information*, 16(1):49.

Agathe Zecevic, Xinyue Zhang, Sebastian Zeki, and Angus Roberts. 2024. Generation and evaluation of

synthetic endoscopy free-text reports with differential privacy. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 14–24, Bangkok, Thailand. Association for Computational Linguistics.

Ziqian Zeng, Jianwei Wang, Junyao Yang, Zhengdong Lu, Haoran Li, Huiping Zhuang, and Cen Chen. 2025. PrivacyRestore: Privacy-preserving inference in large language models via privacy removal and restoration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10821–10855, Vienna, Austria. Association for Computational Linguistics.

Huixin Zhan, Kun Zhang, Chenyi Hu, and Victor Sheng. 2021. Multi-objective privacy-preserving text representation learning. In *Proceedings of the 30th acm international conference on information & knowledge management*, pages 3612–3616.

Juhua Zhang, Zhiliang Tian, Minghang Zhu, Yiping Song, Taishu Sheng, Siyi Yang, Qiunan Du, Xinwang Liu, Minlie Huang, and Dongsheng Li. 2025. Dyntext: Semantic-aware dynamic text sanitization for privacy-preserving llm inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 20243–20255, Vienna, Austria. Association for Computational Linguistics.

Lexin Zhang, Changxiang Li, Qi Hu, Jingjing Lang, Sirui Huang, Linyue Hu, Jingwen Leng, Qiuhan Chen, and Chunli Lv. 2023. Enhancing privacy in large language model with homomorphic encryption and sparse attention. *Applied Sciences*.

Renhan Zhang. 2025. Privacy-oriented text generation in llms via selective fine-tuning and semantic attention masks. *Journal of Computer Technology and Software*, 4(8).

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Zihao Zhao and Anjalie Field. 2025a. Controlled generation for private synthetic text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32720–32735, Suzhou, China. Association for Computational Linguistics.

Zihao Zhao and Anjalie Field. 2025b. Controlled generation for private synthetic text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32720–32735, Suzhou, China. Association for Computational Linguistics.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. 2023. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5459–5473, Toronto, Canada. Association for Computational Linguistics.

Derui Zhu, Dingfan Chen, Xiongfei Wu, Jiahui Geng, Zhuo Li, Jens Grossklags, and Lei Ma. 2024a. PrivAuditor: Benchmarking data protection vulnerabilities in LLM adaptation techniques. *Advances in Neural Information Processing Systems*, 37:9668–9689.

Youxiang Zhu, Ning Gao, Xiaohui Liang, and Honggang Zhang. 2024b. Exploiting privacy preserving prompt techniques for online large language model usage. In *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, pages 4304–4309.

A Complete Literature Review Analysis Results

Category	Evaluation Method	#	Supporting Papers
Empirical privacy & attacks	Attack success rate (ASR)	28	(Macko et al., 2024; Tong et al., 2025b; Li et al., 2024b; Zeng et al., 2025; Vats et al., 2024; Meisenbacher et al., 2025d; Uchendu et al., 2023; Chen et al., 2024; Singh, 2024; Aditya et al., 2024; Feyisetan et al., 2019; Patsakis and Lykousas, 2023; Shao et al., 2024a; Chow et al., 2009; Du et al., 2023; Mai et al., 2024; Khadam et al., 2020; Gu, 2024; Fernandes et al., 2019; Khadam et al., 2021; Mimoto et al., 2020; Gardiner et al., 2024; Ben Cheikh Larbi et al., 2023; Ren et al., 2025; Meisenbacher et al., 2025b; Tong et al., 2025a; Kuo et al., 2025; Yang et al., 2025)
	(Carlini) exposure (rate)	12	(Sun et al., 2024; Li et al., 2024c,a; Shi et al., 2022a,b; Anil et al., 2022; Hoory et al., 2021; Smith et al., 2024; Wu et al., 2023; Sousa and Kern, 2023; Gu, 2024; Zhang, 2025)
	Plausible deniability: N_w (prob. word returned unperturbed)	8	(Meisenbacher et al., 2024a; Huang et al., 2024c; Feyisetan et al., 2019, 2020; Xu et al., 2020; Yue et al., 2021; Arnold et al., 2023a,b)
	Plausible deniability: S_w (support of perturbing a word)	6	(Meisenbacher et al., 2024a; Feyisetan et al., 2019, 2020; Xu et al., 2020; Yue et al., 2021; Arnold et al., 2023b)
	PII count / percentage	7	(Vinod et al., 2025; Mancera et al., 2025; Vats et al., 2024; Smith et al., 2024; Qu et al., 2021; Meisenbacher et al., 2026; Zhao and Field, 2025a)
	De-anonymization success / PII reconstruction rate	8	(Zarski and Janicki, 2025; Wang et al., 2024; Zhou et al., 2023; Zhao and Field, 2025b; Sun and Guo, 2026; Kalari et al., 2026; Pilán et al., 2025; Manzanares-Salor et al., 2026)
	Membership / attack advantage (TPR – FPR)	3	(Vu et al., 2024; Wunderlich et al., 2022; Smith et al., 2024)
Utility	Accuracy / Precision / Recall / F1	147	(Meisenbacher et al., 2024a; Liu et al., 2024; Salemi and Zamani, 2025; Yu et al., 2022; Sun et al., 2024; Zarski and Janicki, 2025; Li et al., 2024c; Higashi and Nakai, 2025; Wang et al., 2024; Li et al., 2024b; Mancera et al., 2025; De Faveri et al., 2024; Staufer et al., 2024; Loiseau et al., 2025; Plant et al., 2025; Kamoji et al., 2024; Uchendu et al., 2023; Chen et al., 2024; Singh, 2024; Zhang et al., 2023; Aditya et al., 2024; Resende et al., 2022; Hessel and Schofield, 2021; Smith et al., 2024; Hao et al., 2022; Feyisetan et al., 2019; Patsakis and Lykousas, 2023; Li et al., 2023b; Zhan et al., 2021; Senge et al., 2022; Yoon et al., 2022; Kassem et al., 2023; Bannihatti Kumar et al., 2023; Wunderlich et al., 2022; Shao et al., 2024a; Du et al., 2023; Shree and Kiran, 2020; Mishra et al., 2024; Prabhumoye et al., 2018; Carvalho et al., 2023; Tokpo and Calders, 2022; Zhou et al., 2023; Yu et al., 2024; Chen et al., 2022; Li et al., 2018; Ponomareva et al., 2022; Lai et al., 2022; Hassan et al., 2023; Yu et al., 2017; Beigi et al., 2019; Shi et al., 2022a; Tang et al., 2024; Wu et al., 2024; Duan et al., 2023; Xie et al., 2024; Mattern et al., 2022a; Qu et al., 2021; Anil et al., 2022; Plant et al., 2021; Canfora et al., 2018; Yue et al., 2021; Bo et al., 2021; Maheshwari et al., 2022; Sousa and Kern, 2023; Ye et al., 2024; Liu and Su, 2024; Huang et al., 2024a; Matzken et al., 2023; Gu, 2024; Ben Cheikh Larbi et al., 2023; Nagy et al., 2023; Yun et al., 2023; Wiest et al., 2024; Yin and Habernal, 2022; Pépin et al., 2024; Moghaddam et al., 2024; Song et al., 2024; Feyisetan and Kasiviswanathan, 2021; Badawi et al., 2020; Feyisetan et al., 2020; Adelani et al., 2020; Macko et al., 2024; Asimopoulos et al., 2024; Tchouka et al., 2024; Meisenbacher et al., 2024c; Meisenbacher and Matthes, 2024a; Zeng et al., 2025; Li et al., 2024a; Meisenbacher et al., 2025d; Raffetseder et al., 2025; Meisenbacher and Matthes, 2024b; Hathurusinghe et al., 2021; Utpala et al., 2023; Panov et al., 2022; Iwendi et al., 2020; Adelani et al., 2021; Yermilov et al., 2023; Alawad et al., 2021; Igamberdiev and Habernal, 2022; Bannour et al., 2022; Luo et al., 2024; Meehan et al., 2022; Weggenmann and Kerschbaum, 2018; Sánchez and Batet, 2017; Hoory et al., 2021; Meisenbacher et al., 2024b; Juez-Hernandez et al., 2023; Biesner et al., 2022; Sánchez and Batet, 2016; Igamberdiev and Habernal, 2023b; Weggenmann et al., 2022; Xie and Tian, 2023; Zecevic et al., 2024; McMurry et al., 2013; Gardiner et al., 2024; Mouhammad et al., 2023; Bao and Carpuat, 2024; Carranza et al., 2024; Sánchez et al., 2014; Hou et al., 2025a; Mishra et al., 2025; Meisenbacher et al., 2026; Isaoglu and Orman, 2026; Hu et al., 2026; Zhao and Field, 2025b; Wiest et al., 2025; Li et al., 2025; Sun et al., 2025c; Karousos et al., 2025; Ren et al., 2025; Frikha et al., 2025; Shahriar et al., 2025; Madaan et al., 2026; Papadopoulou et al., 2026; Psarra and Stefanidis, 2026; Meisenbacher et al., 2025b; Mancera et al., 2025; Shokri et al., 2025; Tobia et al., 2025; Ochs and Habernal, 2025; Shi et al., 2025; Yang et al., 2025; Tian et al., 2026; Meisenbacher et al., 2025c; Manzanares-Salor et al., 2026; Huang et al., 2025b; Abbasalizadeh and Narain, 2025)
	AUC-ROC	15	(Macko et al., 2024; Li et al., 2024a; Smith et al., 2024; Li et al., 2024b, 2023b; Wunderlich et al., 2022; Feyisetan et al., 2020; Nagy et al., 2023; Feyisetan and Kasiviswanathan, 2021; Carvalho et al., 2023; Tang et al., 2024; Duan et al., 2023, 2024; Mishra et al., 2025; Isaoglu and Orman, 2026)
	Exact match (task EM)	9	(Zeng et al., 2025; Khoje, 2024; Alabdulkareem et al., 2024; Ponomareva et al., 2022; Smith et al., 2024; Hou et al., 2025a; Kuo et al., 2025; Shi et al., 2025; Tian et al., 2026)
	True Positive Rate (TPR)	6	(Higashi and Nakai, 2025; Li et al., 2024a; Smith et al., 2024; Li et al., 2023b; Maheshwari et al., 2022; Li et al., 2024b)
	False Positive Rate (FPR)	7	(Higashi and Nakai, 2025; Li et al., 2024a; Resende et al., 2022; Smith et al., 2024; Li et al., 2023b; Zhao and Field, 2025b; Wiest et al., 2025)
	MCC (Matthews Correlation Coefficient)	5	(Hessel and Schofield, 2021; Song et al., 2024; Luo et al., 2024; Mattern et al., 2022b; Matzken et al., 2023)
	MRR (Mean Reciprocal Rank)	4	(Li et al., 2024c; Bao and Carpuat, 2024; Wu et al., 2023; Feyisetan et al., 2020)
	Confusion matrices	5	(Senge et al., 2022; Wiest et al., 2024; Pépin et al., 2024; Mishra et al., 2025; Karousos et al., 2025)
	Utility preservation (%) from information content (IC)	3	(Hassan et al., 2023; Sánchez and Batet, 2016; Sánchez et al., 2014)
Text quality & semantic preservation	Perplexity / Generation perplexity gap	37	(Awon et al., 2025; Flemings and Annavaram, 2024; Flemings et al., 2024; Vinod et al., 2025; Higashi and Nakai, 2025; Vats et al., 2024; Meisenbacher et al., 2025d; Meisenbacher and Matthes, 2024b; Turan et al., 2022; Li et al., 2023a; Wu et al., 2023; Kerrigan et al., 2020; Aditya et al., 2024; Liu and Kalinli, 2024; Smith et al., 2024; Kassem et al., 2023; Tokpo and Calders, 2022; Mattern et al., 2022b; Matzken et al., 2023; Lai et al., 2022; Shi et al., 2022a; Majmudar et al., 2022; Ginart et al., 2022; Tian et al., 2022; Weggenmann et al., 2022; Li et al., 2024b; Meisenbacher et al., 2026; Isaoglu and Orman, 2026; Zhao and Field, 2025a; Madaan et al., 2026; Meisenbacher et al., 2025b; Ochs and Habernal, 2025; Kuo et al., 2025; Huang et al., 2025b)
	Cosine similarity (embedding-based)	28	(Awon et al., 2025; Meisenbacher et al., 2024c; Zhu et al., 2024b; Tong et al., 2025b; De Faveri et al., 2024; Staufer et al., 2024; Meisenbacher et al., 2025d; Meisenbacher and Matthes, 2024b; Uchendu et al., 2023; Tokpo and Calders, 2022; Mattern et al., 2022b,a; Sasada et al., 2021; Weggenmann et al., 2022; Bo et al., 2021; Meisenbacher and Matthes, 2024a; Jiang et al., 2009; Gu, 2024; Meisenbacher et al., 2026; Zhao and Field, 2025b; De Faveri et al., 2025b; Sun et al., 2025a; Meisenbacher et al., 2025a,b; Kim, 2025; Tian et al., 2026; Meisenbacher et al., 2025c; Pilán et al., 2025)
	ROUGE-1 / ROUGE-2 / ROUGE-L	25	(Salemi and Zamani, 2025; Yu et al., 2022; Hou et al., 2025b; Liang et al., 2024; Huang et al., 2025a; Zeng et al., 2025; Huang et al., 2024c; Staufer et al., 2024; Li et al., 2023a; Smith et al., 2024; Yermilov et al., 2023; Zhou et al., 2023; Yu et al., 2024; Zecevic et al., 2024; Gu, 2024; Gardiner et al., 2024; Wu et al., 2024; Hou et al., 2025a; Zhao and Field, 2025a; Li et al., 2025; Sun and Guo, 2026; Frikha et al., 2025; Madaan et al., 2026; Meisenbacher et al., 2025b; Huang et al., 2025b)
	BLEU	21	(Yu et al., 2022; Higashi and Nakai, 2025; Huang et al., 2024c; Meisenbacher et al., 2025d; Carranza et al., 2024; Meisenbacher and Matthes, 2024b; Smith et al., 2024; Feng et al., 2020; Prabhumoye et al., 2018; Igamberdiev and Habernal, 2023b; Igamberdiev et al., 2024; Weggenmann et al., 2022; Ye et al., 2024; Gu, 2024; Hou et al., 2025a; Li et al., 2025; Sun and Guo, 2026; Sun et al., 2025a; Shao et al., 2026; Huang et al., 2025b)
	MAUVE	12	(Tong et al., 2025b; Vinod et al., 2025; Mattern et al., 2022a; Wang et al., 2024; Carranza et al., 2024; Xie et al., 2024; Uchendu et al., 2023; Isaoglu and Orman, 2026; Hu et al., 2026; Zhao and Field, 2025a; Sun et al., 2025c; Zhang et al., 2025)
	METEOR	7	(Yu et al., 2022; Higashi and Nakai, 2025; Fisher et al., 2024; Liang et al., 2024; Uchendu et al., 2023; Adelani et al., 2021; Sun et al., 2025a)
	BERTScore / BARTScore	7	(Liang et al., 2024; Huang et al., 2024c; Utpala et al., 2023; Isaoglu and Orman, 2026; Sun et al., 2025a; Kim, 2025; Shi et al., 2025)
	Word Error Rate (WER)	3	(Vats et al., 2024; Turan et al., 2022; Liu and Kalinli, 2024)
Regression / error & correlation	Pearson / Spearman ρ / Correlation	8	(Hessel and Schofield, 2021; Feyisetan et al., 2019; Li et al., 2023b; Song et al., 2024; Luo et al., 2024; Mishra et al., 2024; Chen et al., 2022; Matzken et al., 2023)
	MAE (Mean Absolute Error)	3	(Salemi and Zamani, 2025; Liu and Su, 2024; Gu, 2024)
	Signal to noise ratio (PSNR)	3	(Yin and Habernal, 2022; Khadam et al., 2021, 2020)
System / efficiency	Runtime (ms) / Runtime volume	13	(Resende et al., 2022; Hao et al., 2022; Senge et al., 2022; Vatsalan et al., 2021; Yin and Habernal, 2022; Sánchez and Batet, 2017; Yu et al., 2017; Morabito et al., 2023; Liu and Su, 2024; Huang et al., 2024a; Gu, 2024; Sánchez et al., 2014; Chakaravarthy et al., 2008)
	Communication volume / overhead / message sizes	7	(Hao et al., 2022; Liang et al., 2024; Luo et al., 2024; Feng et al., 2020; Liu and Su, 2024; Huang et al., 2024a; Zhao and Field, 2025b)
	Speed	6	(Liang et al., 2024; Song et al., 2024; Badawi et al., 2020; Liu and Su, 2024; Huang et al., 2024a; Karamitsos et al., 2025)
	Total inference time / Per-operation time	4	(Liang et al., 2024; Song et al., 2024; Luo et al., 2024; Gu, 2024)
	(Comp. / Comm. / Encrypt./Decrypt.) Latency	3	(Zhang et al., 2023; Badawi et al., 2020; Huang et al., 2024a)
	Memory/disk footprint / RAM usage (engineering metrics)	3	(Bannihatti Kumar et al., 2023; Vatsalan et al., 2021; Moghaddam et al., 2024)
Others	LLM score / LLM-Judge (undefined rubric varies across papers)	9	(De Faveri et al., 2025a; Awon et al., 2025; Wang et al., 2024; Huang et al., 2025a; Zeng et al., 2025; Matzken et al., 2023; Frikha et al., 2025; Kuo et al., 2025; Huang et al., 2025b)
	Relative gain	8	(Meisenbacher et al., 2024c; Meisenbacher and Matthes, 2024b; Weggenmann and Kerschbaum, 2018; Mattern et al., 2022b; Meisenbacher et al., 2024b, 2026, 2025a,b)

Table 3: A summary of the literature review analysis, grouped by (categories of) text privatization evaluation metrics.

B Detailed System Architecture of PrivBench

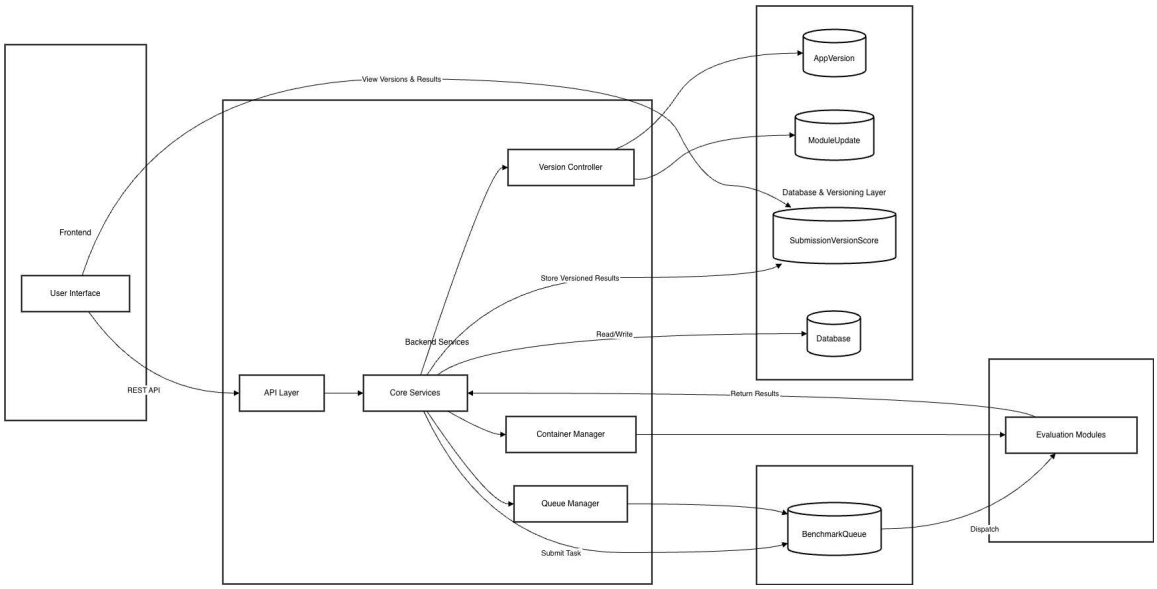

Figure 4: Component-level architecture of PrivBench.

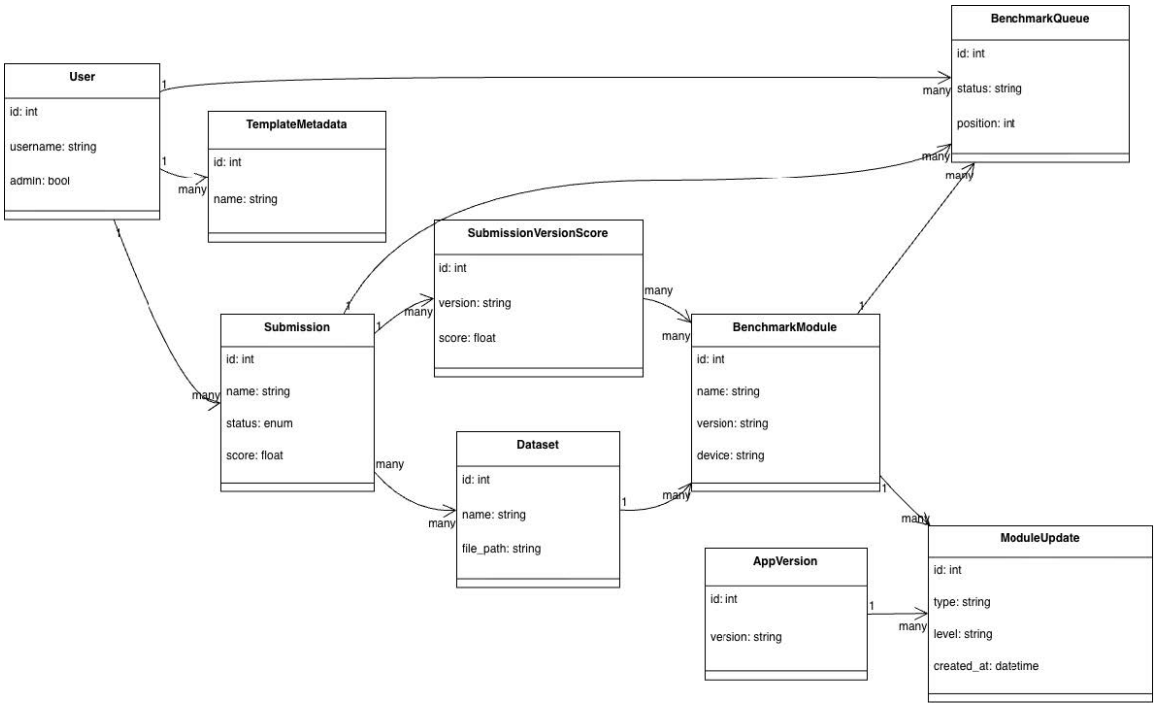

Figure 5: Simplified UML class diagram of the PrivBench web application.