How do we measure privacy in text? A survey of text anonymization metrics

Anonymous ACL submission

Abstract

In this work, we aim to clarify and reconcile metrics for evaluating privacy protection in text through a systematic survey. Although text anonymization is essential for enabling NLP research and model development in domains with sensitive data, evaluating whether anonymization methods sufficiently protect privacy remains an open challenge. In manually reviewing 43 papers that report privacy metrics, we identify and compare six distinct privacy notions and analyze how the associated metrics capture different aspects of privacy risk. We then assess how well these notions align with legal privacy standards (HIPAA and GDPR), as well as user-centered expectations grounded in HCI studies. Our analysis offers practical guidance on navigating the landscape of privacy evaluation approaches further and highlights gaps in current practices. Ultimately, we aim to facilitate more robust, comparable, and legally aware privacy evaluations in text anonymization.

1 Introduction

Text anonymization—through methods such as redaction, rewriting, or data synthesis—has become a critical tool for mitigating the risks of sharing or training models on sensitive data (Lison et al., 2021). When done effectively, anonymization can enable access to valuable resources like clinical records, legal texts, or social media content without endangering individual privacy. However, text anonymization is inherently difficult: high-dimensional data in general is vulnerable to re-identification (Narayanan and Shmatikov, 2008), and even mechanisms that offer formal guarantees can still fail against practical deanonymization attacks (Mattern et al., 2022; Tong et al., 2025; Pang et al., 2025). As LLMs further heighten concerns about memorization of training data (Carlini et al., 2021) and re-identification of sensitive attributes (Staab et al., 2024), rigorous privacy evaluation has become a fundamental requirement for responsible data sharing and model deployment.

Despite the importance of evaluation, measuring the effectiveness of text anonymization systems remains an open challenge. Current evaluations span a wide range of tasks and assumptions, reflecting divergent notions of privacy. Papers focusing on redacting direct identifiers, such as (Hassan et al., 2019; Lison et al., 2021; Pilán et al., 2022), implicitly use different notions of privacy than papers focusing on synthesizing text (Meisenbacher et al., 2024a; Yue et al., 2023; Wang et al., 2023), and even papers targeting the same notion of privacy use different metrics. In many cases, metrics are poorly connected to legal and social notions of privacy, leaving researchers and practitioners with limited guidance on what risks a given evaluation actually measures, or what constitutes sufficient protection in practice.

In this work, we aim to enhance understanding of privacy evaluation in text by conducting a systematic survey of metrics. Through keyword searching and citation links, we identify 43 papers published since 2019 that report metrics for measuring privacy in text, and we manually categorize metrics into high-level notions of privacy. Unlike existing surveys of anonymization techniques (Pawar et al., 2018; Mahendran et al., 2021) or general privacy principles (Wagner and Eckhoff, 2019), we specifically target quantified metrics. While we focus primarily on metrics for evaluating privacy in text that has been anonymized (e.g., through redaction, rewriting, or synthesis), our analysis also has relevance to privacy in models trained on text.

Our analysis reveals six privacy notions underlying specific metrics—identifier removal effectiveness, dataset membership, attribute inference risk, reconstruction attacks, semantic inference risk, and theoretical privacy bounds, which we discuss in more depth in §2.2. We further discuss how these notions map to legal privacy standards, specifically

HIPAA and GDPR (§3), and social expectations derived from user-centered research (§4). Finally, we conclude by discussing open challenges in privacy evaluation and opportunities for future work.

Overall, we aim to aid researchers and practitioners in understanding the landscape of privacy evaluations for text, offering a structured view of what existing metrics capture and what they overlook. By organizing metrics by privacy objective and examining their assumptions, we offer practical guidance for metric selection and reveal gaps in legal alignment, social relevance, and evaluation consistency. We ultimately aim to improve consistency in evaluation and encourage future work on the development of new standardized metrics.

2 Existing privacy metrics in NLP

2.1 Scope and Methodology

We define the scope of this survey as metrics for evaluating privacy in anonymized text outputs—i.e., evaluations that directly assess how much sensitive information remains exposed in the text itself after anonymization. Our goal is to characterize the landscape of privacy evaluations applied to generated or modified text, focusing on settings where both original and anonymized versions are typically available.

We include papers published from 2019 onward that explicitly report one or more quantitative metrics for evaluating privacy in anonymized or synthetic text. This time window ensures our review focuses on recent methods relevant to current NLP pipelines and data-sharing concerns. We identified papers using a combination of keyword-based search and backward citation tracking. Specifically, we searched ACL Anthology and Google Scholar using combinations of the terms "`text anonymization`", "`text sanitization`", and "`synthetic text generation`".

To be included, a paper must satisfy all of the following:

- Focus on natural language text (not images, structured tabular data, or speech);
- Contain anonymized or privatized text outputs (not just internal model embeddings or representations);
- Report at least one privacy evaluation metric (beyond utility, fluency, or readability);
- Be peer-reviewed or publicly available as a preprint since January 2019.

2.2 Survey of Evaluation Metrics

We identify six high-level privacy objectives that the surveyed papers aim to evaluate: identifier removal, dataset membership, attribute inference, reconstruction attacks, semantic inference, and theoretical bounds. These objectives are summarized in Figure 1. Each objective reflects a different aspect of privacy risk and corresponds to distinct families of evaluation metrics. Some papers target a single objective, while others report metrics spanning multiple categories. In the subsections that follow, we describe each objective in turn, outline the types of metrics used to evaluate it, and discuss how these metrics are applied in practice. A detailed mapping of papers to privacy objectives and associated evaluation metrics is provided as a spreadsheet in the supplementary materials.

2.2.1 Identifier Removal Effectiveness

Identifier removal effectiveness asks whether anonymized text still exposes directly identifying information, such as names, addresses, or contact details, that should have been masked during anonymization. The most common evaluation approach compares detected spans against gold standard annotations using token-level or span-level precision, recall, and F_1 scores. These metrics dominate because they clearly reflect two types of failure: masking too little (false negatives) and masking too much (false positives) (Lison et al., 2021).

Metrics vary in how they evaluate disclosure. For example, entity-level recall treats an identifier as successfully protected only if all its mentions are masked across a document or corpus, while other metrics distinguish between identifiers that uniquely point to individuals (e.g., names) and those that may only do so in combination (e.g., age or ZIP code) (Pilán et al., 2022). To handle cases where the predicted masked span does not exactly match the annotated boundary—either by masking too little or too much—tagging schemes (IOB-Exact or IOB-Partial) have been used to address partial masking and boundary mismatch (Hassan et al., 2019; Manzanares-Salor et al., 2022).

Approximate-match metrics credit redactions that are incomplete but still effective—such as changing "John Smith" to "Jonathan" or using paraphrases. Edit-distance-based scores (e.g., Leven-

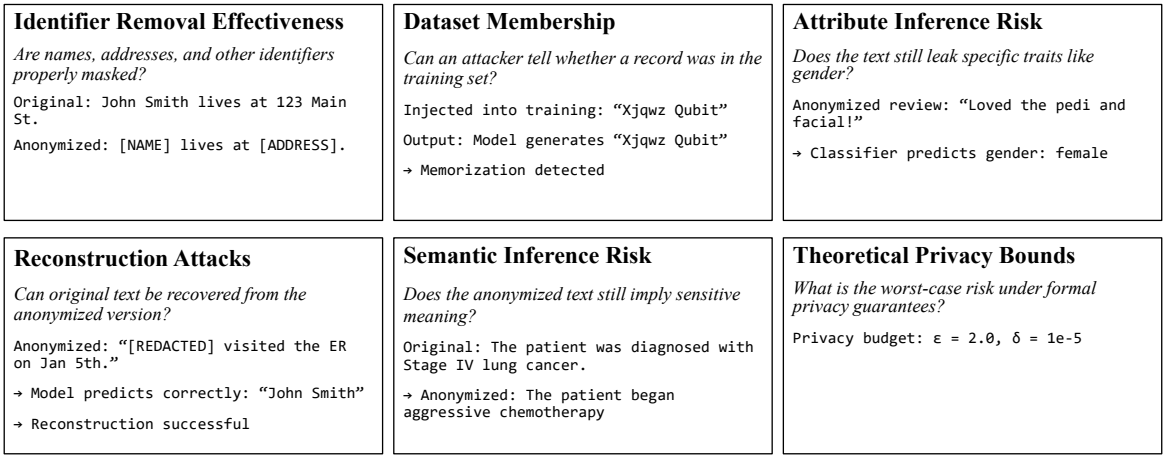

Figure 1: Overview of six privacy objectives used in text privacy evaluation. Each panel summarizes the privacy notion and provides an illustrative example.

shtein Recall) and token-level lexical divergence quantifies how different the anonymized span is from the original (Alves et al., 2024; Xin et al., 2025). At a higher semantic level, metrics like PRIVACY_NLI asks whether the anonymized sentence still implies the original using textual entailment models, while SPRIVACY reports human judgments of whether personal information remains (Huang et al., 2024).

Together, these metrics form a progression from surface-level removal to deeper notions of semantic obfuscation. While span-level F_1 remains the most common metric, newer work shows that semantic or corpus-level assessments may better capture residual privacy risks, especially in cases when anonymization involves rewriting rather than redaction. Importantly, identifier-only metrics may underestimate leakage: successfully masking names does not guarantee protection if the text still allows an individual to be inferred through other cues, which motivates more robust evaluation frameworks under different notions of privacy.

2.2.2 Dataset Membership

Dataset membership metrics assess whether an adversary can determine if a specific record was part of the data used to train or generate an anonymized output. This metric is strongly tied to the concept of privacy in models and is most commonly used to assess synthetic data generation approaches, as most redaction-based approaches would trivially fail this test. When the notion of a "record" is well defined—as in the entire clinical record for a patient—successfully hiding membership may offer broad protections (Salem et al., 2023).

Standard evaluations use shadow or reference models to estimate membership inference accuracy, F_1, or AUC (Arnold et al., 2023; El Kababji et al., 2023). Variants include confidence-threshold and entropy-threshold attacks on privatized embeddings, with success rate indicating leakage (Du et al., 2023). However, membership inference attacks remain fragile and context-sensitive: their performance is highly influenced by attack design, dataset construction, and the nature of the reference data (Naseh and Mireshghallah, 2025; Duan et al., 2024).

While membership inference traditionally focuses on entire data points (e.g., if a full document was included in a dataset used to train a synthetic data generator), work focused on synthetic text generation specifically has also evaluated privacy through canary-injection experiments (Carlini et al., 2019; Yue et al., 2023; Ramesh et al., 2024), which assess memorization and leakage through a partial notion of membership. In canary experiments, unique phrases are inserted into the data used to train synthetic text generators. The leakage rate and perplexity rank of these "canaries" serve as indicators of the risk of data leakage, e.g., a low perplexity indicates a canary was likely a member of the training data (Carlini et al., 2019).

Dataset membership testing is common in evaluations of generation methods that claim differential privacy guarantees (Arnold et al., 2023; Du et al., 2023; Yue et al., 2023). These studies often pair em-

pirical leakage measures with formal (ε, δ) budgets to assess whether theoretical protections translate into practical robustness.

2.2.3 Attribute Inference Risk

Attribute inference metrics evaluate whether sanitized text still reveals sensitive traits, that is, can a reader infer gender, age, or diagnosis more accurately than chance? These metrics are often used to evaluate rewriting-based methods (Meisenbacher et al., 2024b; Meisenbacher and Matthes, 2024) and synthetic text generation approaches (Wang et al., 2023; Wang and Sun, 2022). Most published attacks pursue attributes of the text author: reviewer gender or age in Trustpilot, political leaning in tweets, stylistic cues in blog posts (Meisenbacher et al., 2024b; Chim et al., 2025). A smaller but growing line of work targets attributes of people mentioned within the text, for example patient sex or comorbidities in synthetic EHRs generated by DP-RVAE or PromptEHR (Wang et al., 2023; Wang and Sun, 2022).

A common approach is to train classifiers on both original and anonymized text and compare their ability to predict protected traits. A drop in accuracy or F_1 is interpreted as evidence of improved privacy. For example, Meisenbacher et al. (2024b) use Privacy F_1 under static and adaptive attackers to quantify how retraining DP-MLM rewrites limits attribute leakage. Multi-attribute settings extend this to keyword-inference accuracy, Gender-F_1, and Age-F_1 in synthetic EHRs (Wang et al., 2023; Meisenbacher and Matthes, 2024).

Beyond individual attribute prediction, El Kababji et al. (2023) model sequential attacks in which an adversary first links synthetic clinical trial records to real patients and then predicts sensitive attributes such as tumor grade. These approaches capture different facets of attribute leakage and can be applied to both token and embedding-level representations.

These metrics are valuable for quantifying residual leakage that might persist even after identifiers are removed. However, a drop in inference accuracy does not guarantee that private attributes are fully protected (Du et al., 2023; Chim et al., 2025). Attribute inference thus plays a complementary role in privacy evaluation: it highlights forms of privacy leakage that are not captured by identifier masking alone, but does not ensure broader protection on its own against reconstruction or membership disclosure.

2.2.4 Reconstruction Attacks

Reconstruction attacks pose a different question: after anonymization, can an adversary re-create verbatim or near-verbatim portions of the original document, and thus link them back either to the author or to the individuals mentioned? Even with names removed, rare phrases or consistent style can suffice for re-identification.

The most widely reported reconstruction metrics operate at the document level. Retrieval-based metrics (e.g., BM25, Jaccard, or ensemble linking) count how often an anonymized text's original counterpart is retrieved from the candidate pool, exposing residual uniqueness in wording or topic (Xin et al., 2025; Ben Cheikh Larbi et al., 2023; Morris et al., 2022). In clinical domains, manual re-identification studies simulate the process by which humans might trace rewritten notes back to the patients described (Casula et al., 2024).

More automated approaches for re-identification include bounding worst-case leakage rates across tokens (Tong et al., 2025), and estimating how easily masked tokens are guessed by models like BERT (Chen et al., 2023). Other metrics highlight unique or memorized content: span surprisal (Papadopoulou et al., 2022), plausible-deniability set size (Yue et al., 2021), ROUGE overlap (Zecevic et al., 2024), and rare-token counts (Meisenbacher et al., 2024c).

Together, these metrics range from coarse retrieval to fine-grained content recovery. Choosing among them depends on whether the primary concern is full-document retrieval or recovery of sensitive snippets, and whether the at-risk party is the author of the text, the individual described, or both.

2.2.5 Semantic Inference Risk

While reconstruction metrics focus on verbatim overlap, semantic inference metrics ask a broader question: does the anonymized text still convey the same meaning as the original? If so, an adversary may infer sensitive information, even in the absence of explicit identifiers. Metrics focused on semantic inference differ from those focused on attribute inference in that they primarily assess similarity between original and anonymized text, and do not necessarily target the prediction of specific personal traits. Instead, they flag risks when anonymized content retains enough topical, relational, or narrative structure to support inference.

Most evaluations begin with embedding-based similarity. SBERT cosine scores are commonly

used to quantify alignment between original and anonymized text (Meisenbacher et al., 2024a). To move beyond raw cosine scores, Xin et al. (2025) introduce two refinements: a lexical divergence score, which filters out superficial rewording, and a semantic alignment score, which uses language model prompts to judge factual consistency. Both metrics help identify cases where surface anonymization fails to hide deeper meaning, especially in clinical contexts where sensitive events remain recognizable.

When dense encoders are unavailable, simpler lexical metrics provide a coarse but practical alternative. Igamberdiev and Habernal (2023) reinterpret corpus-level BLEU as a privacy indicator: large n-gram overlap implies that substantial original wording and therefore potential leakage remains. Meisenbacher et al. (2024c) report the Perturbation Percentage (PP), the fraction of tokens altered during anonymization, and show low PP often aligns with successful author-attribute inference.

These metrics vary in granularity and precision, but all reflect a central tradeoff: preserving utility often means preserving meaning, which may leave privacy at risk. For applications where downstream utility is paramount, practitioners may tolerate relatively high similarity scores, whereas in high-sensitivity domains, such as clinical text, even subtle semantic similarities can pose privacy risks.

2.2.6 Theoretical Privacy Bounds

In contrast to the more empirical measurements of privacy described in the preceding sections, differentially private methods provide strict mathematical guarantees that set an upper bound on the maximum permissible privacy leakage. The tightness of this bound depends on the user-specified parameters that define the privacy budget, and it can be reported alongside other empirical measures. As such, we regard these theoretical bounds as distinct metrics that directly quantify the extent of privacy protection.

Differentially private methods generally involve the addition of noise to model representations or gradient updates to reduce the risk of membership inference. The level of noise added is carefully calibrated to adhere to the specified privacy budget, typically formalized by (ε, δ) differentially private guarantees. Both text synthesis and text rewriting approaches have incorporated DP guarantees, including SANTEXT, DP-BART, DP-MLM, and DP-RVAE, and papers typically report results under varying levels of ε (Yue et al., 2021; Igamberdiev and Habernal, 2023; Meisenbacher et al., 2024b; Wang et al., 2023; Du et al., 2023).

In addition to global ε values, some studies analyze privacy at the token level to better understand the behavior of specific mechanisms. The self-substitution rate N_w measures the probability that a token survives the mechanism unchanged, whereas the support size S_w counts how many distinct outputs the mechanism may emit for that token (Meisenbacher et al., 2024a,c; Arnold et al., 2023). These two metrics together characterize the output entropy of the substitution process: when tokens are frequently altered and drawn from a large set of alternatives, an adversary faces greater uncertainty about the original content.

Importantly, theoretical guarantees do not replace empirical testing. They only hold if methods are correctly implemented and the data satisfies necessary assumptions. Furthermore, while theoretical metrics can precisely describe which setup offers better protection, they typically lack human interpretability, e.g. under differential privacy, $\varepsilon = 4$ implies better protection than $\varepsilon = 8$, but it is not clear what either metric actually means for leakage risks nor which value should be used. Recent studies report DP parameters alongside reconstruction or membership attack results, enabling readers to verify whether the empirical results respect the advertised guarantees (Meisenbacher et al., 2024a; Chen et al., 2023; Wang et al., 2023; Meisenbacher et al., 2024b; Zecevic et al., 2024; Arnold et al., 2023; Du et al., 2023; Yue et al., 2023; Wang and Sun, 2022).

3 Are current metrics sufficient to meet legal standards?

Modern privacy regulations articulate rigorous requirements for anonymization that are not always reflected in current technical evaluations. In this section, we assess whether commonly used evaluation metrics in text anonymization align with the legal definitions, using the two most influential frameworks as case studies: the U.S. HIPAA Privacy Rule and the EU General Data Protection Regulation (GDPR). Drawing from the survey in §2.2, we analyze where current practices fall short, and what improvements are necessary for legal defensibility.

3.1 HIPAA: Emphasis on Identifier Removal and Expert Judgment

The HIPAA Privacy Rule defines two standards for de-identification of data: (1) Safe Harbor, which mandates removal of 18 enumerated identifiers; and (2) Expert Determination, in which a statistical expert attests that the risk of re-identification is "very small" given anticipated use (Office for Civil Rights (OCR), 2012).

Identifier Removal metrics (§2.2.1) align well with Safe Harbor. These metrics appeared in 15 of the reviewed papers, and directly measure how effectively models detect and mask identifiable tokens.

However, current evaluation datasets are rarely annotated according to HIPAA standards. Annotation of generic named entity types misses more domain-specific identifiers, especially since HIPAA's list includes quasi-identifiers like geographic information and dates. Entity-level recall metrics (Pilán et al., 2022) better quantify HIPAA compliance than span-level metrics by requiring consistent masking across contexts, but few evaluations use them.

The Expert Determination pathway implies the need for holistic risk modeling—evaluations that simulate adversarial re-identification or analyze residual inference risks. While attribute inference, reconstruction attacks, and semantic inference risk have the potential to mimic expert determinations, only a few studies attempt such modeling, and very few studies investigate how attack models compare to real experts. Exceptions include human-in-the-loop evaluations, such as the TILD framework (Mozes and Kleinberg, 2021), which uses "motivated intruder" tests to assess whether humans can re-identify entities given background knowledge.

While current evaluation metrics cover some aspects of HIPAA, especially Safe Harbor, they fall short of the broader requirements implied by Expert Determination, which demand more comprehensive and adversary-aware assessments.

3.2 GDPR: Contextual Risk and Semantic Inference

GDPR requires that anonymized data be such that individuals are "not identifiable by any means reasonably likely to be used" by an adversary (European Parliament and Council, 2016). This contextual standard evaluates identifiability not just by direct identifiers but also by semantic clues, auxiliary data, and task-specific inference.

Reconstruction metrics (§2.2.4) simulate adversarial behavior and are among the most legally aligned with GDPR. However, most studies adopt a single fixed attacker and rarely vary the knowledge base or background assumptions, limiting their robustness as legal evidence.

Attribute inference metrics (§2.2.3) also relate directly to GDPR concerns, as they measure the extent to which sensitive traits can be recovered from anonymized text. Yet few evaluations test multiple attributes.

Metrics from the Semantic Inference Risk (§2.2.5) category indirectly assess the residual information in the text. High semantic similarity may indicate exposure of sensitive attributes or events. Yet these proxies do not directly evaluate whether an attacker could infer private information, as required under GDPR.

GDPR compliance requires adversarial thinking and evaluation of contextual identifiability. Most current metrics fall short on this front: Identifier Removal metrics overlook quasi-identifiers and risks of re-identification; Reconstruction metrics are rarely diversified across attack strategies; and semantic similarity scores do not map cleanly onto real-world inference risks. Broader adoption of human-intruder studies and diverse reconstruction attacks and attribute inference probes are needed to bridge this gap.

4 User-Centered Privacy and Contextual Integrity

While technical metrics dominate text anonymization research, they often overlook a central question: to what extent do these metrics reflect what people actually care about in privacy? Human-centered literature on HCI and social computing suggests that users' privacy perceptions depend on more than whether names or attributes are masked. Users' privacy expectations are shaped by the information context, agency, and perceived coherence of privatized text. This section explores key themes from user-centered privacy research, identifying gaps between current evaluation practices and the lived concerns of users.

The theory of Contextual Integrity, introduced by Nissenbaum (2004), suggests that privacy is not about secrecy or control in the abstract, but about appropriate flows of information: who sends what to whom, under what conditions, and for what

purpose. In practice, whether a particular data sharing is acceptable depends on if it aligns with the norms in the associated context. For example, acceptance of COVID-19 contact tracing applications and vaccination-certification systems depends on whether the information flows are bounded by expectations about recipients, use purpose, and retention time, all of which go beyond simply removal of identifiers or risks of re-identification (Feng et al., 2024; Zhang et al., 2022). Through a user study with 721 participants, Meisenbacher et al. (2025) show that users care about data sensitivity, mechanism type, and reason for data collection in the specific context of differentially private text, as suggested by contextual integrity theory more generally.

While NLP systems and evaluation practice have minimally drawn from contextual integrity theory, HCI studies have leveraged it by treating privacy as alignment between users' disclosure preferences and the contextual demands, rather than as fixed rules or outputs. Several systems aim to support users in managing what they share, rather than deciding for them. For instance, Rescriber lets users rewrite or hide sensitive parts of their messages to language models, based on what the user feels is appropriate in the moment (Zhou et al., 2025). Other tools like CLEAR and Contextual Privacy Policies adapt the way data is handled depending on factors like location, app behavior, or who the recipient is (Chen et al., 2025; Pan et al., 2024).

Evaluations of privacy in text could similarly integrate context. Currently, metrics focus narrowly on identifier recall, leakage, or attack success, without assessing whether the anonymized text reflects an information flow that is appropriate for the context, whether users feel in control of disclosures, or whether the outputs align with their privacy expectations. As a result, systems may score well on standard benchmarks yet still fail to earn user trust or meet real-world standards of privacy acceptability. Context-sensitive metrics could entail, for example, explicitly defining the scenario where each metric is appropriate. Future work could also develop new metrics that take context or user preferences as input variables that influence the type of assessment.

5 Discussion

While our survey focuses on evaluating privacy in text itself, a related line of research concerns the privacy risks of models trained on text, with recent work focusing on large language models (LLMs). We briefly highlight how our survey can inform research in this setting as well, and generally suggests that better reconciling text and model privacy can advance both areas.

Model privacy literature typically investigates whether trained models can memorize, reveal, or allow inference about sensitive training data (Neel and Chang, 2024). Although the evaluation target differs from text anonymization, the two areas share some similar privacy notions, such as membership inference and reconstruction attacks. Specific metrics for model privacy overlap with metrics used to evaluate privacy in synthetic text, including canary attacks and success rate of membership inference attacks, where evaluation often targets the synthetic text generato, not just the output text. In particular, membership inference attacks (MIAs) have been widely studied in both black-box and white-box settings (Carlini et al., 2022; Shokri et al., 2017), with recent work adapting them to few-shot and in-context learning (Wen et al., 2024; Jiménez-López et al., 2025).

Beyond leakage of training data, Staab et al. (2024) demonstrate an additional model privacy risk in LLMs specifically: that they can infer sensitive traits through attribute inference attacks. This risk is quantified using metrics like classifier accuracy and profiling success, which also appear in anonymization work (Frikha et al., 2025). Although a privacy risk, the potential for LLMs to be powerful de-anonymizers also offers an opportunity for empirical evaluation: LLMs may serve as strong adversaries in empirically conducting reconstruction attacks, attribute inference risks, and semantic inference risks.

Model privacy literature includes several standardized benchmarks. Mireshghallah et al. (2024) apply theories of contextual integrity to evaluate privacy in terms of normative expectations, echoing similar calls in user-centered anonymization metrics. PrivLM-Bench evaluates privacy risks such as PII exposure and attribute inference across standardized tasks (Li et al., 2024). Probing tools like ProPILE and targeted black-box attacks offer practical approaches to assess leakage without requiring internal model access (Kim et al., 2023; Abascal et al., 2024). These methods highlight how information can leak through paraphrases or semantic proxies, a challenge also present in text anonymization. As privacy risks in NLP span both model

behavior and textual output, bridging the two literatures could support more robust and transparent evaluation frameworks. These would incorporate attacker simulations, contextual analysis, and metrics grounded in real-world privacy concerns.

6 Recommendations and Open Challenges

Our survey reveals several gaps in current privacy evaluation practices for text anonymization and highlights opportunities for future work. In this section, we synthesize key takeaways into actionable recommendations and outline open research directions for building more robust and comparable evaluation frameworks.

Align metrics with stated goals. Privacy metrics should reflect the intended privacy guarantees of a method. For example, approaches designed to minimize re-identification risk should not be evaluated solely with identifier-removal F1 scores, which ignore indirect leakage. Similarly, methods that aim to reduce semantic inference should adopt task-specific probes or classifier-based evaluations, not just surface similarity metrics. Articulating intended use cases and mapping them to appropriate metrics is essential for meaningful evaluation.

Design comparable and use-case-grounded evaluations. The field would benefit from standardized evaluation pipelines that apply uniformly across anonymization strategies. Currently, text anonymization methods are frequently evaluated under different notions of privacy. For example, while redaction approaches are evaluated for identifier removal, synthetic data generation methods are evaluated using membership inference attacks. The lack of standardization makes it difficult to compare the practical usability of these approaches. Evaluation protocols should be grounded in realistic scenarios and expected use cases, rather than tailored to probing the specific proposed method.

Support human-centered and context-aware evaluation. Current metrics often overlook privacy risks that arise from context or user expectations. Approaches such as motivated intruder tests-where a human tries to re-identify records using web searches or domain knowledge—contextual acceptability judgments, and scenario-based probing can help capture privacy violations not visible through token-level leakage scores. While these methods are expensive, they offer high-fidelity signals that better reflect real-world privacy concerns.

Bridge technical metrics with legal standards. Technical evaluations should be interpretable in light of legal definitions of identifiability and risk, recognizing that strong performance on token-level metrics may not satisfy privacy laws or user expectations. Integrating adversarial simulations, auxiliary knowledge tests, and plausibility-based linkage metrics can help ensure evaluations better reflect regulatory expectations. At the same time, current policies often lag behind emerging threats. Over time, robust and transparent evaluation metrics, especially those grounded in real-world risks, should inform the development of improved legal standards and regulatory benchmarks.

Scale and structure human-in-the-loop evaluation. Manual re-identification or attribute inference studies offer valuable insights, but are costly and difficult to reproduce. To make them more reproducible and scalable, future work should develop annotation protocols, intruder test guidelines, and hybrid heuristics that combine automation with targeted human review. Establishing norms for reporting such studies would also support transparency and comparison.

By addressing these issues, future research can move toward a more comprehensive, reliable, and socially grounded framework for evaluating privacy in text anonymization.

7 Conclusion

Text anonymization remains an essential yet difficult component of privacy-preserving NLP. Our survey identifies six distinct privacy objectives reflected in existing metrics and highlights gaps between current evaluation practices and the broader legal, social, and practical standards that define meaningful privacy protection.

To move toward more rigorous and relevant evaluation, we call for clearer alignment between stated privacy goals and chosen metrics, greater attention to adversarial and contextual risks, and stronger integration of human-centered perspectives. As privacy risks grow with increasingly powerful generative models, a structure and context-aware evaluation framework will be key to ensuring responsible data sharing and model deployment.

Limitations

This survey focuses exclusively on post hoc evaluation metrics for privacy in text anonymization. We do not assess the effectiveness of anonymization methods themselves. We also do not conduct a thorough review of other privacy paradigms, such as model privacy (except where they relate to our work) or federated learning.

Our inclusion criteria require papers to explicitly report at least one privacy metric, which may bias our sample toward works that adopt quantifiable evaluation practices. Finally, while we discuss legal and social notions of privacy, our analysis is necessarily interpretive and does not constitute formal legal guidance.

References

John Abascal, Stanley Wu, Alina Oprea, and Jonathan Ullman. 2024. TMI! Finetuned Models Leak Private Information from their Pretraining Data. *arXiv preprint*. ArXiv:2306.01181 [cs].

Vasco Alves, Vitor Rolla, João Alveira, David Pissarra, Duarte Pereira, Isabel Curioso, André Carreiro, and Henrique Lopes Cardoso. 2024. Anonymization Through Substitution: Words vs Sentences. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 85–90, Bangkok, Thailand. Association for Computational Linguistics.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. 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.

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.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership Inference Attacks From First Principles. *arXiv preprint*. ArXiv:2112.03570 [cs].

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. *arXiv preprint*. ArXiv:2012.07805 [cs].

Camilla Casula, Elisa Leonardelli, and Sara Tonelli. 2024. Don't Augment, Rewrite? Assessing Abusive Language Detection with Synthetic Data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11240–11247, Bangkok, Thailand. Association for Computational Linguistics.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. 2025. CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications. *arXiv preprint*. ArXiv:2410.13387 [cs].

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.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating Synthetic Data Generation from User Generated Text. *Computational Linguistics*, 51(1):191–233. Place: Cambridge, MA Publisher: MIT Press.

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, pages 2349–2359, New York, NY, USA. Association for Computing Machinery.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*.

Samer El Kababji, Nicholas Mitsakakis, Xi Fang, Ana-Alicia Beltran-Bless, Greg Pond, Lisa Vandermeer, Dhenuka Radhakrishnan, Lucy Mosquera, Alexander Paterson, Lois Shepherd, Bingshu Chen, William E. Barlow, Julie Gralow, Marie-France Savard, Mark Clemons, and Khaled El Emam. 2023. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clinical Cancer Informatics*, (7):e2300116.

European Parliament and Council. 2016. Regulation - 2016/679 - EN - gdpr - EUR-Lex. Doc ID: 32016R0679 Doc Sector: 3 Doc Title: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection

Regulation) (Text with EEA relevance) Doc Type: R Usr_lan: en.

Yuanyuan Feng, Brad Stenger, and Shikun Zhang. 2024. Contextual Acceptance of COVID-19 Mitigation Mobile Apps in the United States: Mixed Methods Survey Study on Postpandemic Data Privacy. *Journal of Medical Internet Research*, 26(1):e57309. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

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. *arXiv preprint*. ArXiv:2407.02956 [cs].

Fadi Hassan, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer. 2019. Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365. ISSN: 2324-9013.

Shuo Huang, William MacLean, Xiaoxi Kang, Anqi Wu, Lizhen Qu, Qiongkai Xu, Zhuang Li, Xingliang Yuan, and Gholamreza Haffari. 2024. NAP^2: A Benchmark for Naturalness and Privacy-Preserving Text Rewriting by Learning from Human. *arXiv preprint*. ArXiv:2406.03749 [cs].

Timour Igamberdiev and Ivan Habernal. 2023. 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.

Daniel Jiménez-López, Nuria Rodríguez-Barroso, M. Victoria Luzón, Javier Del Ser, and Francisco Herrera. 2025. Membership Inference Attacks Fueled by Few-Shot Learning to Detect Privacy Leakage and Address Data Integrity. *Machine Learning and Knowledge Extraction*, 7(2):43. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing Privacy Leakage in Large Language Models. *arXiv preprint*. ArXiv:2307.01881 [cs].

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024. PrivLM-Bench: A Multi-level Privacy Evaluation Benchmark for Language Models. *arXiv preprint*. ArXiv:2311.04044 [cs].

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Darshini Mahendran, Changqing Luo, and Bridget T. Mcinnes. 2021. Review: Privacy-Preservation in the Context of Natural Language Processing. *IEEE Access*, 9:147600–147612.

Benet Manzanares-Salor, David Sánchez, and Pierre Lison. 2022. Automatic Evaluation of Disclosure Risks of Text Anonymization Methods. In *Privacy in Statistical Databases*, pages 157–171, Cham. Springer International Publishing.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. 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.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024a. 1-Diffractor: Efficient and Utility-Preserving Text Obfuscation Leveraging Word-Level Metric Differential Privacy. *arXiv preprint*. ArXiv:2405.01678 [cs].

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024b. 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, Alexandra Klymenko, Alexander Karpp, and Florian Matthes. 2025. Investigating User Perspectives on Differentially Private Text Privatization. *arXiv preprint*. ArXiv:2503.09338 [cs].

Stephen Meisenbacher and Florian Matthes. 2024. 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. 2024c. A Comparative Analysis of Word-Level Metric Differential Privacy: Benchmarking the Privacy-Utility Tradeoff. 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.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and

Yejin Choi. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. *arXiv preprint*. ArXiv:2310.17884 [cs].

John Morris, Justin Chiu, Ramin Zabih, and Alexander Rush. 2022. Unsupervised Text Deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maximilian Mozes and Bennett Kleinberg. 2021. No Intruder, no Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization. *arXiv preprint*. ArXiv:2103.09263 [cs].

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. ISSN: 2375-1207.

Ali Naseh and Niloofar Mireshghallah. 2025. Synthetic Data Can Mislead Evaluations: Membership Inference as Machine Text Detection. *arXiv preprint*. ArXiv:2501.11786 [cs].

Seth Neel and Peter Chang. 2024. Privacy Issues in Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.06717 [cs].

Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review*, 79(1):119.

Office for Civil Rights (OCR). 2012. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Last Modified: 2025-02-03T17:57:01-0500.

Shidong Pan, Zhen Tao, Thong Hoang, Dawen Zhang, Tianshi Li, Zhenchang Xing, Xiwei Xu, Mark Staples, Thierry Rakotoarivelo, and David Lo. 2024. A {NEW} {HOPE}: Contextual Privacy Policies for Mobile Applications and An Approach Toward Automated Generation. pages 5699–5716.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, and Minhui Xue. 2025. Reconstruction of differentially private text sanitization via large language models. *Preprint*, arXiv:2410.12443.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural Text Sanitization with Explicit Measures of Privacy Risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

Ambika Pawar, Swati Ahirrao, and Prathamesh P. Churi. 2018. Anonymization Techniques for Protecting Privacy: A Survey. In *2018 IEEE Punecon*, pages 1–6.

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.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Kopf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Beguelin. 2023. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning . In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 327–345, Los Alamitos, CA, USA. IEEE Computer Society.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks against Machine Learning Models. *arXiv preprint*. ArXiv:1610.05820 [cs].

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy Via Inference with Large Language Models. *arXiv preprint*. ArXiv:2310.07298 [cs].

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. 2025. On the Vulnerability of Text Sanitization. *arXiv preprint*. ArXiv:2410.17052 [cs].

Isabel Wagner and David Eckhoff. 2019. Technical Privacy Metrics: a Systematic Survey. *ACM Computing Surveys*, 51(3):1–38. ArXiv:1512.00327 [cs].

Yuyang Wang, Xianjia Meng, and Ximeng Liu. 2023. Differentially Private Recurrent Variational Autoencoder For Text Privacy Preservation. *Mob. Netw. Appl.*, 28(5):1565–1580.

Zifeng Wang and Jimeng Sun. 2022. PromptEHR: Conditional Electronic Healthcare Records Generation with Prompt Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2885, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. 2024. Membership Inference Attacks Against In-Context Learning. *arXiv preprint*. ArXiv:2409.01380 [cs].

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2025. A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage. *arXiv preprint*. ArXiv:2504.21035 [cs].

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.

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.

Shikun Zhang, Yan Shvartzshnaider, Yuanyuan Feng, Helen Nissenbaum, and Norman Sadeh. 2022. Stop the Spread: A Contextual Integrity Perspective on the Appropriateness of COVID-19 Vaccination Certificates. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 1657–1670, Seoul Republic of Korea. ACM.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28, Yokohama Japan. ACM.