# The Conundrum of Trustworthy Research on Attacking Personally Identifiable Information Removal Techniques

Sebastian Ochs[1,2*] and Ivan Habernal[1,3]

[1]Trustworthy Human Language Technologies
www.trusthlt.org
[2]Technical University of Darmstadt, Germany
sebastian.ochs@tu-darmstadt.de
[3]Research Center for Trustworthy Data Science and Security,
Ruhr University Bochum, Germany
ivan.habernal@ruhr-uni-bochum.de

*Removing personally identifiable information (PII) from texts is necessary to comply with various data protection regulations and to enable data sharing without compromising privacy. However, recent works show that documents sanitized by PII-removal techniques are vulnerable to reconstruction attacks. Yet, we suspect that the reported success of these attacks is largely overestimated. We critically analyze the evaluation of existing attacks and find that data leakage and data contamination are not properly mitigated, leaving the question whether or not PII removal techniques truly protect privacy in real-world scenarios unaddressed. We investigate possible data sources and attack setups that avoid data leakage and conclude that only truly private data can allow us to objectively evaluate vulnerabilities in PII removal techniques. However, access to private data is heavily restricted—and for good reasons—which also means that the public research community cannot address this problem in a transparent, reproducible, and trustworthy manner.*

## 1. Introduction

Nobody wants their private text data, such as personal messages, medical notes, or presence in court rulings, to be freely accessible by the general public. Simultaneously, research on sensitive text domains like law and healthcare is necessary for developing technologies that benefit humanity (Demner-Fushman, Chapman, and McDonald 2009; Sudlow et al. 2015; Zhong et al. 2020). To conduct research on private text data while

* Corresponding author.

Action Editor: Xuanjing Huang. Submission received: 12 June 2025; revised version received: 25 January 2026; accepted for publication: 4 March 2026.

https://doi.org/10.1162/COLI.a.615

© 2026 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

complying with regulations and privacy laws, personally identifiable information (PII) must be removed from affected documents. Manually removing PII from texts is a slow and tedious process (Dorr et al. 2006) that does not scale well given the amounts of sensitive documents available in certain domains (Keuchen and Deuber 2022). PII removal tools (Sweeney 1996; Kleinberg, Davies, and Mozes 2022) aim to protect the privacy of individuals to comply with the legal frameworks, while enabling publishing useful data that no longer contains sensitive information. However, PII removal tools do not guarantee privacy in any formal way. Instead, they mimic how humans usually process text documents to identify and remove private information from the data. This leaves the PII removed texts vulnerable to attacks, as humans can partially reconstruct the removed information with significant effort and background knowledge (Carrell et al. 2020; Deuber, Keuchen, and Christin 2023). Meanwhile, large language models (LLMs) are becoming increasingly more powerful and are capable of inferring private information from text (Staab et al. 2024). This raises concerns about whether LLMs may also be able to infer sensitive information from PII removed documents.

Existing approaches that attack PII removed texts claim that they successfully recover parts of the private information (Nyffenegger, Stürmer, and Niklaus 2024; Patsakis and Lykousas 2023; Lukas et al. 2023; Charpentier and Lison 2025). However, it remains unclear whether their attacks succeed because the PII removal method itself is insufficient, or because the simulated attacker already has access to the original data, for example through applying the attack on well-known public data or by data contamination and data memorization by the attack model. Moreover, there is a lack of guidance on how to properly design, execute, and evaluate adversarial attacks against PII removal tools.

We therefore pose two research questions. First, is the evaluation of existing methods for PII reconstruction attacks fundamentally flawed? In this article, we show what contributes to the inflation of re-identification scores and overestimation of the general ability of the proposed attacks. Second, is it possible for public researchers to address these potential flaws without access to real, sensitive data? We argue that without access to private data, we cannot reliably determine whether the attack models used in previous research with undisclosed pretraining data, especially proprietary LLMs, are genuinely inferring sensitive information from PII removed texts or simply reproducing text spans they memorized during pretraining. To highlight the current shortcomings of PII reconstruction attacks, we further conduct and analyze two experiments on real-world data from Czech court announcements (legal domain) and from personal English travel videos (YouTube vlogs), both extremely unlikely being leaked to any dataset for LLM pre-training.

In the following, we present and briefly discuss related work (Section 2), and then provide background on PII in texts, regulations, and its removal techniques (Section 3). Before addressing our research questions, we point out flaws in PII-related regulations and removal tools (Section 4). We then answer our first research question (Section 5) and turn to the second research question by outlining a valid attack setup (Section 6) and why real private data is needed for proper evaluation (Section 7).

## 2. Related Work

While most of the related work deals with PII removal techniques or attacks, we are not aware of any existing work that directly questions whether the experimental design is inherently flawed. For example, some surveys in the medical domain review the

application of PII removal techniques to electronic health records from the US before the deep learning era (Meystre et al. 2010; Ford et al. 2016). Recently, Sousa and Kern (2023) categorize several privacy-preserving NLP methods based on deep learning, including PII removal, and discuss their utility to comply with data protection laws.

Lison et al. (2021) identify several issues inherent to PII removal techniques (which they term "text anonymization"). The authors highlight that removing predefined categories of PII from text documents is insufficient to provide any formal privacy guarantees, and that human annotators often disagree about which text spans contain private information. Furthermore, Lison et al. (2021) compare typical PII removal methods based on named entity recognition to privacy-preserving data publishing methods from other domains, such as *k*-anonymity (Sweeney 2002), C-sanitize (Sánchez and Batet 2016), and differential privacy (Dwork 2006) and discuss possible difficulties when applying these methods to text. The authors argue against solely relying on PII removal techniques and emphasize the need to incorporate context, utility, and re-identification metrics into the private data-sharing pipeline. However, they do not question the potential flaws in attacks against PII removal techniques or investigate the role of information leakage in these attacks.

In this article, we therefore not only argue that there exist flaws in PII removal techniques, but there are also flaws in the attacks against them.

## 3. Personally Identifiable Information in Texts

PII refers to any attributes that can be used—either alone or in combination—to uniquely identify a person. Data, especially documents from the medical, legal, or financial domain, often contain names, addresses or phone numbers that can single out an individual on their own, called *direct identifiers*. Besides direct, there also exist *indirect identifiers* or *quasi-identifiers*, which can distinguish a person when they are combined with each other. For example, date of birth, ZIP codes, gender, or occupation are not unique to a person, but linking databases together where those attributes appear may lead to the identification of individuals (Sweeney 2000). In contrast to traditional databases, natural language texts can also reveal personal information through syntactic features, such as writing style (Koppel, Argamon, and Shimoni 2002; Verhoeven, Daelemans, and Plank 2016), or semantic content that can be linked to PII (Elazar and Goldberg 2018; Staab et al. 2024).

### 3.1 PII in Regulations

When a dataset is shared with third parties, e.g., for data analysis or research, it is essential that the privacy of contributing persons is protected. Therefore, legislators around the world created different legal frameworks for handling private data. While these regulations often establish similar definitions of PII and require the removal of both direct and indirect identifiers from texts, they differ in the scope of what must be removed or conditions under which some identifiers may be retained.

The General Data Protection Regulation (GDPR) (European Commission 2016), for example, describes how personal data (synonymous with PII) of individuals in the EU must be protected. It lists several types of personal data, particularly the "name, an identification number, location data, an online identifier" and "factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity" of a person (GDPR, Art. 4(1)) (European Commission 2016). Additionally, the GDPR also defines categories that require stricter protection, such as the "racial or ethnic origin,

political opinions, religious or philosophical beliefs" of an individual (GDPR, Art. 9(1)) (European Commission 2016).

In contrast to the GDPR's broad coverage, the U.S. Health Insurance Portability and Accountability Act (HIPAA) (U.S. Congress 1996) applies more specifically to personal information found in medical data. It defines the "Safe Harbor" method, which requires the removal of 18 specific types of protected health information (PHI: including names, location, and phone numbers) for medical documents to be considered de-identified. Once the method is applied to health-related documents and all PHI are removed, the data is no longer subject to the HIPAA privacy rules and may be published.

Similar to how HIPAA regulates removing PHI of medical data in the United States, the GDPR serves as a legal framework to ensure the protection of individuals' privacy when publishing documents. For example, decisions of the Court of Justice of the European Union are "anonymized" by replacing names of natural persons with synthetically generated, fictitious names and released to the public, (apparently) meeting GDPR standards.[1] In recent years, algorithmic tools are increasingly applied to remove PII from court documents, which speeds up the publication process while still complying with GDPR privacy protection across EU jurisdictions (Terzidou 2023). However, the implications of taking the GDPR seriously for PII removal and NLP in general can be disastrous. Weitzenboeck et al. (2022) show that if one wants make sure text data are anonymized according to the most stringent interpretation of the GDPR, the redacted text output becomes just useless (Weitzenboeck et al. 2022, Table 2 on p. 197). This stringent interpretation of the GDPR should prevent any linkage attacks; in other words, it must be impossible to link the anonymized documents to any other documents that contain personal information. But here we encounter an unsolvable problem: Weitzenboeck et al. (2022) anonymize court documents against an adversary *who already owns the original unredacted documents*. Therefore, as long as someone has a legal obligation to keep the original document, which is the case for many datasets, publishing a useful anonymized version violates the GDPR on the one hand, but a strict anonymization is useless on the other hand.

### 3.2 Definitions

In related works, the terminology associated with the removal of PII is often ambiguous. We therefore define several key terms to clarify their usage in the context of natural language processing in our article.

*PII Removal.* We use **PII removal** as a general term for the practice of removing a predefined set of private information from texts to mitigate privacy risks.

*Anonymization.* **Anonymization** is the process of removing or modifying PII from text documents, such that re-identification is not possible with current technology. As stated in the GDPR, data is considered *anonymized* if re-identification of a particular person cannot be reasonably achieved by linking the data to other datasets or applying any currently existing adversarial attacks at the time of the data processing (European Commission 2016, Recital 26). It should be noted that this definition does not account for potential future attacks.

1 See "Fictitious names in anonymised cases" at `https://curia.europa.eu/jcms/jcms/p1_3869098/en/`.

*De-identification.* According to HIPAA, **de-identification** involves altering or removing PII from text documents so that either re-identification risks are low, according to a human expert, or all occurrences of the 18 designated PHI are removed (U.S. Government 2013). In contrast to anonymization, *de-identification* is therefore less strict as HIPAA accepts a low level of risk, if the effort required for re-identification is deemed sufficiently high by the expert's judgment. For example, the electronic health records aggregated by CancerLinQ are de-identified using expert determination (Potter et al. 2020, p. 3).

*Sanitization.* Text **sanitizing** describes the process of obscuring sensitive information in texts while preserving their utility (Jiang et al. 2009; Papadopoulou et al. 2022; Olstad, Papadopoulou, and Lison 2023). The privacy protection mechanisms of *sanitization* are not necessarily limited to removing or modifying PII in texts, but may also be based on differentially private rephrasing methods (Yue et al. 2021; Chen et al. 2023). Unlike anonymization or de-identification, *sanitization* does not follow a unified re-identification risk standard like HIPAA or GDPR, but may utilize existing frameworks.

### 3.3 Typology of PII Removal Techniques

*Redaction.* Deleting PII from text documents is referred to as **redaction**. This process can disrupt the syntactic coherence of sentences as grammatical constituents such as names or other entities are completely removed.

*Masking.* Masking obscures PII by substituting them with non-semantic placeholders, for example mask tokens. These placeholders reveal the placements of PII in the original texts but do not disclose any information beyond their position (Berman 2003, Table 2).

*Replacement.* Preserving the contextual meaning of sensitive data is achieved by *replacing* PII with more relevant but general information. For example, names could be replaced by a special token [name] or locations with [location] (Neamatullah et al. 2008, Table 3.5).

*Pseudonymization.* The GDPR defines **pseudonymization** as replacing PII with substitutes so that it cannot be connected to the affected person without the use of additional data (European Commission 2016, Art. 4 (5)). Under this definition, the relation between the PII and their surrogates can be stored to enable later reconstruction of the original data. However, the primary goals of *pseudonymization* in NLP are privacy preservation and utility, not the ability to reconstruct the original texts. We therefore deviate from the GDPR in this paper and describe *pseudonymization* as substituting PII with synthetically generated data without keeping the link between the synthetic and original information. In NLP, *pseudonymization* is often supported by context-aware language models to retain text utility for downstream tasks while privacy is preserved (Eder et al. 2022).

We provide examples of each described PII removal technique applied to a sentence containing ficticious PII in Table 1.

### 3.4 PII Removal Tools

The shift towards automatic PII removal from texts came in response to earlier practices. Before automatization, PII was typically removed by manually blacking out their

**Table 1**
We generated a message from a ficticious person with ChatGPT 4o mini (`https://chatgpt.com/share/681de483-181c-800e-ae40-88cc2fefb0f9`) and applied common PII removal techniques to the same input sentence with Microsoft Presidio (`https://huggingface.co/spaces/presidio/presidio_demo`).

| Technique | Example |
|---|---|
| **Original** | Hi, I'm Jane Doe, and I live at 123 Maple Street, Springfield. You can reach me at (555) 123-4567 or email me at janedoe@email.com. |
| **Redaction** | Hi, I'm, and I live at 123, . You can reach me at or email me at . |
| **Masking** | Hi, I'm ********, and I live at 123 ************, **********. You can reach me at ************ or email me at **************om. |
| **Replacement (Tag)** | Hi, I'm <PERSON>, and I live at 123 <LOCATION>, <LOCATION>. You can reach me at <PHONE_NUMBER>or email me at <EMAIL_ADDRESS>. |
| **Pseudonymization** | Hi, my name is Alex Carter, and I live at 123 Oak Drive, Rivertown. You can reach me at (206) 482-7743 or via email at acarter88@mailnet.com. |

occurrences in text documents, a task that requires a lot of effort and is time-consuming (Dorr et al. 2006). Even nowadays, PII are manually removed from German court decisions, resulting in the publication of only 2% of the rulings issued by German courts (Keuchen and Deuber 2022).

*Early Methods.* One of the earliest tools was the Scrub system (Sweeney 1996), which used a rule-based approach to detect sensitive information and replaced it with handcrafted values from a lookup table. Through pattern matching and handcrafted rules, Scrub replaced specific types of PII such as names, addresses, dates, and phone numbers from medical discharge summaries. However, Scrub struggled with context-dependent entities or ambiguous terms, for example, PII with references to family relationships or medical conditions. In contrast, systems based on named entity recognition (NER), such as HIDE (Gardner and Xiong 2008), were reported to be more accurate and flexible in detecting a wider range of private information. Rule-based techniques still play a significant roll in NLP to remove PII from text corpora, especially number-based information such as phone and credit card numbers or IP addresses; see Laurençon et al. (2022, Section 3.3) or Soldaini et al. (2024, Section 5.3).

*Contemporary Methods.* Currently, a popular open-source text de-identification tool is Presidio,[2] a software that combines both rule-based and NER techniques to detect and replace PII. Presidio offers support for multiple languages and different PII replacement methods, as displayed in Table 1, and Kotevski et al. (2022) suggest its use for de-identifying medical data. One drawback of Presidio is lacking coreference resolution, which can lead to inconsistent de-identification when one piece of private information is referenced multiple times and in different forms within a document. Other methods like Textwash (Kleinberg, Davies, and Mozes 2022) use coreference resolution, but only for specific PII types that are easily traceable, such as names and locations. Beyond using

2 `https://github.com/microsoft/presidio`.

NER and rule-based systems, Liu et al. (2023) propose to de-identify texts with GPT-4 in a zero-shot setting. While they achieve high accuracy on a clinical notes dataset, it is questionable whether private information can be sent to an online LLM provider in the first place, as it inadvertently leaks the sensitive data.

*Benefits of PII Removal Tools.* Automated PII removal tools are very efficient at removing well-structured sensitive information from free-form texts. With the integration of more sophisticated detection models based on NER, their performance on context-dependent PII has significantly improved compared to earlier rule-based approaches (Pilán et al. 2022). PII removal seems to offer an empirically sound defense against attacks, as it currently takes significant effort and extensive background knowledge to re-identify persons from de-identified texts (Carrell et al. 2020; Deuber, Keuchen, and Christin 2023). Tools such as Presidio and Textwash also support data holders to comply with the privacy protection standards defined in legal frameworks.

### 3.5 Defense Methods Beyond PII Removal

A key limitation of PII removal is the degradation of text utility. By redacting all explicit identifiers, the resulting text often loses key features that are essential for meaningful data analysis (Pal et al. 2024). While Presidio and Textwash offer different strategies to replace PII with semantic placeholders (pseudonymization), researchers also explored other options such as text rewriting to better preserve the semantic meaning of documents while removing PII.

*Adversarial Text Rewriting.* Staab et al. (2025) propose to iteratively rewrite a text document with GPT-4 to obscure the sensitive information detected by an adversarial GPT-4 model. Yang, Zhu, and Gurevych (2025) extend this approach by also including an LLM aiming to preserve the utility of the rewritten text for downstream tasks. These methods, however, raise some privacy concerns, as the data is sent to the closed-source GPT-4 model for inference, inevitably sharing the private data with a closed-source company before the sanitization process. Even if adversarial text rewriting methods empirically defend more strongly against the selected adversaries in the respective paper than PII removal techniques, they still lack any formal privacy guarantees. Given these limitations, and since our main focus is PII reconstruction attacks, we do not discuss adversarial text rewriting methods further in our article.

*PII Removal Combined with DP.* One hybrid approach that combines PII removal with differential privacy is presented by Mouhammad et al. (2023). The authors first detect PII within texts via an NER-based approach, replace the identified spans with special tokens, and then rewrite the resulting text under differential privacy guarantees (Igamberdiev, Arnold, and Habernal 2022). However, their results show that text utility suffers considerably, even when using very high values of $\epsilon$ that offer little to no meaningful privacy protection. We therefore did not consider such approaches for further discussion in our position paper.

## 4. PII-related Privacy Regulations and Removal Tools Have Inherent Flaws

Despite their advantages, PII regulations and removal techniques have limitations that can prevent them from fully protecting the privacy of individuals.

Interpretations of what exactly constitutes personal data differ across legal frameworks from various countries. This poses a challenge to the developers of PII removal tools, as complying with one regulation does not guarantee compliance with others. In the United States, for example, text de-identification systems built for HIPAA focus on removing the 18 types of PHI from medical documents to adequately protect patient privacy in accordance with the legislation. But when these tools are then applied to other domains, such as financial data, they may fail to detect unique identifiers from other domains, e.g., financial codes like IBANs or SWIFT/BIC codes, which fall outside of the scope of HIPAA. Beyond domain-dependency, many countries—such as China and Brazil—have adopted privacy laws inspired by the GDPR.[3] However, the subtle differences between different regulations complicate the generalization of PII removal tools beyond the adaptation to a new language. For example, the GDPR explicitly lists sexual orientation and political beliefs as sensitive personal information (European Commission 2016, Art. 9(1)), while the Chinese Personal Information Protection Law (PIPL) does not.[4]

*Unbounded List of PII.* More challenging is the fact that the list of PII requiring removal from texts is potentially unbounded, as personal data is broadly interpreted as "any information relating to an identified or identifiable natural person" according to Art. 4(1) of the GDPR (European Commission 2016). While the regulation provides some examples of personal data, it deliberately leaves the set of PII open-ended. This significantly increases the difficulty for data holders to share their data with the research community, as the degree of safeguarding required by the GDPR (European Commission 2016, Art. 89) is undefined (Peloquin et al. 2020; Vlahou et al. 2021). In the text domain, successful anonymization is therefore subject to interpretation by the legislators, with data holders facing the risk of non-compliance, trial in courts and potential fines when personal data is exposed after anonymized release.[5] Given a strict interpretation of WP 216[6] (a guidance document on anonymization issued by EU regulators before the GDPR, which remains influential), Weitzenboeck et al. (2022) even argue that anonymization of unstructured data is essentially impossible.

*Vague Capabilities of the Attacker.* In cryptography, encryption methods are designed by considering future resources of the attacker to ensure the protection of data (Diffie and Hellman 1976, Section VI). For example, existing research on cryptographic algorithms is motivated by the future threat to the existing encryption standards posed by quantum computing.[7] In contrast to cryptography, where future advancements are considered, the capabilities of the attacker on texts protected by PII removal techniques are not sufficiently defined by regulations when it comes to protecting privacy in textual data. While the GDPR states that current state of the art re-identification methods and their technological development have to be taken into consideration at the time of the text anonymization process (European Commission 2016, Recital 26), it does not define the resources of the attacker beyond the criterion of "means reasonably likely to be used" for re-identification. This makes it difficult for data holders to adequately address

3 https://iapp.org/news/a/three-years-in-gdpr-highlights-privacy-in-global-landscape.
4 https://personalinformationprotectionlaw.com/PIPL/article-28/.
5 https://openjur.de/i/openjur_wird_verklagt.html.
6 https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
7 https://csrc.nist.gov/projects/post-quantum-cryptography.

vulnerabilities emerging from the advancements in NLP and LLMs, which threaten the privacy of the individual, such as training data extraction (Carlini et al. 2021) or LLM inferences of private information (Staab et al. 2024). Without a clear attack formalization to follow against text anonymization, researchers improvise their own attack setups, potentially leading to flawed attack scenarios, a topic we explore in more detail in Section 5.

*Lack of Guarantees Against Future Attacks.* The major drawback of PII removal techniques, however, is that they do not provide formal privacy guarantees that would be impossible to overcome, under the formally specified technique and the explicitly stated attacker capabilities. In contrast to differential privacy (Dwork 2006), which offers mathematically provable (yet probabilistic) privacy protection for data, **the privacy provided by PII removal is solely based on empirical risk evaluations that cannot account for potentially stronger attack methods**.

## 5. Critical Analysis of Adversarial Attacks on Data Protected by PII Removal

Early attempts at de-identifying structured data containing textual information led to inadequate privacy protection, which resulted in the exposure of personal information of individuals who provided their data to platforms, certainly under the assumption that it would remain anonymous (Narayanan and Shmatikov 2008; Sweeney et al. 2017).[8] Such data leaks may have influenced the decision of legislators to implement stricter privacy regulations to protect user data, such as HIPAA or the GDPR. Current state-of-the-art de-identification methods therefore remove more PII from texts than required by previously implemented privacy regulations. But with the growing reasoning capabilities of LLMs (Mao et al. 2024), redacted PII might be at risk to be revealed through attacks. However, *are these attacks successful due to stronger model abilities, such as inferences from context, or due to inadequate attack setups and potential side-channel leaks?*

To address our first research question, whether existing evaluations of PII reconstruction attacks are fundamentally flawed, we focus on recent work that explicitly targets documents where PII removal was applied and LLMs are the main attack method, either through retrieval-augmented generation, direct prompting, or infilling. To identify such publications, we resorted to an extensive manual search, as systematic, keyword-based searches proved to be ineffective, due to the small yet fragmented research area, where terminology is used inconsistently across papers. While this selection is not exhaustive, it is representative of the current directions in adversarial attacks on PII-removed texts with LLMs.

### 5.1 Leakage Through Media Reporting

We exemplify a particular type of side-channel leakage using the following paper. Nyffenegger, Stürmer, and Niklaus (2024) proposed an attack on PII-sanitized texts, namely, on court rulings of the federal supreme court of Switzerland, which were sanitized by masking PII through human annotators assisted by an automatic PII detection system. The authors manually re-identified persons involved in seven court decisions to create a gold standard for the attack by manually linking the rulings to Swiss news articles that contain relevant information about each case, for example, utilizing the

8 https://www.cnet.com/tech/tech-industry/aol-apologizes-for-release-of-user-search-data/.

penalty or case file numbers mentioned in the media. Using the additional information, the researchers link more news articles to the case until they successfully re-identify the name of seven individuals that were involved in the court rulings. For the final dataset, the researchers also mix in irrelevant articles to ensure that the resulting dataset does not only include relevant data for the court decisions. The attack itself is an approach based on retrieval-augmented generation with LLMs with the goal of recovering the name of the case participants. The authors retrieve the top five relevant articles for each sanitized court ruling and prompt GPT-4 to re-identify the person from the given context, which was successful for 5 out of the 7 cases.

The attack framework presented by Nyffenegger, Stürmer, and Niklaus (2024) does not suffice to disprove the effectiveness of PII removal methods. As the authors report, when providing GPT-4 with relevant news articles that contain the real name of a litigant in the sanitized court ruling, the model is able to recognize the person. However, in these cases, the news articles reporting on the court rulings are already public knowledge and are the decisive factor that lead to the re-identification of affected parties. *PII removal techniques cannot offer protection in such cases, as the underlying data is already publicly available information before the text sanitization, and the names of the case participants are no longer private.*

### 5.2 Leakage Through Public Knowledge

Another attack on sanitized text documents is proposed by Patsakis and Lykousas (2023). For their attack, the authors utilize a dataset published by Textwash that contains manually rewritten Wikipedia biographies of celebrities. Textwash replaced the PII in the biographies with corresponding named entity tags. The researchers attacked the sanitized biographies with the aim to recognize the original celebrity. Using GPT-3 (Brown et al. 2020), the attack successfully recognized 59% of famous people. An alternatively tested strategy using PII with pseudonymization to intentionally mislead the attacker led to a drop in re-identification, down to 50%, after revising the attack prompt.

However, a major limitation of the proposed attack by Patsakis and Lykousas (2023) is that it was conducted on well-known, public information about famous people such as actors, musicians, and politicians. Although the rewritten Wikipedia biographies from Textwash may not appear verbatim in the GPT-3 training data, the model was definitely trained on the original Wikipedia entries (Brown et al. 2020, Section 2.2), as well as several other information sources about the "anonymized" celebrities (Table 2). Since a vast amount of PII of public figures is already publicly available information, it is possible to re-identify famous people using the semantic context of documents that is left after PII removal (Lohr et al. 2024).

We demonstrate the flaws of this benchmark in Figure 1, which shows that only a few trigger words are enough to retrieve all PII from a language model. Moreover, Table 3 shows another problem in the evaluation. Both solutions of the PII reconstruction are possible, as both Emma Watson and Angelina Jolie would perfectly fill the gaps. Yet the benchmark seems to be skewed towards more famous persons who might be also over-represented in the Web data fed into LLM pre-training, leading to artifacts in the benchmark and the attack evaluation. In other words, if Angelina Jolie were in the gold standard data, the PII attack filling in Emma Watson would fail, thus lowering the attack score from which we might conclude that the PII removal technique is safer. But both for Angelina Jolie and Emma Watson, the PII removal tool would produce the *very same output*.

**Table 2**
Shortened example from Table 5, Patsakis and Lykousas (2023) of GPT-4 recognizing Emma Watson from a PII removed text (`https://chatgpt.com/share/6825cf22-06e4-800e-9de1-edf923c9dc23`).

| | |
|---|---|
| **System** | Find which celebrity matches the following description and match the missing information |
| **User** | PERSON_FIRSTNAME_1 PERSON_LASTNAME_2 is a NUMERIC_4 year old actress, born in LOCATION_2 but raised in LOCATION_1. PRONOUN is most famously known for playing PERSON_FIRSTNAME_2 PERSON_LASTNAME_1 in the OTHER_IDENTIFYING_ATTRIBUTE_2 franchise. [...] PRONOUN is very involved in activism and is a very outspoken feminist. |
| **Assistant** | Emma Watson is a 31 year old actress, born in Paris but raised in the UK. She is most famously known for playing Hermione Granger in the Harry Potter franchise. [...] she is very involved in activism and is a very outspoken feminist. |

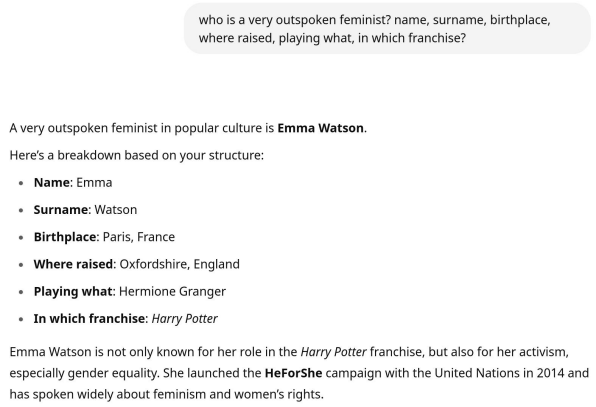


**Figure 1**
Solving the task by Patsakis and Lykousas (2023) by querying ChatGPT-4o with a few keywords only (cf. Table 2).

**Table 3**
Both Emma Watson and Angelina Jolie would be a valid solution of the PII reconstruction attack by Patsakis and Lykousas (2023).

| | |
|---|---|
| **Gold standard** | Emma Watson is a 31 year old actress, born in Paris but raised in the UK. She is most famously known for playing Hermione Granger in the Harry Potter franchise. [...] she is very involved in activism and is a very outspoken feminist. |
| **Alternative solution** | Angelina Jolie is a 49 year old actress, born in Los Angeles but raised in the Palisades, New York. She is most famously known for playing Lara Croft in the Lara Croft franchise. [...] she is very involved in activism and is a very outspoken feminist. |

We therefore argue that the re-identification success of GPT-3 is not the fault of the de-identification method per se, but rather a consequence of the public nature of the utilized data. *When the original and anonymized documents share semantically similar or even identical content, recognizing the corresponding famous person becomes a pattern-matching task instead of a privacy breach.*

### 5.3 Leakage Through Memorization

Lukas et al. (2023) propose a PII reconstruction attack on sanitized documents. Here, the attacker has access to an LLM fine-tuned on the original documents. The researchers fine-tune GPT-2 models (Radford et al. 2019) with and without differentially private stochastic gradient descent (DP-SGD, a training method that injects noise into gradients and therefore limits the influence of any single data sample on the model) on three datasets; ENRON (Klimt and Yang 2004), Yelp-HEALTH,[9] and cases of the European Court of Human Rights (ECHR, Chalkidis, Androutsopoulos, and Aletras (2019)). After fine-tuning, the PII in each text of the above-mentioned datasets is masked with Presidio. The authors then attack the sanitized documents with the fine-tuned language models by filling in the masked information. As expected, the GPT-2 models trained with privacy guarantees using DP-SGD perform worse at reconstructing the masked PII, achieving under 1% accuracy on the "person" entity type on each dataset. In contrast, their non-private counterparts are able to guess the correct "person" with significantly greater accuracy, up to 15% accuracy on ENRON and up to 18% on the ECHR dataset, posing a greater risk to privacy. The authors also observe that larger model versions of GPT-2 generally leak more PII when attacked with the same method.

Although Lukas et al. (2023) claim that PII removal techniques are insufficient to prevent PII leakage, their attack does not prove that PII leak from sanitized documents because of the weak protection offered by PII removal methods. As the attacking GPT-2 models were fine-tuned on the original data *before* PII removal, the sensitive information in the data is already memorized by the model at the start of the attack and it can use the sanitized texts as context to accurately recall PII based on its training objective of next token prediction. Consequently, the ability of the attack to reconstruct PII from sanitized texts should be attributed to *memory leakage instead of limitations in the PII removal method itself*. This is further proven by the authors themselves, as GPT-2 models fine-tuned on the PII removed instead of original documents achieve 0% reconstruction accuracy.

Charpentier and Lison (2025) face a similar issue in their approach. The authors attack sanitized Wikipedia biographies, ECHR cases, and synthetic clinical notes by combining a retrieval model with an LLM-based infilling method to find relevant passages in background documents and infer the original masked text spans. However, the unsanitized versions of the sanitized Wikipedia and ECHR texts are most likely in the pretraining data of their utilized infilling models, as Charpentier and Lison (2025) acknowledge in their limitations. The authors also vary the access of background knowledge the attacker has in their setup. They show in their appendix section E that an attacker without access to the original ECHR case linked to a sanitized document, but with access to all other original cases, is able to reconstruct the name of Polish government agent J. Wołąsiewicz. However, J. Wołąsiewicz also appeared in other ECHR cases as a Polish public official. This does not show that the attack was successful in this case, but rather that (1) the same PII in ECHR court cases can be found in multiple documents

9 https://business.yelp.com/data/resources/open-dataset/.

and (2) public figures were incorrectly labeled as private entities. Again, we argue that the supposedly successful reconstruction attack does not demonstrate that PII removal techniques are a weak privacy protection method, but that attack performances are overestimated through memorization and other factors, such as mislabeling.

Even though Morris et al. (2022) do not focus on PII reconstruction attacks, they encounter an interesting phenomenon in their adversarial re-identification metric. The authors propose an unsupervised de-identification method based on *k*-anonymity, where documents are continuously redacted until the probability of re-identification estimated by an adversary falls below a threshold, depending on how many people share similar attributes. The main de-identification target is a small subset of Wikipedia biographies. However, the re-identification models of Morris et al.'s (2022) method are based on RoBERTa (Liu et al. 2019), which was pretrained on large corpora that include Wikipedia. The authors observe that in some cases the adversarial model is able to identify people from biographies where 95% of the words are masked (Morris et al. 2022). While the authors state in their limitations that memorization "is unlikely to be happening", we are not convinced by this claim, especially since the re-identification models are also fine-tuned on the biographies. It is highly probable that the adversarial model memorized the exact sentence structure of these individuals, which allows re-identification even when most of the text is replaced with masked tokens. In these cases, the de-identification method has no chance to protect the privacy of the re-identified persons, as the attacker has already memorized the original biographies and their distinctive phrasing. This is the same kind of memory leakage we also observed in Lukas et al.'s (2023) work.

### 5.4 Adversarial Attacks on PII Removed Texts Are Flawed

Overall, the estimated privacy risks posed by contemporary attacks on PII removed texts are overestimated, as either the attack models had already memorized the supposedly "private" data during training (Patsakis and Lykousas 2023; Lukas et al. 2023) or the "private" information was already publicly available in news articles (Nyffenegger, Stürmer, and Niklaus 2024). While we believe that the existing attacks are valuable, we argue that data leakage is the most likely reason for the success of these adversarial attacks against sanitized documents, giving the attacker an unrealistic advantage against PII removal techniques. Therefore, the answer to our first research question is yes: There are inherent flaws in the evaluation of current PII reconstruction attacks.

## 6. The Setup of Adversarial Attacks on PII Removed Texts

We now address our second research question: Can the flaws we identified in the previous section be resolved without access to real, sensitive data? Given the uncertainty whether previous attacks are successful because of weaknesses in PII removal or through unintentional data leakage, we first need to outline the conditions a valid attack setup must satisfy to avoid exaggerating the attack results. We cover both the defense method and threat model, before turning to the data issue in Section 7.

### 6.1 Defense Methods

When attacking PII removal methods to discover potential weaknesses, the defensive mechanism itself has to reflect real-world deployments. It must be highly accurate at detecting PII in texts to demonstrate its privacy protection abilities, while maintaining

the text utility for downstream tasks to a reasonable degree. Current state-of-the-art methods combine pattern matching with regular expressions and NER, though NER has significantly improved in recent years due to the integration of context-sensitive embeddings (Jehangir, Radhakrishnan, and Agarwal 2023). Therefore, the defense mechanism should at least use such a hybrid, state-of-the-art method that ideally aligns with legal frameworks, supporting GDPR or HIPAA compliance. Furthermore, the text de-identification technique should be open-source, ensuring that its PII detection and replacement strategies are transparent and comprehensible. This also allows researchers to identify and address potential weaknesses in the defense mechanism. In contrast, using closed-source, cloud-based services similar to Amazon Comprehend[10] or Microsoft Azure[11] to remove PII from documents would make such an error analysis more difficult. Furthermore, if the documents contain sensitive private data, it may not be advisable to share that data with proprietary services for PII removal.

### 6.2 Threat Model

Previous attacks do not demonstrate that PII removal is ineffective for privacy protection, but rather show that PII removal cannot prevent leakage when the model already memorizes the underlying data beforehand. However, we *do* believe that PII removal methods may be vulnerable to attacks where failures can be attributed entirely to PII removal.

In recent years, the ability of LLMs to infer information from context has improved significantly, partially due to better reasoning capabilities and larger model sizes (Wei et al. 2022). Furthermore, stylometric features, such as writing style, syntactic patterns, and unique phrasings, can uniquely identify individuals, as famously demonstrated by the FBI in the capture of the Unabomber (Leonard, Ford, and Christensen 2017, Section II). When provided with free-form text written by humans, LLMs are also capable of accurately attributing authors to the corresponding documents (Ai et al. 2022; Huang, Chen, and Shu 2024), which could be leveraged to infer sensitive user information. Additionally, LLMs can be exploited to extract private information from text data that is often not explicitly stated, such as age, gender, and occupation (Staab et al. 2024). LLMs can also infer locations from human-written texts based on implicit geographic features, namely, landmarks, dialects, or local cuisine (Zhang, Luo, and Huang 2025). Even the Big Five personality traits (Costa and McCrae 2008) of users can be predicted from social media posts with reasonable reliability, as demonstrated by Peters and Matz (2024).

PII removal methods are well-suited for modifying or removing explicitly stated private information in texts. However, it remains uncertain how much of the data may be recovered from the implicit information left behind. Given the capabilities of LLMs to infer various sensitive information from textual data, even when such attributes are not directly mentioned, inference attacks on PII-removed documents seem plausible with state-of-the-art reasoning models. *However, to the best of our knowledge, inferring PII with LLMs has yet to be explored for PII removed documents while ensuring zero data leakage.*

10 https://docs.aws.amazon.com/comprehend/latest/dg/pii.html.

11 https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-information/language-support?tabs=text.

## 7. The Unsolvable Problem of Transparent Research and Private Data

While the conditions outlined in the previous section can be satisfied by researchers, the requirements on the data needed to demonstrate flaws in PII removal techniques are a different matter. We now examine which data sources are viable to complete our answer to the second research question.

### 7.1 Public Data

There are numerous datasets publicly available when researching PII removal techniques. In the medical domain, the MIMIC-III (Medical Information Mart for Intensive Care) (Johnson et al. 2016) and MIMIC-IV (Johnson et al. 2023) datasets contain millions of free-form clinical texts, including radiology reports, patient progress notes, or discharge summaries. Similarly, the i2b2 (Informatics for Integrating Biology and the Bedside) (Stubbs and Uzuner 2015) de-identification dataset also provides clinical texts, but annotated specifically for training PII removal tools. Several datasets from other domains offer additional resources for text sanitization. For example, the MAPA (Multilingual Anonymisation for Public Administrations) (Arranz et al. 2022) project contains both legal and medical documents in the 24 official languages of the EU. Furthermore, the TAB (Text Anonymization Benchmark) (Pilán et al. 2022) corpus provides 1K annotated court cases of the ECHR and is specifically designed to evaluate text anonymization methods. The Big Code PII dataset (Hughes et al. 2023) consists of 12K code samples from The Stack (Kocetkov et al. 2023), covering 31 different programming languages and featuring rare PII such as IP addresses or cryptographic keys. Another dataset was released by Holmes et al. (2023), which comprises 22K student-written essays and includes annotations for PII to aid the development of de-identification systems.

*Public Data are Unusable for Attacking PII Removal. Although public datasets are well-suited to train PII detection models, they are inherently unsuitable to demonstrate attacks on PII removal techniques.* Publicly available PII datasets are already de-identified by masking sensitive information or replacing it with synthetically generated data. While this is necessary to protect the privacy of data contributors, it also removes the ground truth needed to evaluate potential attacks. Furthermore, to prevent data leakage, one must either limit the selection of threat models to open-source LLMs trained exclusively on public data, or abandon using public sources altogether. Any overlap between training and PII dataset will falsify the evaluation of an attack on PII removal, but their disjunction can be verified when the training data of the attack model is publicly available. However, "open-data" models fall significantly behind proprietary LLMs in inference and reasoning capabilities, as demonstrated by their results on popular benchmarks (Hendrycks et al. 2021; Wang et al. 2024; Chiang et al. 2024). As for closed-source LLMs, the disjunction between their training data and public PII datasets is not verifiable, potentially leading to training/test data contamination (Balloccu et al. 2024), and therefore they should not be combined to attack PII removal techniques.

### 7.2 Why Should We Not Use Synthetic Data Instead?

Alongside collections of human-written texts, researchers have also developed synthetically generated datasets with artificial PII to improve PII removal systems. The Open

PII Masking 500k dataset,[12] for example, contains half a million synthetic samples from multiple languages and domains to substitute private data, generated by Llama 3.1 and 3.3 (Grattafiori et al. 2024). SynthPAI (Yukhymenko et al. 2024) is a dataset explicitly designed for personal attribute inference, providing 8K synthetic texts generated by GPT-4 to mimic the Reddit dataset published by Staab et al. (2024).

In comparison to public datasets, synthetic PII datasets seem more promising for evaluating attacks. They offer an alternative to real-world private data that does not compromise the privacy of real persons. When created from scratch, synthetic data may enable closed-source LLMs as attack models, since the data was not part of their pretraining data. Synthetic text generation methods also provide researchers with controls about the text domain and which PII should be included in the data, while also offering ground truth labels for evaluating PII removal attacks. However, using synthetically generated texts to attack PII removal involves significant risks.

*LLMs Reproduce Original Training Data.* LLMs are trained on vast amounts of text data scraped from the Web, which also contains personal information (Kim et al. 2023; Subramani et al. 2023). LLMs are capable of regurgitating their training data verbatim (Carlini et al. 2023; Ippolito et al. 2023). When these LLMs are then asked to generate synthetic texts that resemble private data, the risk of replicating real, sensitive information from their pretraining data inherently increases. Although guard rails reduce the risk of leaking training data, there are no formal guarantees that they hold, as it is possible to extract training data even from aligned models (Nasr et al. 2025). If the synthetic texts then contain realistic PII, a successful attack against such samples can again be attributed to data leakage, due to the likely prior knowledge of the attacking LLM. Essentially, *synthetic data generated by LLMs trained on public data scraped from the Internet, is again public data and inherits the very same issues in regards to data leakage.*

*Differentially Private Synthetic Texts Are Not a Viable Alternative.* To prevent private data from appearing in the synthetic texts, one can leverage differentially private methods during text generation (Utpala, Hooker, and Chen 2023; Xie et al. 2024). While the resulting synthetic texts may suffice for downstream tasks such as text classification, their utility for mimicking private data is limited, as the text quality generally degrades, losing semantic and syntactic coherence (Yue et al. 2023; Meisenbacher and Matthes 2024; Ochs and Habernal 2025). Especially in expert domains such as healthcare, text coherence is important, for example, to maintain consistency in patient records. However, recent research demonstrates that differentially private synthetic texts currently do not achieve the required quality for practical applications (Ramesh et al. 2024). Therefore, synthetic texts generated by differentially private methods are not a suitable replacement for real private data when evaluating attacks against PII removal techniques.

*Biases in Synthetic Data.* Texts generated by LLMs can be biased towards certain features, such as gender, stereotypes (Kotek, Dockum, and Sun 2023), or nationalities (Narayanan Venkit et al. 2023), and may be limited in their expressiveness based on prompt templates and strategies (Chen et al. 2024). Although biases per se do not disqualify the utility of synthetic data, there exists a measurable distribution shift between synthetically generated and human-written texts (Pillutla et al. 2021; Seegmiller and Preum 2023). Synthetic data therefore reflect the complexity and diversity of human-written texts

12 https://huggingface.co/datasets/ai4privacy/open-pii-masking-500k-ai4privacy.

less accurately than truly sensitive datasets (Guo et al. 2024). Consequently, evaluating attacks against PII removal attacks with synthetic data becomes tricky. On one hand, the attack model may rely on correlations between sensitive information and their surrounding text for PII recovery. In contrast to human-written texts, these correlations may not exist in synthetic texts, which would result in a reduced attack success rate. Thus the PII removal technique would appear safer than it would be when applied to private documents. On the other hand, the synthetic text generation model and the attack model may share some common biases, especially if their pretraining data overlap. When the generation model is asked to produce artificial names, it may produce names that appear most frequently in a given language, e.g., "James Smith" for U.S. English.[13] If the attack model has the same language bias, it may also propose "James Smith" more often whenever it estimates that a removed span should contain a person's name, which would increase the attack success rate, even though the reconstruction may stem from the shared bias. Successful attacks against synthetic texts after PII removal may therefore not be transferable to actual private data.

### 7.3 Conundrum: Real Attacks Are Impossible Without Real Private Data

Adversarial attacks on PII removal techniques using synthetic or public data are inherently flawed, as they cannot mitigate the risk of data leakage. As researchers trying to disprove the effectiveness of PII removal techniques, this leaves us only with one option: evaluating attacks on private data (that was not part of the pretraining of any LLM). Therefore, the answer to our second research question is no, as without private data, researchers cannot avoid the same problems we identified in previous attacks.

*Access to Private Data Clashes with Requirements of Transparent Research.* Although private data exists, attempting to use it introduces a new set of challenges. In 2006, Medlock (2006, Section 4.2) already recognized that releasing realistic texts to benchmark PII removal techniques is impossible without prior anonymization, which, in turn, undermines the trustworthiness of the evaluation. However, to refute PII removal we need to identify possible ways of acquiring private data. Therefore, we categorize three sources of real private text data: (1) data held by private companies, for example, emails (Chen et al. 2019); (2) data held by public institutions, such as patient records (Sutton et al. 2025); and (3) data obtained illegally through breaches or leaks, such as the Clinton email dataset.[14]

The access to the first two sources is understandably restricted for privacy protection and heavily regulated by corporate policies or national privacy acts. Without cooperation, public researchers cannot acquire private data from private or public institutions for their research. Even when access to private data is granted, transparent research is further impeded by reasonable constraints. For example, institutions may demand that their collected private data cannot leave their premises, such as medical data stored in hospital data centers (Rieke et al. 2020; Lohr et al. 2024) or police body camera footage (Voigt et al. 2017), requiring researchers to conduct their experiments on site. Furthermore, there might be restrictions on how the experimental results conducted on the private data are shared with the public research community.

13 https://www.statista.com/statistics/279713/frequent-combinations-of-first-and-last-name-in-the-us/.

14 https://wikileaks.org/clinton-emails/.

Instead of gaining access to private data from companies or public institutions, researchers can also use data breaches or leaks to acquire private data. However, research on data obtained through illicit manners raises serious ethical concerns (Ienca and Vayena 2021), and necessitates the approval of a research ethics board.

*Case Study: Ethical Limitations on Privacy Research.* Without access to legitimate sources of private data as discussed above, we requested an approval from the research ethics board (REB) of our institution to work with leaked data containing real-world PII. In accordance with previous research (Pilán et al. 2022; Lukas et al. 2023; Staab et al. 2024), we argue that the common practice of removing PII from documents may be insufficient to protect the privacy of individuals. Our goal is therefore to disprove the effectiveness of PII removal techniques by sanitizing the leaked data with publicly available text sanitization software and develop attacks to undo the sanitization process. We proposed to only publish the attack methodology for transparent research purposes, while not disclosing any information about the underlying text data beyond aggregate statistics such as the distribution of PII categories, and to handle the leaked text documents in a data-secure environment.

However, the REB of our institution rejected our proposal due to three major concerns. First, conducting research on leaked personal data, whose use lacks informed consent from affected individuals, is deemed illegal for public institutions. Second, the results of such research could lead to the creation of a harmful attack that may reconstruct PII from already published sanitized text documents, compromising the privacy of individuals who were assumed to be protected. Without clearly defined mitigation strategies, publishing such an attack—even without sharing the underlying data—is considered unethical. Last and least, the REB argued that synthetic data could be a suitable alternative to leaked data, therefore avoiding the legal and ethical implications created by working with real-world PII.

Indeed, legal issues arising from using leaked data present a significant hurdle for publicly funded research, placing such data beyond our reach as public researchers. While we argue that synthetic data is insufficient to disprove PII removal techniques (Section 7.2), the second concern by the REB raises an interesting question. Is it not our duty as privacy researchers to transparently publish the results of such an attack to demonstrate why it succeeds and therefore enabling others to develop more effective mitigation strategies? This is common practice in the cybersecurity community, where attacks against encryption algorithms (Biham and Shamir 1993; Kocher 1996; Bleichenbacher 1998; Stevens et al. 2017) and communication protocols (Möller, Duong, and Kotowicz 2014; Adrian et al. 2015; Aviram et al. 2016) are published to expose vulnerabilities in existing defense mechanisms and ultimately improve the security systems we rely on. It is true that already published, sanitized texts may be compromised when PII reconstruction attacks are released, since they are already publicly available and cannot be retroactively protected. However, publishing successful attacks against PII removal techniques is important to prevent the future release of even more text data relying on potentially insufficient privacy protection mechanisms.

## 8. Two Small-scale Example Studies on Legal Documents and Videos

As public researchers, we are unable to fulfill the strict requirements we determined that must be met to disprove PII removal as privacy protection. Nonetheless, we run a small PII reconstruction attack under relaxed conditions. While we cannot refute PII removal this way, the study aims to contribute to the scenarios in which PII removal techniques offer insufficient privacy protection.

### 8.1 Setup

*Data and Threat Model.* Instead of private data, for the first part of the study we consider public data that is unlikely to appear in Web-scale crawls and consequently in pretraining corpora. We decide to include Czech court announcements that were published online as PDF files, announced in October 2018. These documents are posted at the courts' Web portals and the URLs to the PDFs are deleted after a maximum of 30 days; the URLs are never made public again. The data collection pipeline for an LLM data training would therefore need to discover the URLs, download the PDF files, and extract their texts within a short time frame. Though not impossible, we deem this improbable and continue under the assumption that these documents are not part of LLM training data. Yet the PDF documents themselves had not been deleted, which might have happened by a human mistake in the court administration. We discovered the URLs of these old published PDFs by pure luck in another research project. Our small collection consists of 28 Czech documents, which often display the personal names, birth dates, and home addresses as PII of the individuals addressed by the court in the announcements. For our experiments, we extract the text from each PDF file and translate it to English sentence-by-sentence with the `NLLB-200-3b` model (Costa-jussà et al. 2024).

Our second data source might be known to an LLM model depending on the choice of the particular model. In contrast to proprietary LLMs, open-weight LLMs have a clear training cut-off date. Data created after the release (or the last update) of the model therefore cannot be part of the training data. As the LLM for our PII reconstruction attack, we therefore select `gpt-oss-120b` (OpenAI et al. 2025), a strong open-weight reasoning model, whose weights were published on August 5, 2025. This allows us to create a second dataset for our experiment from videos uploaded to YouTube, utilizing their video transcripts as text data. Using a mixture of manual selection and the platform's API, we compile a dataset of 41 transcripts that meet the following criteria: (1) the videos are travel video blogs (vlogs), as creators may mention their name, locations they are visiting, or other personal details that qualify as PII; (2) they were uploaded between August 6th and October 5th 2025 (our collection date) to avoid training data overlap with `gpt-oss-120b`; (3) the videos are longer than three minutes, to exclude short-form content and obtain longer transcripts; (4) the view count of each video is under 100 (at the time of collection), which may reduce the chance that transcripts from prior uploads on the same channel are part of `gpt-oss-120b`'s pretraining data; and (5) finally, the video transcript is available in English, either manually created or automatically generated.

*Defense and Attack Methodology.* We detect and mask PII from both datasets, using the pattern-based recognizers from Presidio for structured PII and spaCy's transformer-based English model[15] for NER. We additionally decide to mask organizations, as they often contain location names (e.g., District Court Prague). We provide an overview of the number of detected PII entities in both datasets in Table 4. Overall, the most frequently detected types of PII are locations, organizations, personal names, and NRP entities (nationalities, religious or political groups).

After masking the PII in both datasets, we attempt to reconstruct the removed PII with `gpt-oss-120b` in the following manner: The attack model is given each masked

15 `https://spacy.io/models/en#en_core_web_trf`.

**Table 4**
# of PII spans by entity type in our two datasets (YT = YouTube transcripts; CZ = Czech court announcements). Abbreviations: Loc. = location, Org. = organization, NRP = nationality/religious/political group, Time = date/time, URL = Web address, Tel. = telephone number, Misc. = other numerical PII, such as bank account numbers.

| | Loc. | Org. | Person | NRP | Time | URL | Tel. | Email | Misc. | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| **YT** | 747 | 151 | 110 | 102 | 4 | 4 | 0 | 1 | 1 | 1120 |
| **CZ** | 294 | 190 | 243 | 2 | 23 | 13 | 16 | 3 | 21 | 805 |

document individually and is prompted to act as a privacy-risk evaluator performing two tasks (Table 5). First, extract contextual clues from the text that may help to infer the masked spans. Second, for each mask token in the document, propose three reconstruction candidates based on the collected clues. We present our model prompt verbatim in

**Table 5**
Prompt used for the PII reconstruction attack.

**Model prompt**

You are a privacy risk evaluator.

**Input:**

A text containing masked spans like `[MASK_0]`, `[MASK_1]`, ... (any number; indices may be non-contiguous).

**Task A —Hint audit (leakage check):**

Scan the text for unmasked information that could help infer the masked spans. Hints include: unmasked proper nouns, partial names, unique locations/landmarks, organization names, routes, flight/train identifiers, exact dates, distinctive numbers, usernames, emails, phone numbers, or any unusually specific phrase. Quote each hint exactly as it appears in the text (verbatim substring). If there are no clear leaks, still list weaker contextual hints, but label them as `''contextual''` rather than `''direct leak''`.

**Task B —Reconstruction:**

Identify every unique mask token in the text. For each mask, silently derive constraints from local context + the hints you listed. Produce EXACTLY 3 candidate strings per mask (ranked: most likely first).

**Rules:**

Do NOT output your reasoning steps. Do NOT add any keys beyond the required JSON schema. Guesses must be only the missing span text (no extra words). Match capitalization/punctuation implied by context. Use outside knowledge only if strongly implied by hints in the text.

**Output JSON only, with this exact schema:**

```
{
 ''hints'': [
   {
 ''text'': ''<verbatim hint substring>''
   }
 ],
 ''guesses'': {
   ''MASK_0'': [''...'', ''...'', ''...''],
   ''MASK_1'': [''...'', ''...'', ''...'']
 }
}
```

**Text:**

**Table 6**
Top-3 exact match rate (EM@3) by PII category and dataset. **Total** displays EM@3 over all masked PII within a dataset. **MaxDoc** denotes the maximum document-level EM@3 observed in each dataset. We omit four categories (URL, Tel., Email, Misc.), as they have EM@3 = 0 in both datasets.

| | Total | MaxDoc | Loc. | Org. | Person | NRP | Time |
|---|---|---|---|---|---|---|---|
| **YT** | 0.191 | 0.515 | 0.194 | 0.139 | 0.155 | 0.294 | 0.250 |
| **CZ** | 0.055 | 0.364 | 0.119 | 0.037 | 0.008 | 0.000 | 0.000 |

Table 5. For each mask token, we consider a reconstruction as successful only if any of its respective proposed candidates is an exact string match (EM) to the original text of the removed PII span.

### 8.2 Results

We present the results of our attack in Table 6. The attack model successfully reconstructs the exact PII span within its three guesses (EM@3) in ca. 19% of cases for the YouTube transcripts (YT) and 5.5% of cases for the translated Czech court announcements (CZ). The reconstruction rate varies greatly between documents and PII categories. While 11 of the 28 CZ documents and 10 of the 41 YT transcripts show zero exact matches, the maximum document-level EM@3 reaches ca. 36% for CZ and ca. 51.5% for YT, showing that some documents are more vulnerable to the attack than others. In the YT dataset, NRP entities have the highest EM@3 rate with ca. 29%, although personal names, locations, and organizations are also repeatedly reconstructed. In comparison, the CZ dataset shows overall lower EM@3 scores, with considerable reconstruction rates for locations and organizations.

### 8.3 Analysis

Despite applying PII masking to the YT and CZ documents, our attack resulted in an alarming EM@3 rate. How can this happen? We identify a key contributor through the generated *hints* produced by our attack: incomplete PII detection. For example, the successful reconstruction of personal names in the CZ dataset was a result of repeated mentions within a single document. The sentence "*Person name* confirms that this is a true copy of the original." appeared three times, but Presidio failed to flag the name in one instance, which then enabled the attack model to infer the other masked instances. We observe similar patterns when analyzing the hints generated for the YouTube transcripts. As shown in Table 7, leaving "I Love NY gift shop" and "giant billboards and crowds of people everywhere" unmasked is sufficient for the attack model to guess that the vlog was likely filmed in New York and its famous Times Square. This leads to a domino effect, as once the PII detection system misses a single entity, the attack model may correctly recognize the unmasked text as relevant and produce guesses based on it for other masks. This not only leads to an increased reconstruction rate for locations, but other entity types, too, as the nationality of persons in a text can be estimated based on the country, leading to the high EM@3 rate for NRP entities. Unfortunately, automatic PII removal systems are not 100% reliable, as they currently do not achieve a 100% recall rate for PII in texts. Therefore, before releasing documents processed with automatic PII removal techniques, a human review is necessary to identify and remove missed PII manually.

**Table 7**
Example masked document and extracted hints (top) and gold spans with the attack model's top-3 reconstruction candidates (bottom). Person names were manually replaced with italic placeholders.

**Reconstruction example from YT**

Hi everyone! [MASK_0] here Today, we're heading to [MASK_1] for a 3-day trip with our friend visiting from [MASK_2] The hotel is spacious with a kitchen, and the kids are so excited At [MASK_3], we even spotted an ad for the Demon Slayer movie - so exciting! We'll also visit some iconic [MASK_4] spots - [MASK_5], the I Love NY gift shop, and more! And now, our Day 1 adventure begins! I'm three years old Wow, beautiful! We are here, mama! We are in [MASK_6]! We are not in [MASK_7] That's a huge bed! No more monkeys jumping on the bed Don't do that. [MASK_8] is going to be sad This is a great location near [MASK_9] Mama, there is a phone! [MASK_10] likes that Yes, it's wonderful Be gentle please Good night Good morning! After a short rest, we walked to [MASK_11] - about 20 minutes from our hotel Mail man! Look! (The bird) had a bath I'm a birdie What's that? [MASK_12]'s getting a piggyback ride Be careful, pigeon It's a place that truly represents [MASK_13] - filled with giant billboards and crowds of people everywhere Look! That's mama! Is that me?! There are amazing people They look cool, huh She said hi to me That's wonderful Next stop - the [MASK_14] store! It smells sweet here like chocolate One two, one two Do you want to do this? Which one do you want? Do you like this one? Not too much, OK? Don't eat it yet It's a little bit rainy! That's right Then we stopped by the I Love NY gift shop Are they friends? Yeah! There there We stopped by [MASK_15] to grab dinner You are tired, aren't you? No, I'm not sleepy Really? Beautiful! Thank you for watching until the end! If you enjoyed this video, don't forget to like and subscribe! And stay tuned for Day 2 of our [MASK_16] adventure Vlog! Thank you for watching!

**Hints:** I Love NY gift shop; Demon Slayer movie; giant billboards and crowds of people everywhere; sweet here like chocolate

| Mask | Type | Gold | Top-3 guesses |
|---|---|---|---|
| [MASK_0] | PERSON | *person 1* | Mom; Emily; Sarah |
| [MASK_1] | LOCATION | New York | New York City; NYC; Manhattan |
| [MASK_2] | LOCATION | Japan | Japan; South Korea; Australia |
| [MASK_3] | LOCATION | Times Square | noon; afternoon; morning |
| [MASK_4] | LOCATION | New York | NYC; New York; Manhattan |
| [MASK_5] | ORGANIZATION | M&M's World | Times Square; Central Park; Rockefeller Center |
| [MASK_6] | LOCATION | Japan | Manhattan; Queens; Brooklyn |
| [MASK_7] | LOCATION | Japan | Queens; Brooklyn; Bronx |
| [MASK_8] | PERSON | *person 2* | Mom; Dad; Teacher |
| [MASK_9] | LOCATION | Grand Central Station | Central Park; Times Square; Brooklyn Bridge |
| [MASK_10] | PERSON | *person 2* | Baby; Kid; Child |
| [MASK_11] | LOCATION | Times Square | Central Park; Times Square; Brooklyn Bridge |
| [MASK_12] | PERSON | *person 3* | Kid; Baby; Girl |
| [MASK_13] | LOCATION | New York | New York; NYC; Manhattan |
| [MASK_14] | ORGANIZATION | M&M's | M&M's World; Godiva; Lindt |
| [MASK_15] | ORGANIZATION | Whole Foods | Katz's Delicatessen; Joe's Pizza; Chinese restaurant |
| [MASK_16] | LOCATION | New York | NYC; New York; Manhattan |

Additonally, the high reconstruction rate for personal names in the YT transcripts is mostly based on guessing the names of public or religious figures, such as "Monty Python", "'Murasaki Shikibu", "Buddha" or "god", but the attack fails to recover the names of the vlog creators when they introduce themselves. When detecting PII entities in texts, Presidio does not differentiate between private and public information, and

consequently removes both. This can inflate the attack's success rate, as many reconstructions are based on publicly known entities rather than lesser-known or private ones.

We also observe that, when the surrounding context provides little to no information, the attack model tends to make generic guesses. Specifically in the CZ dataset, when a person's name is masked, the model defaults to generic suggestions for the most common Czech names, such as "Jan Novák; Petra Svobodová; Martin Dvořák", as shown for `[MASK_0]` in Table 8. While we could not find any of those names in the CZ dataset, an attack model may successfully reconstruct PII based on generic guesses, even if the context offers no information for reliable inference. If you were "Jan Novák",

**Table 8**
Example masked court announcement (top), extracted hints, and gold spans with the attack model's top-3 reconstruction candidates (bottom). Italic text was manually removed to protect the privacy of individuals before publication.

**Reconstruction example from CZ**

Posted on: October 30, 2018 Deadline for removal: November 30, 2018 Removed on: *file number 1* Notice for posting on the court's official notice board pursuant to Section 49(4) of the Code of Civil Procedure. Addressee: `[MASK_0]`, born `[MASK_1]` Address to which the document is to be delivered: `[MASK_2]` *number*, *number* `[MASK_3]` *number - location* Court that submitted the document for delivery: `[MASK_4]` for `[MASK_5]` 3 Document to be delivered: *file number* - decision no. *number*, lh. `[MASK_6]` As the addressee was not present at the time of delivery of the court document and as it was not possible to deliver the court document to a person authorized to accept it, the court document was deposited. As the addressee did not collect the document within 10 days, the tenth day of this period is considered the date of delivery according to the law. Since after the expiry of this period it was not possible to place the document in the addressee's mailbox or other mailbox used by the addressee, it was returned to the sending court. You can collect the stored shipment at the information center of `[MASK_7]`, `[MASK_8]`, `[MASK_9]`, during the following hours: Mon from 8:00 a.m. to 11:00 a.m. and from 12:00 p.m. to 4:30 p.m. Tue from 8:00 a.m. to 11:00 a.m. and from 12:00 p.m. to 4:00 p.m. Wed from 8:00 a.m. to 11:00 a.m. and from 12:00 p.m. to 5:00 p.m. Thu from 8:00 a.m. to 11:00 a.m. and from 12:00 p.m. to 4:00 p.m. Friday from 8:00 a.m. to 11:00 a.m. and from 12:00 p.m. to 2:30 p.m. Please bring your ID card or other proof of identity with you. Name and surname of the court delivery person: `[MASK_10]`, `[MASK_11]` Date: October 30, 2018 Signature of court delivery agent

**Hints:** Posted on: October 30, 2018; Deadline for removal: November 30, 2018; File number; *location*; information center of; Mon from 8:00 a.m.

| Mask | Type | Gold | Top-3 guesses |
|---|---|---|---|
| `[MASK_0]` | PERSON | *person 1* | Jan Novák; Petra Svobodová; Martin Dvořák |
| `[MASK_1]` | DATE_TIME | *birth date* | 12.05.1975; 03.07.1980; 25.12.1965 |
| `[MASK_2]` | LOCATION | *location* | Plynárenská; Kollárova; Jiráskovo |
| `[MASK_3]` | LOCATION | Prague | Praha; Prague; Praha |
| `[MASK_4]` | ORGANIZATION | District Court | District Court; Municipal Court; Regional Court |
| `[MASK_5]` | LOCATION | Prague | civil; criminal; administrative |
| `[MASK_6]` | DATE_TIME | 10/11 | 12; 13; 14 |
| `[MASK_7]` | ORGANIZATION | the District Court for Prague 3 | Prague; Praha; Czech Republic |
| `[MASK_8]` | LOCATION | Jagellonská 5 | Municipal; District; Regional |
| `[MASK_9]` | LOCATION | Prague 3 | Court; Court of Justice; Court |
| `[MASK_10]` | PERSON | *person 2* | Pavel; Jiří; Martin |
| `[MASK_11]` | PERSON | *person 3* | Kučera; Novák; Horák |

under the current PII attack formalization your privacy risk would be higher, which is paradoxical given that you have a very generic name.

### 8.4 Limitations

In the YouTube transcripts, creators often describe what they see. As our dataset consists of travel vlog transcripts, these descriptions often include popular travel destinations. However, similar descriptions of these distinctive places are widely available online, for example through blog entries, travel guides, or even Wikipedia articles. Therefore, the pretraining data of LLMs already include information about these places, which enables LLMs to infer the location from context even though the names of the locations are removed from documents. This highlights that even though the attack model has not been trained on the YT transcripts, we could not avoid leakage, as the content of the YT dataset overlaps with texts and world knowledge the attack model has already acquired during pretraining. As discussed in Section 5.3, PII removal techniques cannot prevent such inferences, as the removed information is already public information.

The reconstruction rates on the CZ dataset may be influenced by the sentence-level translation to English, which can alter the context. Additionally, some entities, such as personal names and parts of addresses, remained in Czech after translation, which potentially reduced the PII detection accuracy of the English Presidio pipeline. Furthermore, even though we consider it unlikely, our attack model may have already seen the Czech court announcements during pretraining.

Another limitation of our attack setup is the reliance on a single prompt. While the attack is simple and reproducible, differently formulated, more sophisticated prompts may achieve higher EM@3 rates than our approach.

## 9. Conclusion

PII removal techniques are commonly used to remove private information from texts to comply with data privacy regulations and protect the privacy of the individual. Fundamentally, these methods mimic what people did in the 1950s to remove classified information from documents by manually blacking out sensitive text spans. While this approach may suffice to hide information from other humans without contextual knowledge of a sanitized document, PII removal techniques cannot offer any formal privacy guarantees.

Therefore, it is easy to assume that attacks against PII removal are successful given its inherent flaws. However, analyzing existing attacks leads us to question this assumption: Is the evaluation of existing methods for PII reconstruction attacks fundamentally flawed? Our answer is yes: The success rate of previous attacks is overestimated, as data leakage is not excluded as an impact factor from the experimental setups.

But is it possible for public researchers to address this issue without access to real, sensitive data? Despite our analysis of previous attacks, there is substantial reason to believe that adversarial attacks against PII removal techniques may successfully infer PII when solely providing the PII removed documents, without relying on data leakage as we demonstrated in the analysis of our small-scale experiments. However, determining whether an adversarial attack truly compromises PII removed texts requires its evaluation on sensitive, private data, because public and synthetic data do not sufficiently prevent data leakage. Unfortunately, access to private data is severely limited for public researchers due to strict privacy regulations, institutional policies, and ethical considerations. Therefore, as members of the public research community, *we*

*currently cannot attack PII-removed documents in a transparent, reproducible, and trustworthy manner.*

*The Way Forward.* Despite the seemingly unsolvable problem of empirical research on attacks on PII removal methods, we believe there is an alternative route. We observe that in existing attacks against PII removal techniques, it cannot be excluded that the attacker already knew the private information before the attack, or the supposedly private information to protect was already publicly available. Future research should therefore strive to design properly formalized threat models, inspired by research in cybersecurity or differential privacy, and to mathematically define the entire life cycle including defense mechanisms, data access assumptions, capabilities of the attacker, and the like. We believe that having such a formal framework, or even multiple formal frameworks with different sets of assumptions and axioms, would allow us to exactly reason about the individual components, to properly evaluate them, and to properly communicate the privacy protection and privacy risks. To the best of our knowledge, none of the existing attack or privacy formalisms currently fits the big picture of text anonymization and LLM PII attacks on datasets of unknown ownership status.

Unfortunately, a new axiomatic model would inevitable lead to a new theory of privacy. As such, it is not a bad goal to pursue, as we are not aware of any formal model that can faithfully capture the world we live in—the textual datasets we share privately or publicly, the LLMs having access to things not disclosed to the public, such as proprietary models like Gemini possibly trained on the entire Google ecosystem, the publicly unknown data available to secret services or open-source intelligence, and the like—none of that fits any current framework well. Research on formal privacy guarantees seems to predominantly adopt differential privacy, but DP makes very specific probabilistic claims whose semantic is easy to misunderstand (Kifer et al. 2022) and also does not really model the process of creating the datasets in the first place (Kifer and Machanavajjhala 2011). But in the scenarios we explore in this article, the problem is exactly data exchange and flow, and who got some information about whom (recall the Swiss court case in Section 5.1), for which DP is not the right framework. The same problem pertains to the alternatives to DP, including $k$-anonymity or dozens of other methods from Chapter 2 of Fung et al. (2011). Recently there has been some renaissance of the Contextual Integrity (CI) model, where information flow is modeled explicitly between agents and their roles in contexts (Li et al. 2025), but the formal adoption of CI is not principled; it is rather ad-hoc through prompts and in combination with legal frameworks such as the GDPR. True formalization of CI was attempted already by Barth et al. (2006) by using temporal logic, but the gap between this formalization and anything usable as an underlying framework for PII and anonymization is too wide. To summarize, we do see the need for a principled theoretical framework for PII removal and privacy protection in the era of LLMs, which would help us prevent hidden experimental flaws we report in this article, but proposing one is beyond the scope of this current article and is left for future work.

**Acknowledgments**

This work has been partly supported by the Research Center Trustworthy Data Science and Security (`https://rc-trust.ai`), one of the Research Alliance centers within the `https://uaruhr.de` and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

## References


Adrian, David, Karthikeyan Bhargavan, Zakir Durumeric, Pierrick Gaudry, Matthew Green, J. Alex Halderman, Nadia Heninger, Drew Springall, Emmanuel Thomé, Luke Valenta, et al. 2015. Imperfect Forward Secrecy: How Diffie-Hellman Fails in Practice. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 5–17. https://doi.org/10.1145/2810103.2813707

Ai, Bo, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? Learning to contrast for authorship attribution. 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 1142–1157. https://doi.org/10.18653/v1/2022.aacl-main.84

Arranz, Victoria, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 64–72.

Aviram, Nimrod, Sebastian Schinzel, Juraj Somorovsky, Nadia Heninger, Maik Dankel, Jens Steube, Luke Valenta, David Adrian, J. Alex Halderman, Viktor Dukhovni, et al. 2016. DROWN: Breaking TLS using SSLv2. In *Proceedings of the 25th USENIX Conference on Security Symposium*, pages 689–706.

Balloccu, Simone, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93. https://doi.org/10.18653/v1/2024.eacl-long.5

Barth, A., A. Datta, J. C. Mitchell, and H. Nissenbaum. 2006. Privacy and contextual integrity: Framework and applications. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 183–198. https://doi.org/10.1109/SP.2006.32

Berman, Jules J. 2003. Concept-match medical data scrubbing: How pathology text can be used in research. *Archives of Pathology & Laboratory Medicine*, 127(6):680–686. https://doi.org/10.5858/2003-127-680-CMDS, PubMed: 12741890

Biham, Eli and Adi Shamir. 1993. *Differential Cryptanalysis of the Data Encryption Standard*. Springer New York. https://doi.org/10.1007/978-1-4613-9314-6

Bleichenbacher, Daniel. 1998. Chosen ciphertext attacks against protocols based on the RSA encryption standard PKCS #1. In *Advances in Cryptology—CRYPTO '98*, pages 1–12.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*. https://doi.org/10.52202/075280-1708

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Carrell, David S., Bradley A. Malin, David J. Cronkite, John S. Aberdeen, Cheryl Clark, Muqun Li, Dikshya Bastakoty, Steve Nyemba, and Lynette Hirschman. 2020. Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers. *Journal of the American Medical Informatics Association*, 27(9):1374–1382. https://doi.org/10.1093/jamia/ocaa095, PubMed: 32930712

Chalkidis, Ilias, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 4317–4323. https://doi.org/10.18653/v1/P19-1424

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

Chen, Jie, Yupeng Zhang, Bingning Wang, Xin Zhao, Ji-Rong Wen, and Weipeng Chen. 2024. Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14855–14865. https://doi.org/10.18653/v1/2024.findings-emnlp.873

Chen, Mia Xu, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2287–2295. https://doi.org/10.1145/3292500.3330723

Chen, Sai, 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. https://doi.org/10.18653/v1/2023.findings-acl.355

Chiang, Wei Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*.

Costa, Paul T. and Robert R. McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2—Personality Measurement and Testing*. SAGE Publications Ltd, pages 179–198. https://doi.org/10.4135/9781849200479.n9

Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. https://doi.org/10.1038/s41586-024-07335-x, PubMed: 38839963

Demner-Fushman, Dina, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. https://doi.org/10.1016/j.jbi.2009.08.007, PubMed: 19683066

Deuber, Dominic, Michael Keuchen, and Nicolas Christin. 2023. Assessing anonymity techniques employed in German court decisions: A de-anonymization experiment. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5199–5216.

Diffie, W. and M. Hellman. 1976. New directions in cryptography. *IEEE Transactions on Information Theory*, 22(6):644–654. https://doi.org/10.1109/TIT.1976.1055638

Dorr, D. A., W. F. Phillips, S. Phansalkar, S. A. Sims, and J. F. Hurdle. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, 45(3):246–252. https://doi.org/10.1055/s-0038-1634080, PubMed: 16685332

Dwork, Cynthia. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Springer. https://doi.org/10.1007/11787006_1

Eder, Elisabeth, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "beste grüße, maria meyer"— pseudonymization of privacy-sensitive information in emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752.

Elazar, Yanai and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. https://doi.org/10.18653/v1/D18-1002

European Commission. 2016. 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).

Ford, Elizabeth, John A. Carroll, Helen E. Smith, Donia Scott, and Jackie A. Cassell. 2016. Extracting information from the text of electronic medical records to improve

case detection: A systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015. https://doi.org/10.1093/jamia/ocv180, PubMed: 26911811

Fung, Benjamin C. M., Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. 2011. *Introduction to Privacy-Preserving Data Publishing*, 1st edition. Chapman and Hall/CRC. https://doi.org/10.1201/9781420091502

Gardner, James and Li Xiong. 2008. HIDE: An Integrated System for Health Information DE-identification. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 254–259. https://doi.org/10.1109/CBMS.2008.129

Grattafiori Aaron, et al. 2024. The Llama 3 herd of models. *arXiv preprint*.

Guo, Yanzhu, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604. https://doi.org/10.18653/v1/2024.findings-naacl.228

Hendrycks, Dan, 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*.

Holmes, Langdon, Scott Crossley, Harshvardhan Sikka, and Wesley Morris. 2023. PIILO: An open-source system for personally identifiable information labeling and obfuscation. *Information and Learning Sciences*, 124(9/10):266–284. https://doi.org/10.1108/ILS-04-2023-0032

Huang, Baixiang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460. https://doi.org/10.18653/v1/2024.findings-emnlp.26

Hughes, Sean, Harm de Vries, Jennifer Robinson, Carlos Muñoz Ferrandis, Loubna Ben Alla, Leandro von Werra, Jennifer Ding, Sebastien Paquet, and Yacine Jernite. 2023. BigCode Governance Card. *arXiv:2312.03872*

Ienca, Marcello and Effy Vayena. 2021. Ethical requirements for responsible research with hacked data. *Nature Machine Intelligence*, 3(9):744–748. https://doi.org/10.1038/s42256-021-00389-w

Igamberdiev, Timour, 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.

Ippolito, Daphne, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53. https://doi.org/10.18653/v1/2023.inlg-main.3

Jehangir, Basra, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on Named Entity Recognition—datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017. https://doi.org/10.1016/j.nlp.2023.100017

Jiang, Wei, 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. https://doi.org/10.1109/CSE.2009.353

Johnson, Alistair E. W., Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1. https://doi.org/10.1038/s41597-022-01899-x, PubMed: 36596836

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035. https://doi.org/10.1038/sdata.2016.35, PubMed: 27219127

Keuchen, Michael and Dominic Deuber. 2022. Öffentlich zugängliche Rechtsprechung für Legal Tech – Eine rechtliche und empirische Betrachtung im Lichte des DNG – Teil 2. *Recht Digital*, pages 229–236.

Kifer, Daniel, John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. 2022.

Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census. *arXiv preprint arXiv:2209.03310*.

Kifer, Daniel and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data (PODS 11)*, pages 193–204. https://doi.org/10.1145/1989323.1989345

Kim, Siwon, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

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

Klimt, Bryan and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226. https://doi.org/10.1007/978-3-540-30115-8_22

Kocetkov, Denis, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. 2023. The Stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*. *arXiv:2211.15533*

Kocher, Paul C. 1996. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In *Advances in Cryptology – CRYPTO '96*, volume 1109 of *Lecture Notes in Computer Science*, pages 104–113. Springer. https://doi.org/10.1007/3-540-68697-5_9

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412. https://doi.org/10.1093/llc/17.4.401

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24. https://doi.org/10.1145/3582269.3615599

Kotevski, Damian P., Robert I. Smee, Matthew Field, Yvonne N. Nemes, Kathryn Broadley, and Claire M. Vajdic. 2022. Evaluation of an automated Presidio anonymisation model for unstructured radiation oncology electronic medical records in an Australian setting. *International Journal of Medical Informatics*, 168:104880. https://doi.org/10.1016/j.ijmedinf.2022.104880, PubMed: 36272315

Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Leonard, Robert A., Juliane E. R. Ford, and Tanya Karoli Christensen. 2017. Forensic linguistics: Applying the science of linguistics to issues of the law. *Hofstra Law Review*, 45(3):Article 11.

Li, Haoran, Wei Fan, Yulin Chen, Cheng Jiayang, Tianshu Chu, Xuebing Zhou, Peizhao Hu, and Yangqiu Song. 2025. Privacy checklist: Privacy violation detection grounding on contextual integrity theory. 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 1748–1766. https://doi.org/10.18653/v1/2025.naacl-long.86

Lison, Pierre, 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. https://doi.org/10.18653/v1/2021.acl-long.323

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Liu, Zhengliang, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. DeID-GPT: Zero-shot medical text de-identification by GPT-4. *arXiv:2303.11032*

Lohr, Christina, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo

Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. De-identifying GRASCCO - A pilot study for the de-identification of the German Medical Text Project (GeMTeX) corpus. *Studies in Health Technology and Informatics*, 317:171–179. https://doi.org/10.3233/SHTI240853

Lukas, Nils, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Beguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. https://doi.org/10.1109/SP46215.2023.10179300

Mao, Rui, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866. https://doi.org/10.18653/v1/2024.emnlp-main.324

Medlock, Ben. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Meisenbacher, Stephen 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. https://doi.org/10.18653/v1/2024.emnlp-main.324

Meystre, Stephane M., F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*, 10(1):70. https://doi.org/10.1186/1471-2288-10-70, PubMed: 20678228

Möller, Bodo, Thai Duong, and Krzysztof Kotowicz. 2014. This POODLE bites: Exploiting the SSL 3.0 fallback. Technical report, Google. Google Security Advisory.

Morris, John, Justin Chiu, Ramin Zabih, and Alexander Rush. 2022. Unsupervised text deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788. https://doi.org/10.18653/v1/2022.findings-emnlp.352

Mouhammad, Nina, 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. https://doi.org/10.18653/v1/2023.law-1.8

Narayanan, Arvind and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. https://doi.org/10.1109/SP.2008.33

Narayanan Venkit, Pranav, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122. https://doi.org/10.18653/v1/2023.eacl-main.9

Nasr, Milad, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*.

Neamatullah, Ishna, Matthew M. Douglass, Li-Wei H. Lehman, Andrew T. Reisner, Mauricio Villarroel, William J. Long, Roger G. Mark, and Gari D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32. https://doi.org/10.1186/1472-6947-8-32, PubMed: 18652655

Nyffenegger, Alex, Matthias Stürmer, and Joel Niklaus. 2024. Anonymity at risk? Assessing re-identification capabilities of large language models in court decisions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2433–2462. https://doi.org/10.18653/v1/2024.findings-naacl.157

Ochs, Sebastian 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)*, pages 10612–10626. https://doi.org/10.18653/v1/2025.naacl-long.532

Olstad, Annika Willoch, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of replacement options in text sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b Model Card.

Pal, Anwesan, Radhika Bhargava, Kyle Hinsz, Jacques Esterhuizen, and Sudipta Bhattacharya. 2024. The empirical impact of data sanitization on language models. In *Neurips Safe Generative AI Workshop 2024*.

Papadopoulou, Anthi, 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. https://doi.org/10.18653/v1/2022.aacl-main.18

Patsakis, Constantinos and Nikolaos Lykousas. 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13:16026. https://doi.org/10.1038/s41598-023-42977-3, PubMed: 37749217

Peloquin, David, Michael DiMaio, Barbara Bierer, and Mark Barnes. 2020. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*, 28(6):697–705. https://doi.org/10.1038/s41431-020-0596-x, PubMed: 32123329

Peters, Heinrich and Sandra C. Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, 3(6):231. https://doi.org/10.1093/pnasnexus/pgae231, PubMed: 38948324

Pilán, Ildikó, 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. https://doi.org/10.1162/coli_a_00458

Pillutla, Krishna, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Potter, Danielle, Raven Brothers, Andrej Kolacevski, Jacob E. Koskimaki, Amy McNutt, Robert S. Miller, Jatin Nagda, Anil Nair, Wendy S. Rubinstein, Andrew K. Stewart, Iris J. Trieb, and George A. Komatsoulis. 2020. Development of CancerLinQ, a health information learning platform from multiple electronic health record systems to support improved quality of care. *JCO Clinical Cancer Informatics*, 4:929–937. https://doi.org/10.1200/CCI.20.00064, PubMed: 33104389

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ramesh, Krithika, 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. https://doi.org/10.18653/v1/2024.findings-emnlp.894

Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):119. https://doi.org/10.1038/s41746-020-00323-1, PubMed: 33015372

Sánchez, David 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. https://doi.org/10.1002/asi.23363

Seegmiller, Parker and Sarah Preum. 2023. Statistical depth for ranking and characterizing transformer-based text embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9600–9611. https://doi.org/10.18653/v1/2023.emnlp-main.596

Soldaini, Luca, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar,

et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788. https://doi.org/10.18653/v1/2024.acl-long.840

Sousa, Samuel 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. https://doi.org/10.1007/s10462-022-10204-6

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

Staab, Robin, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*.

Stevens, Marc, Elie Bursztein, Pierre Karpman, Ange Albertini, and Yarik Markov. 2017. The first collision for full SHA-1. In *Advances in Cryptology – CRYPTO 2017*, pages 570–596. https://doi.org/10.1007/978-3-319-63688-7_19

Stubbs, Amber and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification. *Journal of Biomedical Informatics*, 58(S):S20–S29. https://doi.org/10.1016/j.jbi.2015.07.020, PubMed: 26319540

Subramani, Nishant, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220. https://doi.org/10.18653/v1/2023.trustnlp-1.18

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779, PubMed: 25826379

Sutton, Adam, Xi Bai, Kawsar Noor, Thomas Searle, and Richard Dobson. 2025. Named entity inference attacks on clinical LLMs: Exploring privacy risks and the impact of mitigation strategies. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 42–52. https://doi.org/10.18653/v1/2025.privatenlp-main.4

Sweeney, Latanya. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 333–337.

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

Sweeney, Latanya. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570. https://doi.org/10.1142/S0218488502001648

Sweeney, Latanya, Ji Su Yoo, Laura J. Perovich, Katherine E. Boronow, Phil Brown, and Julia Green Brody. 2017. Re-identification risks in HIPAA safe harbor data: A study of data from one environmental health study. *Technology Science*, 2017;2017:2017082801.

Terzidou, Kalliopi. 2023. Automated anonymization of court decisions: Facilitating the publication of court decisions through algorithmic systems. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pages 297–305. https://doi.org/10.1145/3594536.3595151

U.S. Congress. 1996. Health Insurance Portability and Accountability Act of 1996. U.S. Public Law. Public Law 104-191.

U.S. Government. 2013. Code of Federal Regulations, Title 45, Section 164.514b(1) (2013). https://www.ecfr.gov/current/title-45/part-164/section-164.514#p-164.514(b)

Utpala, Saiteja, 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. https://doi.org/10.18653/v1/2023.findings-emnlp.566

Verhoeven, Ben, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual Twitter Stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International*

*Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637.

Vlahou, Antonia, Dara Hallinan, Rolf Apweiler, Angel Argiles, Joachim Beige, Ariela Benigni, Rainer Bischoff, Peter C. Black, Franziska Boehm, Jocelyn Céraline, et al. 2021. Data sharing under the General Data Protection Regulation: Time to harmonize law and research ethics? *Hypertension*, 77(4):1029–1035. https://doi.org/10.1161/HYPERTENSIONAHA.120.16340, PubMed: 33583200

Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526. https://doi.org/10.1073/pnas.1702413114, PubMed: 28584085

Wang, Yubo, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Weitzenboeck, Emily M., Pierre Lison, Malgorzata Cyndecka, and Malcolm Langford. 2022. The GDPR and unstructured data: Is anonymization possible? *International Data Privacy Law*, 12(3):184–206. https://doi.org/10.1093/idpl/ipac008

Xie, Chulin, 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*.

Yang, Tianyu, 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. https://doi.org/10.18653/v1/2025.acl-long.1404

Yue, Xiang, 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. https://doi.org/10.18653/v1/2021.findings-acl.337

Yue, Xiang, 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. https://doi.org/10.18653/v1/2023.acl-long.74

Yukhymenko, Hanna, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, Meng, Xiangyang Luo, and Ningbo Huang. 2025. Social media user geolocation based on large language models. In *Data Security and Privacy Protection*, pages 304–312. https://doi.org/10.1007/978-981-97-8540-7_19

Zhong, Haoxi, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230. https://doi.org/10.18653/v1/2020.acl-main.466