A Study of Privacy-preserving Language Modeling Approaches

Pritilata Saha 10 and Abhirup Sinha 10 1

Abstract: Recent developments in language modeling have increased their use in various applications and domains. Language models, often trained on sensitive data, can memorize and disclose this information during privacy attacks, raising concerns about protecting individuals' privacy rights. Preserving privacy in language models has become a crucial area of research, as privacy is one of the fundamental human rights. Despite its significance, understanding of how much privacy risk these language models possess and how it can be mitigated is still limited. This research addresses this by providing a comprehensive study of the privacy-preserving language modeling approaches. This study gives an in-depth overview of these approaches, highlights their strengths, and investigates their limitations. The outcomes of this study contribute to the ongoing research on privacy-preserving language modeling, providing valuable insights and outlining future research directions.

Keywords: Privacy-preserving Language Modeling, Differential Privacy, Knowledge Unlearning, Private Representation Learning, Large Language Models

1 Introduction

Recent work has shown that large language models (LLMs) tend to memorize information from training data containing personally identifiable information (PII), and adversaries can extract this information later [Ca21; He22; Na23]. However, everyone has the Right to be Forgotten under the General Data Protection Regulation (GDPR) law [GNG21]. Though current state-of-the-art LLMs perform well in generating human-like text, recent research has revealed the vulnerability of these models to preserve privacy [Ki24; Na23].

Due to the legal obligations and ethical responsibilities associated with using language models, privacy-preserving practices are important. Recent privacy-preserving language models employ various approaches, such as Differential Privacy [Ab16; An21; Sh21; Sh22; WGX22], Knowledge Unlearning [Ja22; YXL23], Data Preprocessing [KWR22; Le21; Li21b], Private Representation Learning [Zh22; Zh23], Federated Learning [Li20; Li21a; Mc17; Re20] to mitigate inherent privacy risks. All the current privacy-preserving methodologies safeguard privacy against different types of attacks, but no single approach can protect against all kinds of privacy attacks alone.

This study reviews the core concepts underlying the most used approaches for mitigating privacy risks, along with their benefits and challenges. This study is divided into four independent research directions.

Department of Computer Science, Paderborn University, Germany, psaha@mail.upb.de, https://orcid.org/0000-0002-7776-1620; abhirup@mail.upb.de, https://orcid.org/0000-0002-6927-5526

- **Differential Privacy (DP)-based approaches:** Differential Privacy-based approaches apply existing DP algorithms, sometimes with some updates, e.g., DP-SGD, to protect privacy information from leaking from training data.
- Private Representation Learning-based approaches: These approaches can preserve the privacy of the representations to avoid text reconstruction attacks.
- Knowledge Unlearning-based approaches: These approaches use algorithms, e.g., negative log-likelihood, to forget specific sequences of tokens from training data.
- Data Preprocessing approaches: These approaches focus on detecting and removing sensitive information from training data to mitigate privacy risks.

This research discusses each approach's merits and shortcomings and suggests future research directions depending on them.

The remainder of this paper is structured as follows. First, it gives an overview of what privacy-preserving language modeling is (Section 2). Then, discuss the four widely used research methods and approaches (Section 3), followed by the key findings (Section 4) where findings of this work are added. Finally, the limitations and future works (Section 5) are followed by the conclusion (Section 6).

2 Why Preserve Privacy in Language Models?

LLMs have recently become an integral part of our lives, and these models can be used for different tasks, e.g., text generation and language translation. These language models have been widely used in chatbots, AI assistance systems, etc. But when it comes to privacy, most of the models struggle to preserve privacy.

2.1 Legal Requirements

According to the Universal Declaration of Human Rights [JSM88], privacy is one of the fundamental human rights, and individuals should not face unwarranted intrusion into their privacy. The General Data Protection Regulation (GDPR) was adopted in 2018 [VV17], and it gives individuals control over their personal data. Every individual has the right to limit the use of their personal information, and the Right to be Forgotten is part of the General Data Protection Regulation (GDPR) law [GNG21; Ma13]. Therefore, addressing privacy concerns in LMs is not only a social responsibility, but there are also legal requirements to ensure compliance with established human rights and data protection laws.

2.2 Privacy Risks

Most current state-of-the-art models can not ensure the privacy of personally identifiable information. Language models are trained with highly sensitive data that contains personally identifiable information (PII), such as names, email addresses, and phone numbers. These models tend to memorize the knowledge from initial pretraining, and recent works have shown that adversaries can extract training data from language models [He22].

Preserving privacy in language models involves implementing different approaches to mitigate privacy risks. Current privacy-preserving methodologies utilize techniques like Differential Privacy (DP), Knowledge Unlearning, Private Representation Learning, etc. These methods aim to protect the disclosure of sensitive information in the training data under privacy attacks. Section 3 discusses some of such available methods in privacy-preserving NLP.

3 Research Methods and Approaches

3.1 Differential Privacy-based approaches

Differential privacy [DR+14; Dw08] is one of the most used approaches to preserving privacy in data. Differential Privacy-based approaches [Ab16; An21; Sh21; Sh22; WGX22] in language modeling aim to protect sensitive information by employing an $(\epsilon; \delta) - DP$ algorithm. An $(\epsilon; \delta) - DP$ algorithm's objective is to limit its output's use to probabilistically determine the presence of a single record in the dataset by a factor of e^{ϵ} . DP-based approaches for preserving privacy in language models use DP-SGD optimization proposed by Abadi et al. [Ab16]. The fundamental concept behind DP-SGD involves clipping each example's gradients and adding Gaussian noise $z \sim \mathcal{N}(0, C^2\sigma^2I)$ during training. The new gradient is calculated as,

$$\tilde{gL_t} = \frac{1}{L} \left(\sum_{x_i} g(x_i) + z_t \right) \tag{1}$$

The adaptive noise technique, suggested by Wu et al. [WGX22], dynamically modifies the noise magnitude based on the privacy probability of an item within the DP-SGD process. Ultimately, a gradient optimization algorithm incorporating adaptive noise is presented in equation 3.

$$\gamma_B = \frac{\sum_{i=1}^L \rho(s_i)}{L} \tag{2}$$

$$z_{\text{Badp}} = \gamma_B \cdot \mathcal{N}(0, C^2 \sigma^2 I^2) \tag{3}$$

Here, γ_B denotes the privacy weights, which is the privacy probability averaged over batch $B = \{s_1, s_2, \dots, s_L\}$ of size L as shown in equation 2. $\mathcal{N}(0, C^2\sigma^2I_2)$ is the Gaussian noise of B, where σ is a noise multiplier, and C is the clipping norm.

While the DP approaches mentioned earlier focus on the overall data, the Selective Differential Privacy approach proposed by Shi et al. [Sh21; Sh22] focuses on the privacy-sensitive portion of the data only to provide a privacy guarantee. This approach uses a policy function F to distinguish between private and non-private attributes inside one data point and, in this way, protect the privacy of the sensitive parts while maintaining model utility. In this approach, the policy function is first used to get the privacy bit matrix P = F(D). Here, for a record $r \in \tau$, the policy function $F: \tau \mapsto \{0,1\}^{n_r}$ finds sensitive attributes by assigning $F(r)_i = 0$ and non-sensitive attributes assigning $F(r)_i = 1$. Here, n_r is the number of attributes in r. Policy function can be defined according to the application. After that, the Selective-DPSGD algorithm is used to train the model. Within this algorithm, the regular stochastic gradient descent (SGD) algorithm is used for non-private updates, and DP-SGD is used for private updates. The private and non-private updates are determined by the privacy bit matrix P.

Benefits: DP can provide strong privacy over the entire dataset, so it can be used to preserve the privacy of PII. Though some research suggested the performance degradation issue, the most recent approaches, e.g., Adaptive DP and Selective DP, are designed to maintain model utility to some extent. So, these new approaches can improve the performance of differentially private models.

Challenges: The DP-based approaches are computationally costly. Additionally, as stated by the [WGX22], implementing the Selective Differential Privacy method requires knowledge of which items in the dataset contain private information. This requirement becomes prohibitively expensive, particularly for large-scale datasets. DP can provide privacy when there is a clear privacy border [Br20] but can not provide privacy in some real-world scenarios like inference attack [SR20; Zh22]. Some studies suggest that DP can cause severe degradation of the model's performance [Ja22].

3.2 Private Representation Learning

To avoid sharing sensitive information, instead of sharing plain text data directly, people can share representations[Li21a]. However, these representations can still be transformed into the original text under a text reconstruction attack [Pa20; SR20]. Private representation learning approaches can preserve inference privacy as they operate at the level of latent vector representations rather than modifying the texts themselves. These approaches can be applied to each word representation, making them indistinguishable or, at the targeted

token, breaking the one-to-one relation between token representations and raw words and hiding private words.

The TextFusion approach by Zhou et al. [Zh22] does not change the basic architecture of the pre-trained language model but introduces a fusion predictor to determine which tokens should be fused. Then, the suitable token representations are fused in the privacy-preserving layer. This approach also tries to mislead the attacker by making the token representation more predictable to a different word with the closest Euclidean distance. The main goal of these fused representations is to make it challenging for privacy attackers to revert the token representations back to raw words, and the misleading training misleads the attacker for both fused and unfused representations.

Another framework, named TextObfuscator, proposed by Zhou et al. [Zh23], used the obfuscation technique to preserve privacy. The task is done in two steps; in the first step, each word is assigned to a corresponding prototype depending on semantic and task-related roles. In the second step for private representation training, the goal is to get a word representation and then make the representation close to its prototype by using the $\mathcal{L}_{\text{close}}$ in equation 4. Here, p_{x_i} is the prototype of x_i and the word representation is $H = \{h_i\}_{i=1}^n$.

$$\mathcal{L}_{\text{close}} = \frac{1}{2} \sum_{i=1}^{n} ||h_i - p_{x_i}||_2^2$$
 (4)

Also, the distance between different prototypes is maintained to avoid collapse during training using prototype distance loss in equation 5. Here, n_p is the number of prototypes.

$$\mathcal{L}_{\text{away}} = \frac{2}{n_p(n_p - 1)} \sum_{i=1}^{n_p} \sum_{j=i+1}^{n_p} ||p_i - p_j||_2^2$$
 (5)

Benefits: According to some studies, DP-based approaches do not fully protect privacy under the text reconstruction attack during inference [SR20; Zh22]. Privatizing token representations during inference can overcome this problem. Also, these approaches can preserve privacy without substantially sacrificing performance.

Challenges: The TextFusion approach relies on getting the predictions for confident representations on the early layer for token classification, which will not be suitable for tasks requiring a large-scale fusion ratio. Also, for token classification, the fusion rate has a greater impact on the task performance. The TextObfuscator requires more training steps compared with fine-tuning, resulting in increased computational cost. Also, the approach was designed for inference privacy and was not tested against other privacy attributes.

3.3 Knowledge Unlearning approaches

Machine unlearning is an approach to overcome data privacy issues in machine learning. It has been mostly used for preserving privacy image classification models. However, this unlearning approach has recently been adapted for forgetting targeted data in language models [Ja22; YXL23]. The unlearning task is more challenging for language models compared with classification tasks due to the larger output space(\sim a few image classes vs. a sequence of tokens that can each be classified into $V \in \mathbb{R}^{-50,000}$)

The knowledge unlearning approach for mitigating privacy by Jang et al. [Ja22] proposes a method to unlearn a specific sequence of tokens for language models. The proposed approach negates the original training objective by training to maximize the negative log-likelihood loss of the token sequence. By going in the opposite direction of the traditional gradient descent, it reverts the effects learned from specific sequences of tokens. In the loss function described in Equation 6, f_{θ} denotes the model with parameters θ , $x = \{x_1, x_2, \dots, x_T\}$ is a sequence of tokens, and $p_{\theta}(x_t \mid x_{< t})$ represents the conditional probability of x_t being the next token given the preceding tokens $x_{< t}$. The target is to maximize the loss \mathcal{L}_{UL} .

$$\mathcal{L}_{UL}(f_{\theta}, x) = -\sum_{t=1}^{T} \log(p_{\theta}(x_t | x_{< t}))$$
(6)

Extraction Likelihood (EL) and Memorization Accuracy (MA) are used to quantify if a token sequence can be considered to be forgotten. As shown in equation 7, given a token sequence $x = \{x_1, x_2, \dots, x_T\}$ to a language model f pretrained with parameters θ , EL_n is the total n-gram overlap of generated and target token sequences calculated by equation 8. Here, $ng(\cdot)$ is the n-grams of a given token sequence. If an n-gram c is present in the n-grams of sequence b, then it is considered to overlap.

$$EL_n(x) = \frac{\sum_{t=1}^{T-n} OVERLAP_n(f_{\theta}(x_{< t}), x_{\ge t})}{T-n}$$

$$(7)$$

$$OVERLAP_n(a,b) = \frac{\sum_{c \in ng(a)} \mathbb{1}\{c \in ng(b)\}}{|ng(a)|}$$
(8)

Memorization Accuracy (MA) is calculated by equation 9, quantifying the memorization of a given token sequence by the language model. It is considered to be memorized if the token predicted by the LM at position t matches the actual token x_t .

$$MA(x) = \frac{\sum_{t=1}^{T-1} \mathbb{1}\{argmax(p_{\theta}(\cdot|x_{< t})) = x_t\}}{T-1}$$
(9)

A token sequence is considered forgotten if both EL_n and MA are lower than the average of token sequences from a validation corpus. In equations 8 and 9, $\mathbb{1}\{\cdot\}$ is the Indicator function i.e., $\mathbb{1}\{True\} = 1$ and $\mathbb{1}\{False\} = 0$.

Another gradient ascent loss-based unlearning approach has been proposed by Yao et al. [YXL23]. The main idea of this approach is that any task where the language model needs to forget the impact of certain training samples can be achieved by unlearning. To perform unlearning, it only requires the negative samples. Then, the gradient ascent loss is used to forget the negative samples.

Benefits: The unlearning approach is useful for making an LLM forget PII and copyright contents. It can preserve privacy under targeted extraction attacks. The approach is also cost-efficient as it doesn't require re-training the whole language model [Ja22]. It only updates the parameters for a few negative samples. Also, the knowledge unlearning approach causes minimal or negligible deterioration in the original LLM's performance. In some cases, it even results in notable enhancements in LLM performance.

Challenges: A study by Carlini et al. [Ca22] suggests that machine unlearning can even degrade others' privacy. According to Carlini et al. [Ca22], if modifications occur to the underlying dataset, a data point that is presently safe from membership inference may later become vulnerable.

3.4 Data preprocessing approaches

The data preprocessing approach to preserving privacy in language models requires re-training an LLM with anonymised data. This section includes overviews of text anonymisation and deduplication, among the data preprocessing approaches.

According to Lison et al. [Li21b], the process of text anonymisation poses a significant challenge, even for human annotators, as it extends beyond simply identifying predetermined categories of entities. The anonymisation can be done via NLP approaches [Pa22; YRC23] or DP approaches [Ch23; Yu21]. In NLP approaches, different techniques are employed to remove or mask privacy-sensitive information, such as de-identification and obfuscation. In de-identification, sensitive information is removed or masked by generic or anonymous identifiers [Pa22; YRC23]. Obfuscation is also done at the level of latent vector representations rather than modifying the text [Hu20; MBL19]. DP-based text anonymisation approaches use differential privacy principles during the anonymisation process. In SANTEXT, proposed by Yue et al. [Yu21], the authors considered the entire document sensitive and sanitised it with a modified MLDP algorithm [Ch13].

Some approaches involve deduplication of the training data to mitigate privacy risks in language models [KWR22; Le21]. Duplication is when a sequence of characters exactly matches with another sequence of characters. Language models have a tendency to regenerate duplicate sequences from the training data, and an adversary can use them to recover a training sequence. According to Lee et al. [Le21], the frequency of generating an N-length

sequence by a model superlinearly increases with the increase of duplication of that sequence in the training data. During deduplication, two types of duplicate sequences can be considered: the exact substring duplication and the approximate or semantic duplication. But for the privacy-preserving task, the authors considered only the exact sequence as it matches the adversary's goal. When two examples, x_i and x_j , share a sufficiently long substring, that substring is removed from one of them. This deduplication approach reduces the amount of training data generated by the models by reducing the regeneration of duplicate sequences.

Benefits: The main advantages of data preprocessing approaches are that they can provide defence against extraction attacks to some extent and maintain overall model performance. Some approaches even improve performance, so these approaches are suitable for tasks where high performance is a critical factor. Also, the approach is less computationally demanding compared with DP-based approaches.

Challenges: Though privacy can be compromised by recovering approximate duplicates, deduplicating approximate duplicates is more challenging and future work can be done in this direction. Also, the preprocessing approach requires re-training the underlying language models each time it wants to stop a new sequence from regenerating, so it is not suitable for tasks where we need to update for only a few token sequences. According to recent studies [Br20], more than preprocessing approaches are required to remove privacy-sensitive data like bank passwords and medical records and provide weaker privacy protection against this type of information.

4 Key findings

Our study on privacy-preserving methods for language modeling approaches has revealed several important insights. These are summarized below.

- Data preprocessing methods cannot fully provide privacy guarantees, which are insufficient for removing personally identifiable information like names, email addresses, and passwords.
- Data preprocessing and Knowledge Unlearning-based approaches do not harm the overall performances, but performance decreases for Differential Privacy-based approaches in some cases.
- Knowledge Unlearning does not require re-training the language model. It needs to
 perform parameter updates for a few tokens, which is faster than other approaches
 involving re-training.
- While other approaches do not fully protect privacy under the text reconstruction attack during inference, Private Representation Learning-based approaches help to preserve privacy in the inference phase.

 DP provides strong privacy, and recent models such as Adaptive DP and Selective DP improve the performance of differentially private models while maintaining privacy.

5 Limitation and Future works

In this section, we focus on some of the major limitations of the discussed approaches, as well as the limitations of this study. Based on these limitations, we also point out future possible research directions, e.g., expansion of methods to cover languages other than English, development of integrated privacy-preserving techniques, etc.

Language Coverage: Most current studies focus on the English language, but expanding research into privacy risks posed by language models working with other languages is important. Also, the recent progress in privacy-preserving language modelling should be reflected in language models operating in multilingual contexts. Future research should aim to bridge the gap by exploring privacy risks and adapting privacy-preserving methodologies to languages beyond English.

Methodological Limitations: From the findings of this study, it becomes evident that no single method provides overall protection against diverse privacy risks. Data preprocessing-based approaches can not fail to remove personally identifiable information. Some studies [SR20; Zh22] suggested that the widely used DP-based approaches can not protect privacy at the inference phase. Also, the Knowledge Unlearning approach does not focus on inference attacks. Private Representation Learning approaches are focused on preserving privacy at the inference phase and can not give a guarantee about other privacy attributes. Some studies suggest combining multiple approaches to enhance privacy [KWR22], but more work is needed to explore how multiple privacy methods can work together. Future research should also aim to develop integrated privacy-preserving strategies that address the limitations of existing methods, ensuring comprehensive protection against diverse privacy attacks.

Comparative Analysis and Unexplored Impacts of Knowledge Unlearning approach:

The existing Knowledge Unlearning studies lack a comprehensive comparison with more recent DP approaches, leaving an avenue for future research to explore and provide a better understanding of the effectiveness of these two privacy-preserving techniques. Also, as discussed in section 3, Knowledge Unlearning can compromise other people's privacy. Further research is needed to investigate the impact of Knowledge Unlearning on other people's privacy.

Limitations of this work: One limitation of this work is that it only focuses on four approaches: Private Representation Learning, Knowledge Unlearning, Data preprocessing,

and Differential Privacy. Due to page limitations, we could not discuss some other available approaches, e.g., Federated Learning and Homomorphic Encryption. Also, this paper's findings are based on previous works, and this study does not include any case studies. These are left for future work.

6 Conclusion

In recent years, LLMs are becoming an integral part of our lives. People are using LLMs with little or no knowledge of how many privacy risks these models pose. This study explains these privacy risks posed by the current LLMs and currently used approaches for mitigating these risks. This paper provides an overview of the approaches and their benefits and challenges. Privacy-preserving approaches in language modeling implement different strategies to mitigate privacy risks in language models. As described in this study, mitigating privacy risks is a difficult task, and various privacy attacks make the task more difficult. This study discusses four kinds of approaches: Private Representation Learning, Knowledge Unlearning, Data preprocessing, and Differential Privacy. All the approaches can preserve privacy against specific attacks, and none of them gives privacy against all types of attacks. The aim of this work was to understand the approaches and find out the gaps to help the community focus on the bigger unresolved questions.

References

- [Ab16] Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. Pp. 308–318, 2016.
- [An21] Anil, R.; Ghazi, B.; Gupta, V.; Kumar, R.; Manurangsi, P.: Large-scale differentially private BERT. arXiv preprint arXiv:2108.01624/, 2021.
- [Br20] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33/, pp. 1877–1901, 2020.
- [Ca21] Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U., et al.: Extracting Training Data from Large Language Models. In: 30th USENIX Security Symposium (USENIX Security 21). Pp. 2633–2650, 2021.
- [Ca22] Carlini, N.; Jagielski, M.; Zhang, C.; Papernot, N.; Terzis, A.; Tramer, F.: The privacy onion effect: Memorization is relative. Advances in Neural Information Processing Systems 35/, pp. 13263–13276, 2022.

- [Ch13] Chatzikokolakis, K.; Andrés, M. E.; Bordenabe, N. E.; Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13. Springer, pp. 82–102, 2013.
- [Ch23] Chen, S.; Mo, F.; Wang, Y.; Chen, C.; Nie, J.-Y.; Wang, C.; Cui, J.: A Customized Text Sanitization Mechanism with Differential Privacy. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp. 5747–5758, 2023.
- [DR+14] Dwork, C.; Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 9/3–4, pp. 211–407, 2014.
- [Dw08] Dwork, C.: Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Springer, pp. 1–19, 2008.
- [GNG21] Graves, L.; Nagisetty, V.; Ganesh, V.: Amnesiac machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. 13, pp. 11516–11524, 2021.
- [He22] Heikkilä, M.: What does GPT-3 "know" about me, 2022.
- [Hu20] Huang, Y.; Song, Z.; Chen, D.; Li, K.; Arora, S.: TextHide: Tackling data privacy in language understanding tasks. arXiv preprint arXiv:2010.06053/, 2020.
- [Ja22] Jang, J.; Yoon, D.; Yang, S.; Cha, S.; Lee, M.; Logeswaran, L.; Seo, M.: Knowledge unlearning for mitigating privacy risks in language models. arXiv preprint arXiv:2210.01504/, 2022.
- [JSM88] Johnson, S. R.; Symonides, J.; Mayor, F.: The Universal Declaration of Human Rights. Concerts for Human Rights Foundation, Incorporated (CHRF), 1988.
- [Ki24] Kim, S.; Yun, S.; Lee, H.; Gubri, M.; Yoon, S.; Oh, S. J.: Propile: Probing privacy leakage in large language models. Advances in Neural Information Processing Systems 36/, 2024.
- [KWR22] Kandpal, N.; Wallace, E.; Raffel, C.: Deduplicating training data mitigates privacy risks in language models. In: International Conference on Machine Learning. PMLR, pp. 10697–10707, 2022.
- [Le21] Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N.: Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499/, 2021.
- [Li20] Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems 2/, pp. 429–450, 2020.

- [Li21a] Lin, B. Y.; He, C.; Zeng, Z.; Wang, H.; Huang, Y.; Dupuy, C.; Gupta, R.; Soltanolkotabi, M.; Ren, X.; Avestimehr, S.: Fednlp: Benchmarking federated learning methods for natural language processing tasks. arXiv preprint arXiv:2104.08815/, 2021.
- [Li21b] Lison, P.; Pilán, I.; Sanchez, D.; Batet, M.; Øvrelid, L.: 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). Association for Computational Linguistics, Online, pp. 4188–4203, 2021.
- [Ma13] Mantelero, A.: The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'. Computer Law & Security Review 29/3, pp. 229–235, 2013.
- [MBL19] Mosallanezhad, A.; Beigi, G.; Liu, H.: Deep reinforcement learning-based text anonymization against private-attribute inference. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Pp. 2360–2369, 2019.
- [Mc17] McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B. A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR, pp. 1273–1282, 2017.
- [Na23] Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; Lee, K.: Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035/, 2023.
- [Pa20] Pan, X.; Zhang, M.; Ji, S.; Yang, M.: Privacy risks of general-purpose language models. In: 2020 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 1314–1331, 2020.
- [Pa22] Papadopoulou, A.; Lison, P.; Øvrelid, L.; Pilán, I.: Bootstrapping Text Anonymization Models with Distant Supervision. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 4477–4487, 2022.
- [Re20] Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; McMahan, H. B.: Adaptive federated optimization. arXiv preprint arXiv:2003.00295/, 2020.
- [Sh21] Shi, W.; Cui, A.; Li, E.; Jia, R.; Yu, Z.: Selective differential privacy for language modeling. arXiv preprint arXiv:2108.12944/, 2021.
- [Sh22] Shi, W.; Shea, R.; Chen, S.; Zhang, C.; Jia, R.; Yu, Z.: Just fine-tune twice: Selective differential privacy for large language models. arXiv preprint arXiv:2204.07667/, 2022.

- [SR20] Song, C.; Raghunathan, A.: Information leakage in embedding models. In: Proceedings of the 2020 ACM SIGSAC conference on computer and communications security. Pp. 377–390, 2020.
- [VV17] Voigt, P.; Von dem Bussche, A.: The EU general data protection regulation (GDPR): A Practical Guide. Springer Cham, 2017, ISBN: 978-3-319-57958-0.
- [WGX22] Wu, X.; Gong, L.; Xiong, D.: Adaptive Differential Privacy for Language Model Training. In: Proceedings of the First Workshop on Federated Learning for Natural Language Processing (FL4NLP 2022). Pp. 21–26, 2022.
- [YRC23] Yermilov, O.; Raheja, V.; Chernodub, A.: Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization. In: Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). Association for Computational Linguistics, Toronto, Canada, pp. 232–241, 2023.
- [Yu21] Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; Chow, S. S. M.: Differential Privacy for Text Analytics via Natural Text Sanitization. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, pp. 3853–3866, 2021.
- [YXL23] Yao, Y.; Xu, X.; Liu, Y.: Large Language Model Unlearning, 2023, arXiv: 2310.10683 [cs.CL].
- [Zh22] Zhou, X.; Lu, J.; Gui, T.; Ma, R.; Fei, Z.; Wang, Y.; Ding, Y.; Cheung, Y.; Zhang, Q.; Huang, X.: TextFusion: Privacy-Preserving Pre-trained Model Inference via Token Fusion. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 8360–8371, 2022.
- [Zh23] Zhou, X.; Lu, Y.; Ma, R.; Gui, T.; Wang, Y.; Ding, Y.; Zhang, Y.; Zhang, Q.; Huang, X.-J.: Textobfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In: Findings of the Association for Computational Linguistics: ACL 2023. Pp. 5459–5473, 2023.