Provably Confidential Language Modelling

Xuandong Zhao [1] **Lei Li** [1] **Yu-Xiang Wang** [1]

Abstract

Large language models are shown to memorize privacy information such as social security numbers in training data. Given the sheer scale of the training corpus, it is challenging to screen and filter all privacy data, either manually or automatically. In this paper, we propose **C**onfidentially **R**edacted **T**raining (CRT), a method to train language generation models while protecting the confidential segments. We borrow ideas from differential privacy (which solves a related but distinct problem) and show that our method is able to *provably prevent* unintended memorization by randomizing parts of the training process. Moreover, we show that redaction with an approximately correct screening policy *amplifies* the confidentiality guarantee. We implement the method for both LSTM and GPT language models. Our experimental results show that the models trained by CRT obtain almost the same perplexity while preserving strong confidentiality.

1. Introduction

Language models (LM) have rich real-world applications in, among others, machine translation (Bahdanau et al., 2015), AI chatbots (Hosseini-Asl et al., 2020), question answering (Kwiatkowski et al., 2019), and information retrieval (Ganguly et al., 2015). The advent of transformers (Vaswani et al., 2017) has fostered a dramatic advancement in the capabilities of generative neural language models, yet they come at a cost to privacy, as the amount of excess parameters in the LM enables it to memorize certain training samples. Recent works show that sensitive user information from the training dataset, such as address and name, can be extracted verbatim from text generation models by querying the LM as an API (Carlini et al., 2019; 2021; Lee et al., 2022). How to train a high-performing language model without memorizing sensitive text has become a major research challenge.

Existing solutions to this problem primarily leverage differential privacy (DP) (Dwork et al., 2006).

Differentially private learning algorithms ensure that an attacker could not infer whether a data point is used for training, let alone extracting the sensitive information within that data point.

However, there are several mismatches between the problem of *privacy* that DP addresses, and our problem of preventing the memorization of sensitive text (henceforth referred to as *confidentiality*). First, confidential information in a natural language dataset is sparse (e.g., the bulk of an email might not carry confidential information). DP's undiscriminating protection for all sentences could be unnecessarily conservative which limits the utility of the trained model. Second, what needs to be protected is the content of the sensitive text, rather than the data context. For example, in the sentence "`My SSN is 123-45-6789.`", it is the actual SSN that we hope to conceal rather than the general information that someone entered her SSN in a chatbot dialogue. Thirdly, the same sensitive content could appear in many data points, which makes the protection of the content more challenging than protecting one data sample. These differences motivate us to treat the problem of confidentiality protection in LM separately with new definitions.

Besides DP, we also consider classical techniques of redaction and deduplication. *Redaction* refers to the process of removing sensitive or classified information from a document prior to its publication in governmental and legal contexts. *Deduplication* is the procedure of detecting and removing identical and nearly identical texts from a corpus. The main challenge of applying these techniques is that it is hard to manually redact a gigantic dataset and automated tools are far from being perfect.

The contribution of this paper is fivefold.

1. We show that in the absence of a perfect screening policy, the risk of a language model memorizing sensitive content is real and can be efficiently exploited with only blackbox access to the model even if the learning algorithm satisfies the recently proposed notion of

[1]UC Santa Barbara. Correspondence to: Xuandong Zhao <xuandongzhao@cs.ucsb.edu>, Lei Li <leili@cs.ucsb.edu>, Yu-Xiang Wang <yuxiangw@cs.ucsb.edu>.

Accepted to the *1st Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA.

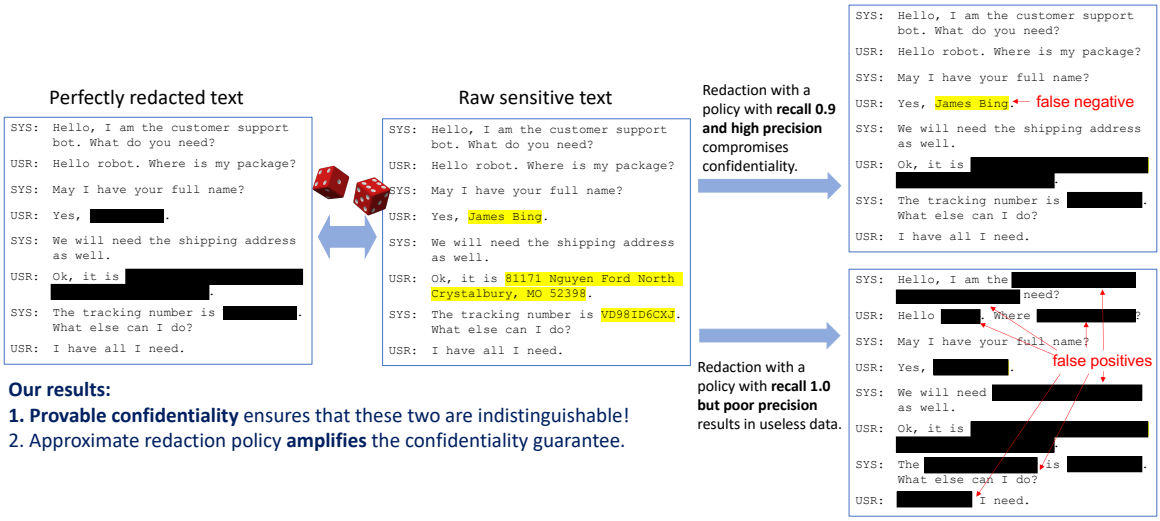

Figure 1. An example from simulated dialog dataset `CustomerSim`. The yellow highlights are confidential content (middle). Left shows the text after *Redaction* by a sequence labeling policy π. However, if the policy is not perfect, there exists false negative or false positive samples as shown on the right.

selective differential privacy (Shi et al., 2021).

2. Inspired by differential privacy, we introduce a new definition of *confidentiality* which precisely quantifies the risk of leaking sensitive text.
3. We propose CRT to train language generation models while protecting confidential text. The method with deduplication and redaction operations work even under imperfect confidential text labeling policies.
4. We theoretically prove that CRT, combined with differentially private stochastic gradient descent (DP-SGD), provides strong confidentiality guarantees.
5. Our experiments on both MultiWOZ 2.2 and CustomerSim datasets show that different models trained by CRT can achieve the same or better perplexity than existing solutions (against the attacks of Carlini et al. (2019; 2021)).

To the best of our knowledge, we are the first that rigorously establish the role of deduplication and redaction in achieving provably stronger confidentiality (or the related differential privacy) guarantees; and the first that achieve provably confidentiality in transformer models with only a mild utility loss.

Conclusion. In this paper, we propose confidentially redacted training (CRT), a method to train language models while protecting the secret texts. We introduce a new definition of confidentiality which quantifies the risk of leaking sensitive content. We prove the effectiveness of CRT both theoretically and empirically on multiple datasets and language models.

Algorithm 1: CRT

Input : Dataset D (after tokenization / splitting), labelling policies π, π_c, number of epochs T

$D' \leftarrow \text{Dedup}(D)$
$D'' \leftarrow \text{Redact}_\pi(D')$
$D^{pri} \leftarrow \{s \in D'' | \exists x \in s \text{ s.t. } \pi(s,x) = 1 \text{ or } \exists x \subset s \text{ s.t. } \pi_c(s,x) = 1\}$
$D^{pub} \leftarrow \{s \in D'' | s \notin D^{pri}\}$
for $e = 1, ..., T$ **do**
  Run one epoch of SGD with D^{pub}
  Run one epoch[1] of DP-SGD with D^{pri}
end

Broader Impact. This work will alleviate ethical concerns of large-scale pre-trained language models. This paper provides one promising solution to an important aspect of NLP: training high quality language models for text generation without compromising confidential information. The current use cases of language models involve pretraining on public web corpus and fine-tuning on individual application data. However, the private application specific data often contains user-generated sensitive information. The proposed method in this paper aims to use as much individual fine-tuning data as possible, while does not leak or memorize any confidential information with provable guarantees. Without the method, one has to either use the general pretraining

[1]DP-SGD uses Poisson-sampled Gaussian mechanisms (with a random batchsize), thus cannot ensure all data points are seen and some data points might be seen many times. One epoch means the number of iterations that in expectation covers $|D^{pri}|$ data points.

LM without fine-tuning or manually filter sensitive information and fine-tuning on the remaining. It can be applied in broader applications that need language models or text generation models.

In our experiments, we use a simulation scheme to mimic confidential content in a real corpus. We did not compromise any real user's confidential information.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. X. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2019.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D. X., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Ganguly, D., Roy, D., Mitra, M., and Jones, G. J. Word embedding based generalized language model for information retrieval. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S., and Socher, R. A simple language model for task-oriented dialogue. *ArXiv*, abs/2005.00796, 2020.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A. P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q. V., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022.

Shi, W., Cui, A., Li, E., Jia, R., and Yu, Z. Selective differential privacy for language modeling. *ArXiv*, abs/2108.12944, 2021.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.