# The Past, Present, and Future of Privacy in Large Language Models

Hello to all! Before diving in, I wanted to first thank you for taking the time to read this article. The paper I wish to share with you today is called Privacy in Large Language Models: Attacks, Defenses and Future Directions. It was published on October 13 of this year and discusses what privacy is as it relates to Large Language Models (LLMs), along with the different strategies that have been used/are being continually evaluated today to mitigate sensitive data leakage and exploitation.

So to begin, let's first find out how privacy relates to LLM's and why it is so important. LLM's require massive amounts of (usually sensitive) training data, which is typically pulled from the internet or other public domains. One important aspect of models like ChatGPT is that they must protect the privacy of the information used to train such models from harmful adversarial prompts which aim to extract sensitive training data from the model itself. This has become a very hot topic for discussion today, as failing to protect such sensitive information can result in costly lawsuits for the companies that develop these models.

The first step in preserving privacy is defining it in mathematical terms so that it can be quantified and measured. A popular framework adopted today is Differential Privacy (DP), which the article defines as incorporating "random noise into aggregated data to allow data mining without exposing participants' private information" (Section 2.2.2). Some of the various problems facing LLMs other than simply exposing training data include inference data privacy (sensitive information about previous prompts/queries stored by the model) and re-identification (adversarial identification of individuals’ information after various benign queries). The greater challenge for LLMs today lies in finding a balance between privacy and utility.

There are a great many different kinds of “attacks” that LLMs can fall victim to today, some of which fall under the category of backdoor attacks, which, as defined in the paper, “Involve the insertion or modification of specific input patterns that trigger the model to misbehave or produce targeted outputs” (Section 3.1). Data poisoning, which is a “less potent attack where only a portion of the training data is manipulated,” offers a vehicle to implement backdoor attacks. More specific examples of this attack are covered in-depth in the paper. Another type of attack, which can be considered a different form of a backdoor attack, is a prompt injection attack, which is once more defined in the paper as an attack that “manipulates or injects malicious content into the prompt or input $p$ given to the model to get the altered pattern $\tilde{p}$, with the aim to influence its behavior or generate unwanted outputs $\mathrm{f}(\tilde{\mathrm{p}})$” (Section 3.2). Additionally, training data extraction attacks are just what they sound like. These attacks depict a scenario in which an adversary attempts to recover a model's memorized training data, typically with what's known as a “jailbreaking prompt”. In contrast, membership inference attacks (MIA) describe scenarios in which an adversary already has information on potential samples used within the training of the model and attempts to determine whether a particular sample was used in the model’s training.

While less common, more advanced attacks are discussed within the paper, where an adversary has access to more information about the model such as gradients and vector representations which include: attribute inference attacks, embedding inversion attacks, prompt extraction attacks, adversarial attacks, side channel attacks and decoding algorithm stealing. Each of these are discussed more in-depth within the paper itself.

As for the defenses to such attacks, we turn back to DP in LLMs, which provide a privacy guarantee for the model. One commonly implemented deep learning algorithm today is Differential Privacy Stochastic Gradient Descent (DPSGD), which attempts to preserve privacy by adding noise to gradients during the training process. The paper covers four distinct types of DP-based LLMs in depth: DP-based pre-training, DP-based fine-tuning, DP-based Prompt tuning, and DP-based synthetic text generation, each of which are described in greater detail in the paper. Secure multi-party computation (SMPC) - based LLMs provide another solution to the privacy problem. SMPC is a “cryptographic technique that allows multiple parties to collaborate in training a machine learning model while maintaining the privacy of their individual data” and which “enables these parties to jointly compute model updates without exposing their private data to others, ensuring that each party can contribute their local data to the training process without disclosing any sensitive information” (Section 4.2). Another privacy solution in the world of LLMs is Federated Learning, which the paper defines as “a privacy-preserving distributed learning paradigm enabling multiple parties to train or fine-tune their LLMs collaboratively without sharing private data owned by participating parties" (Section 4.3). However, this framework comes with its own limitations and vulnerabilities, in that they can leak data privacy if they suffer specific types of inference attacks.

DP based LLMs are widely adopted for data privacy, but they face notable limitations. One limitation is the theoretical worst-case bounding, as DP-tuned LLMs assume a powerful adversary with full control over training data, however, in real world scenarios it is uncommon to come across such a capable adversary, creating a gap between theoretical analysis and practical scenarios, which might include a simple training data extraction attack. Additionally, the utility of DP-based LLMs tends to degrade, especially when applied to complex downstream tasks, despite claims that careful tuning can achieve similar performance to non-DP tuning on simpler tasks. This degradation weakens the motivation for using DP-based fine-tuning.

Toward the end of the paper, the text shifts to outlining several key research areas. First, with regard to prompt injection attacks–which aim to influence LLMs' output–the domain as it relates to privacy is still relatively uncharted, and the need for more robust defenses will be required as current systems continue to fall victim to these attacks. The second area focuses on future improvements in SMPC for privacy-preserving inference in LLMs. Challenges and ongoing efforts involve integrating Model Structure Optimization (MSO) and SMPC Protocol Optimization (SPO) to design efficient, versatile privacy-preserving algorithms. The third area addresses the alignment of privacy with human perception, emphasizing the limitations of current privacy studies and the need for a broader understanding of privacy as defined by societal norms, which may be influenced by social norms, ethnicity, religious beliefs, and privacy laws. The fourth area discusses empirical privacy evaluation, emboldening the need for more nuanced and detailed metrics beyond Differential Privacy parameters and highlighting the importance of assessing privacy in real-world scenarios. The final section discusses contextualized privacy judgment, emphasizing the lack of a general privacy violation detection framework and the need for frameworks with reasoning ability in complex, long-context situations–especially when working with LLM-based chatbots.

In summary, this article navigates through the past, present, and future of privacy in LLMs by examining the critical role of differential privacy and the various types of attacks and defensive strategies employed today. The paper discusses other solutions to privacy as well, such as Federated Learning, along with its limitations. Looking ahead, the article highlights several key research areas within which lie many unanswered questions, emphasizing the evolving nature of privacy challenges and the ongoing efforts to address them.

Li, H., Chen, Y., Luo, J., Kang, Y., Zhang, X., Hu, Q., Chan, C., &amp; Song, Y. (2023, October 16). Privacy in large language models: Attacks, defenses and Future Directions. arXiv.org. https://arxiv.org/abs/2310.10383