

Mitigating Data Leakage in Large Language Models

Problem Statement

Artificial intelligence is swiftly evolving. Today, Large Language Models (LLMs) are among the most advanced forms of technology developed for natural language understanding and generation. The technology is highly effective, yet it also represents a major threat to security and privacy, posing a threat of data leakage, including both sensitive and private information (for example, personal data, PII, confidential business information, system commands)

Recent observations indicate that Large Language Models (LLMs) - particularly those based on transformer architectures - learn to memorize and recall rare or even sensitive input sequences with specific prompts. Although unlikely to be peaked at true entropy, the parametric memory of LLMs significantly increases the risk of data exposure in state-of-the-art settings such as multi-turn conversational models, retrieval-augmented generation (RAG), and multimodal models. This paper investigates such phenomena and develops a rigorous and quantifiable framework to detect them. Building on prior work, we formalize the phenomenon as retrieval and propose a suite of evaluation metrics to comprehensively assess the risk of data leakage. Empirical analyses on both synthetic and real-world datasets—spanning Web Apply, Web Search, Web Vision, and Open Web Arena—demonstrate that models can be tricked into retrieving sensitive information, and causal ablation further supports the causal contribution of interaction history. These findings highlight robust vulnerabilities in contemporary language models and motivate a re-examination of both model architecture and deployment practices.

Privacy preservation is a crucial problem, especially for high-stakes applications in healthcare, finance, and enterprises, where leaking sensitive information can have severe ethical, legal, and operational impacts. As LLMs are increasingly used for high-stakes decision making, ensuring privacy of input data and reliability of models is critical for maintaining user trust and compliance with regulations.

Existing approaches have already started exploring ways to mitigate bias in deep learning-driven private conversational interfaces. Techniques like input filtering or privacy-aware decoding, as well as methods that use differential privacy or, more specifically, intervene on circuit-level, have been proposed and implemented. While some of these have some limitations regarding customization for different model architectures or high computational overhead, they mostly focus on one part of the system—either input, model, or output. There is a lack of more holistic approaches.

Here we discuss why a scalable, integrated, and model-agnostic framework to prevent data leakage for all LLMs and use cases, without compromising performance is

crucial and present a unified solution to the addressed challenges to the state-of-the-art.

Research Objectives

This paper aims to mitigate the problems of data leakage from Large Language Models. To this end, we propose a new comprehensive approach to preventing data leakage and test it on a large model and dataset. The methodology and goals are explained below.

1 .Data Leakage Attacks on Large Language Models

This investigation aims to elucidate the principal mechanisms through which data leakage occurs, incorporating both established and emergent categories such as memorization, prompt-induced and prompt-based attacks, and retrieval-based exposure.

2 .Evaluating current countermeasures to data leakage.

Evaluating existing approaches such as privacy-aware decoding, circuit patching, and input filtering.

These techniques should be examined in terms of strengths and limitations, as well as their implications for scalability and overall model performance.

3 .Design hybrid mitigation techniques for LLM data leakage.

How to combine input-level, model-level and inference-time defenses in a single system. So that multiple leakage vectors are addressed.

4 .To implement and validate the proposed framework

To evaluate the quality of the resulting case work using a set of metrics such as leakage rate reduction, accuracy, and response quality.

Dataset-level analysis should test the developed framework on multiple data-sets for overall robustness of results.

5 .Compare with other related approaches.

We assess the scalability, performance, and the privacy-utility trade-off.

The question that needs to be answered here is: To what extent does this new approach outperform current approaches?

6 .To guide future research and practical applications.

Our future research will concentrate on enriching the above described framework and identifying its limitations.

To arm practitioners with insights for running secure LLMs in realistic scenarios.

References:

1. Prompt Leakage effect and mitigation strategies for multi-turn LLM Applications, <https://aclanthology.org/2024.emnlp-industry.94/>
2. Defeating Cerberus: Concept-Guided Privacy-Leakage Mitigation in Multimodal Language Models - arXiv, ‘<https://arxiv.org/html/2509.25525v1>
3. Privacy-Aware Decoding: Mitigating Privacy Leakage of Large Language Models in Retrieval-Augmented Generation -arXiv.org, ‘<https://arxiv.org/abs/2508.03098>
4. PATCH: Mitigating PII Leakage in Language Models with Privacy-Aware Targeted Circuit Patching - ResearchGate, ‘17 https://www.researchgate.net/publication/396374186_PATCH_Mitigating_PII_Leakage_in_Language_Models_with_Privacy-Aware_Targeted_Circuit_Patching
5. Privacy-Aware Detection for Large Language Models Using a Hybrid BiLSTM-HMM Approach - IEEE Xplore, ‘<https://ieeexplore.ieee.org/iel8/6287639/6514899/11077118.pdf>
6. Investigating How Pre-training Data Leakage Affects Models' Reproduction and Detection Capabilities - ACL Anthology, ‘<https://aclanthology.org/2025.emnlp-main.1201/>