PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps

Ruixuan Liu[§], Tianhao Wang[†], Yang Cao[‡], and Li Xiong[§]

Emory University[§], University of Virginia[†], Tokyo Institute of Technology[‡]

Full Bibliographic Reference

Ruixuan Liu, Tianhao Wang, Yang Cao, and Li Xiong. 2024. PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (CCS '24). Association for Computing Machinery, New York, NY, USA, 3511–3524. https://doi.org/10.1145/3658644.3690279

Link / DOI

DOI/URL: https://doi.org/10.1145/3658644.3690279

Original Abstract

The pre-training and fine-tuning paradigm has demonstrated its effectiveness and has become the standard approach for tailoring language models to various tasks. Currently, community-based platforms offer easy access to various pre-trained models, as anyone can publish without strict validation processes. However, a released pre-trained model can be a privacy trap for fine-tuning datasets if it is carefully designed. In this work, we propose PreCurious framework to reveal the new attack surface where the attacker releases the pre-trained model and gets a black-box access to the final fine-tuned model. PreCurious aims to escalate the general privacy risk of both membership inference and data extraction on the fine-tuning dataset. The key intuition behind PreCurious is to manipulate the memorization stage of the pre-trained model and guide fine-tuning with a seemingly legitimate configuration. While empirical and theoretical evidence suggests that parameter-efficient and differentially private fine-tuning techniques can defend against privacy attacks on a fine-tuned model, PreCurious demonstrates the possibility of breaking up this invulnerability in a stealthy manner compared to fine-tuning on a benign pre-trained model. While DP provides some mitigation for membership inference attack, by further leveraging a sanitized dataset, PreCurious demonstrates potential vulnerabilities for targeted data extraction even under differentially private tuning with a strict privacy budget e.g. $\varepsilon = 0.05$. Thus, PreCurious raises warnings for users on the potential risks of downloading pre-trained models from unknown sources, relying solely on tutorials or common-sense defenses, and releasing sanitized datasets even after perfect scrubbing.

PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps

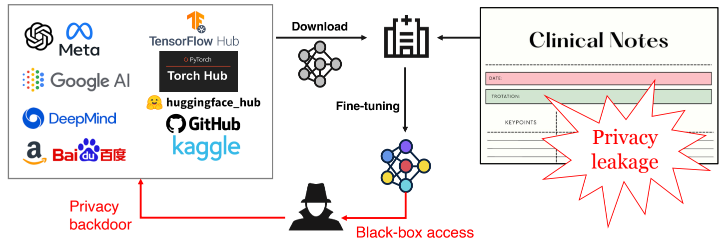
Ruixuan Liu§, Tianhao Wang†, Yang Cao‡, Li Xiong§
Emory University§, University of Virginia†, Tokyo Institute of Technology‡

Published at CCS'2024

Motivation

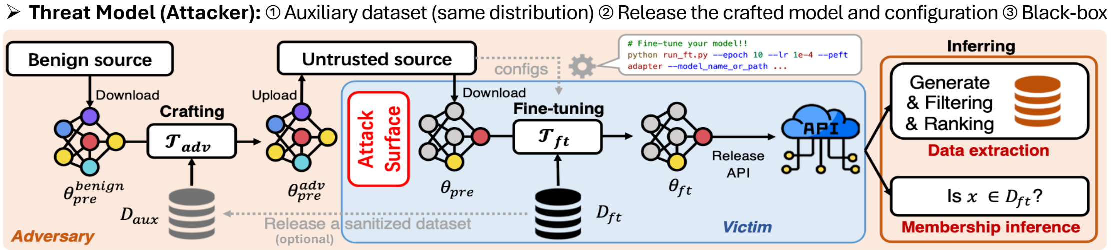
- The pre-train → fine-tune paradigm is standard for domain adaptation (e.g., medical, email, finance).
- Fine-tuning data is often sensitive; models are deployed via APIs with black-box access.
- Community hubs make it easy to download "pre-trained" checkpoints, but model integrity is not guaranteed.

Model supply-chain risk can translate into data privacy risk for downstream fine-tuning



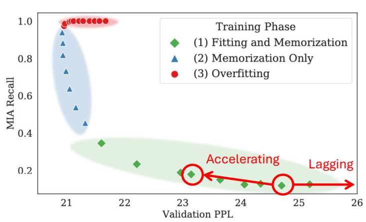
PreCurious Framework Overview

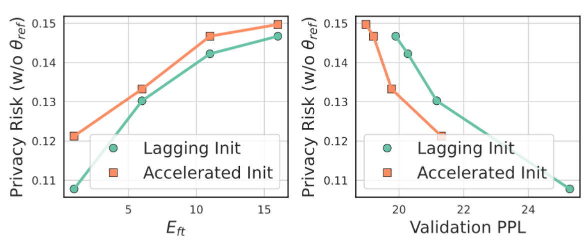
- **Threat Model (Attacker):** ① Auxiliary dataset (same distribution) ② Release the crafted model and configuration ③ Black-box access to the target model



- **Malicious Initialization**: Manipulates open-sourced pre-trained models to "trap" future fine-tuning processes.
- **Smart Adaptation:** Exploits user configs (PEFT, early stopping) to maximize attack success.
- **Amplified Leakage:** Drastically increases MIA and Extraction risks on private fine-tuning datasets.

How to setup the "Privacy Trap"?





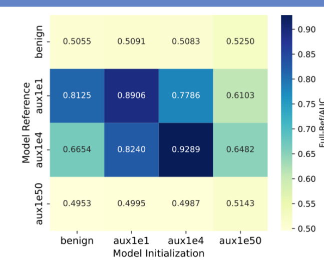
Key Intuition

- Privacy risks of fine-tuning data continuously increases in training
- The final memorization level depends on the starting point on the curve

Two-Case Solutions

- ✓ Case I (Fixed Epochs): Accelerated Init enters memorization earlier → Higher privacy risks
- ✓ Case II (Early Stopping): Lagging Init forces more iterations to converge → Higher privacy risks

A Dual-Use Privacy Trap



The crafted model can be used as:

- The **"pre-trained" model** with privacy trap in fine-tuning
- The **reference model** that calibrates sample-hardness in MIA

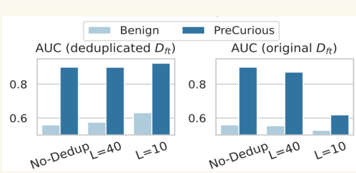
Key Results

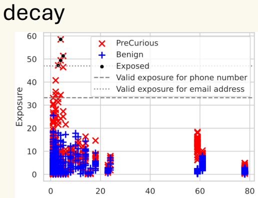
Across datasets and fine-tuning methods, PreCurious amplifies MIA risk (TPR @ 0.01% FPR)

- PubMed: ~**131×** Boost (0.52% → 68.33%)
- PTB: ~**36×** Boost (2.58% → 92.84%)
- Enron: ~**8×** Boost (0.30% → 2.40%)

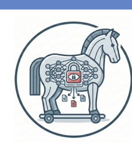
PreCurious **remains effective** under

- DP fine-tuning (overall reduces MIA risk, but still few secrets have valid exposure values)
- Deduplication / Weight decay





Conclusion and Takeaway



Stronger Privacy Auditing

- **New privacy interface of LMs:** a general, effective and stealthy privacy backdoor ***PreCurious***
  - MIA and Data extraction attacks
  - Independent on fine-tuning or model architecture
- **New memorization manipulation**: two strategies to craft privacy backdoor model for comparing by epochs and by performance
- **New understanding of defense:** identify defense vulnerability with active attacker

Paper

Code

Source Verification. Only download verified, official models on open-sourced platforms

Audit Training Dynamics. Monitor memorization patterns even if validation appears normal

Release Sanitized Data with Caution. PreCurious can exploit even perfectly sanitized data to recover secrets

Contact: ruixuan.liu2@emory.edu