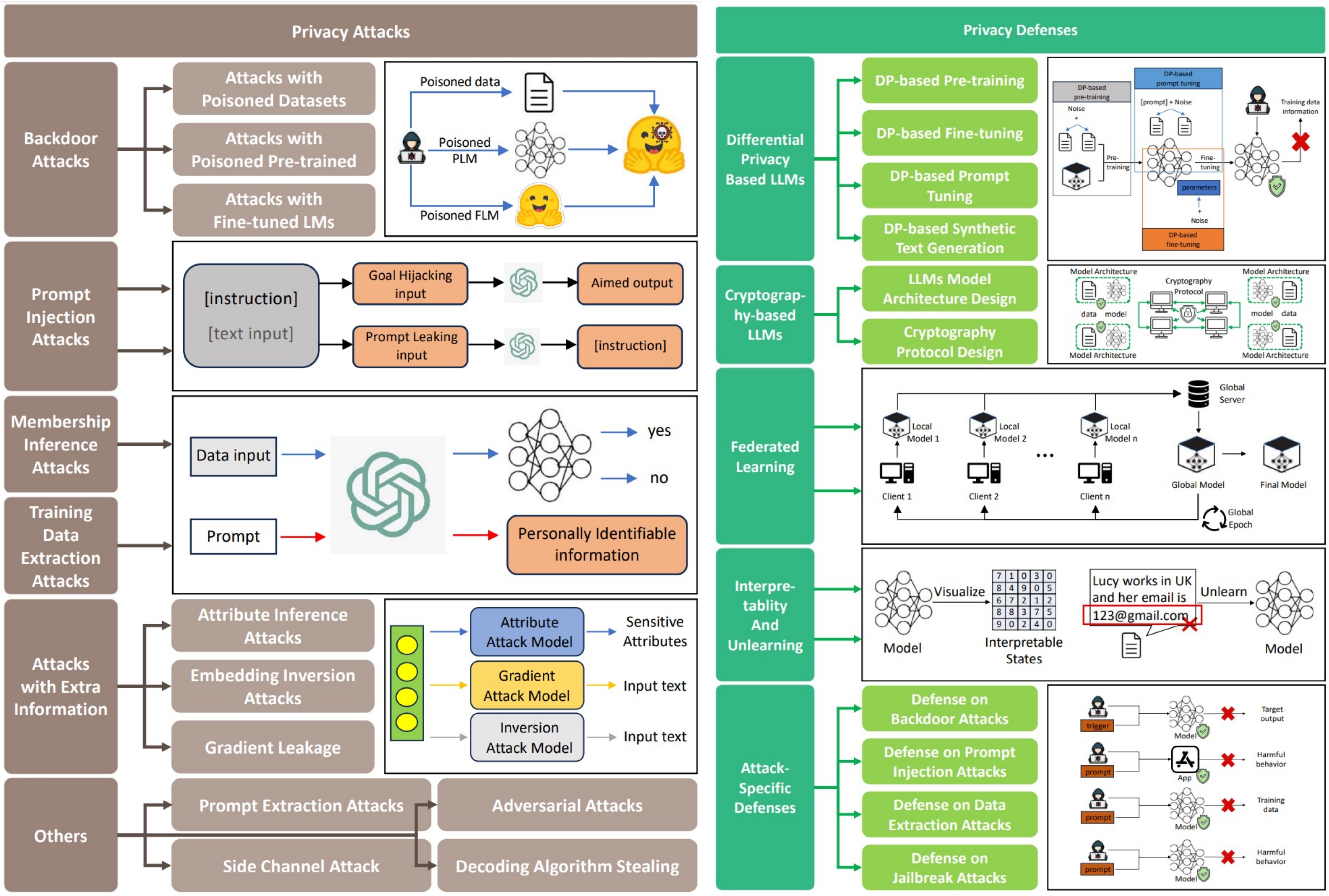
Norwegian Centre for Research-based Innovation

NORCICS

SFI Norwegian Centre for Cybersecurity in Critical Sectors

# Privacy Attacks and Defenses in Large Language Models

Large language models (LLMs) have improved accessibility and usability, enabling various applications like virtual assistants and chatbots, but also raise potential privacy risks. Unrestricted access to these models introduces privacy risks, as their capabilities can be exploited for malicious purposes or unintentionally compromise sensitive information.



Privacy Attacks

Backdoor Attacks
- Attacks with Poisoned Datasets
- Attacks with Poisoned Pre-trained
- Attacks with Fine-tuned LMs

Prompt Injection Attacks

Membership Inference Attacks

Training Data Extraction Attacks

Attacks with Extra Information
- Attribute Inference Attacks
- Embedding Inversion Attacks
- Gradient Leakage

Others
- Prompt Extraction Attacks
- Adversarial Attacks
- Side Channel Attack
- Decoding Algorithm Stealing

Privacy Defenses

Differential Privacy Based LLMs
- DP-based Pre-training
- DP-based Fine-tuning
- DP-based Prompt Tuning
- DP-based Synthetic Text Generation

Cryptography-based LLMs
- LLMs Model Architecture Design
- Cryptography Protocol Design

Federated Learning

Interpretability And Unlearning

Attack-Specific Defenses
- Defense on Backdoor Attacks
- Defense on Prompt Injection Attacks
- Defense on Data Extraction Attacks
- Defense on Jailbreak Attacks

Several key strategies are emerging to preserve privacy in large language models (LLMs). **Differential privacy** is a popular technique that injects noise into the data or model outputs, preventing the identification of individual data points while maintaining model accuracy. **Federated learning** is another approach where data remains localized on user devices, with only model updates shared, thus reducing the risk of exposing raw data. **Model distillation** is used to create a smaller, privacy-preserving model that retains essential knowledge without needing direct access to sensitive data. **Encryption methods**, such as homomorphic encryption, allow computation on encrypted data, adding an additional layer of protection. Finally, **access control** and **auditing mechanisms** can help monitor and restrict the use of sensitive data, ensuring adherence to privacy policies. Each of these approaches has its strengths and challenges, often requiring careful calibration to balance privacy with model performance.

Ahmad Hassanpour email: Ahmad.Hassanpour@ntnu.no