Beihang
University

LoRO: Real-Time on-Device Secure Inference for LLMs via TEE-Based Low Rank Obfuscation

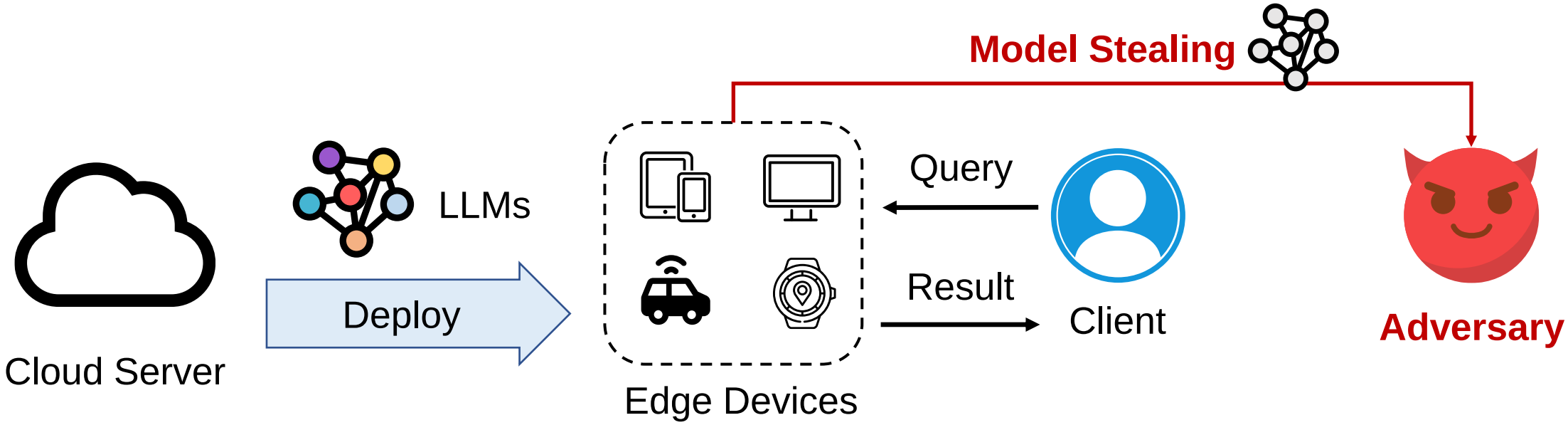
Gaojian Xiong, Yu Sun, Jianhua Liu, Jian Cui, Jianwei Liu

School of Cyber Science and Technology, Beihang University, China

Corresponding authors: Yu Sun (sunyv@buaa.edu.cn)

Background: Edge-Side LLMs Inference Scenario

- **Large Language Models (LLMs) are widely deployed in edge devices to serve for clients**



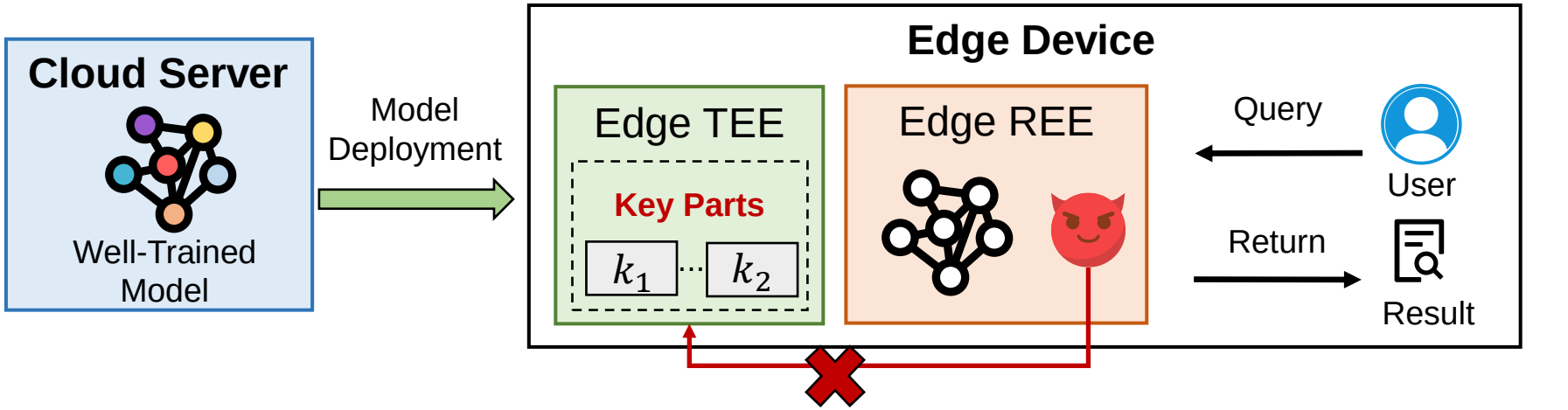
- **Model Theft Concerns:**

  Deployed valuable LLMs are usually **vulnerable to be stolen (intellectual property loss)**

Background: Prior Solutions

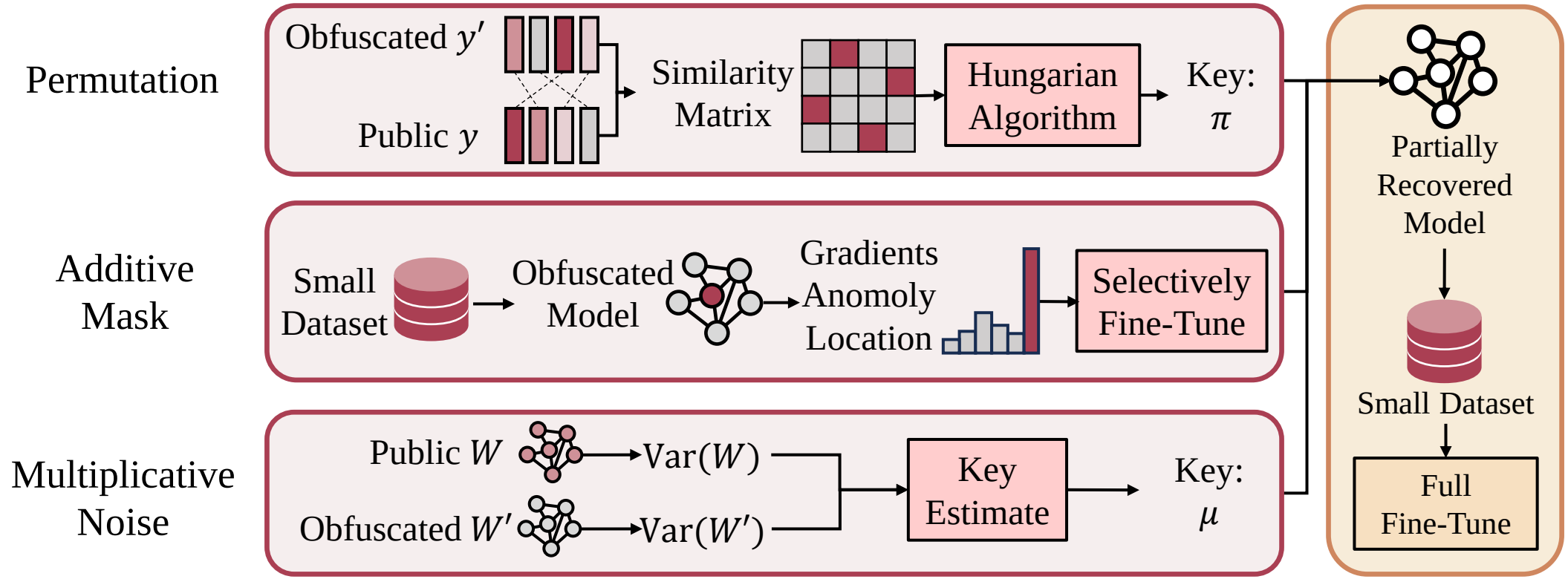
- **Cryptography-based solutions (Multi-Party Computation and Homomorphic Encryption)**
  - Secure but lead to **significant inference latency**

- **Trusted Execution Environment (TEE)-based solutions**
  - **Obfuscate** most parameters and offload them to Rich Execution Environment (REE) for acceleration
  - **Shield** key parts in TEE to recover the inference results
  - Achieve **model confidentiality** while ensuring **inference efficiency.**



Finding: Statistical Vulnerability in Existing Works

- We reveal **statistical vulnerabilities** in existing TEE-shielded protection methods
- Based on this, we propose **Model Stealing attack with Prior (MSP)**



How to defend against MSP?

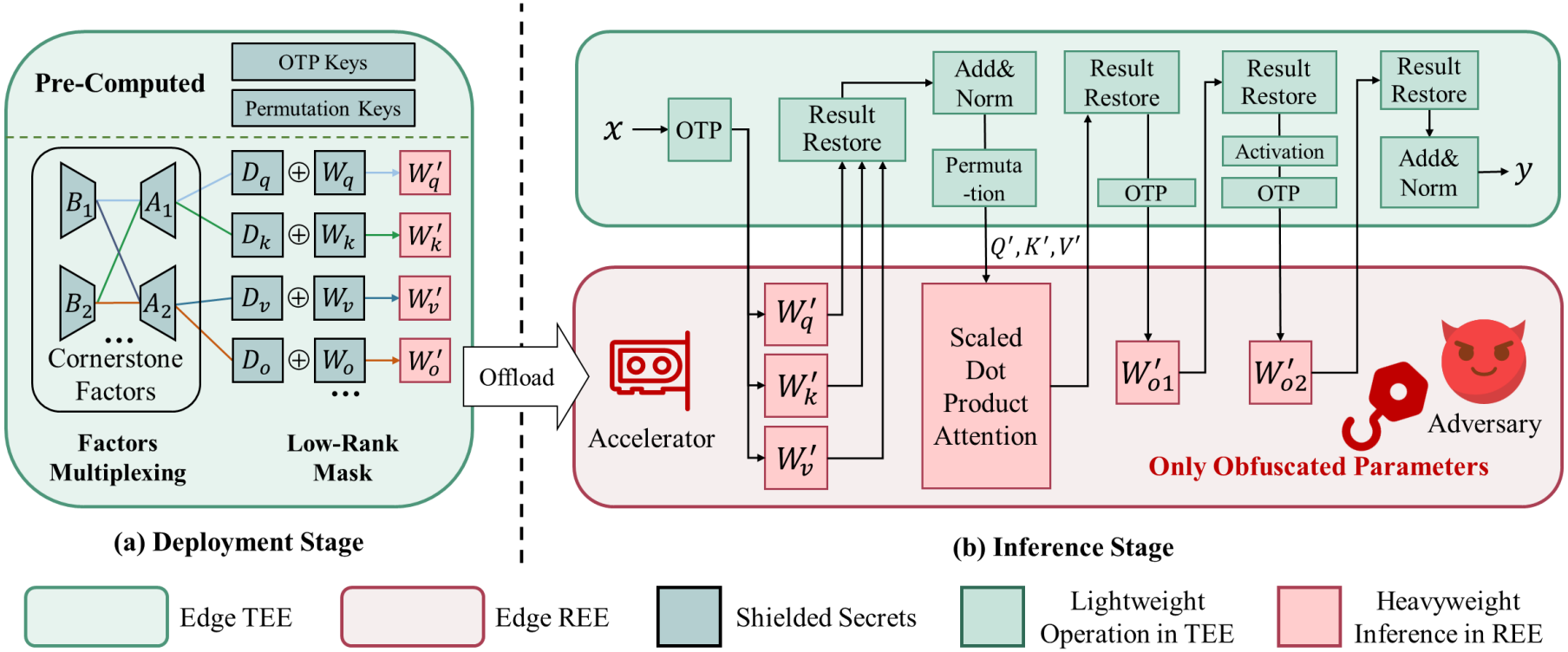
➢ **Dense Mask:** Theoretically, Dense mask is considered secure against MSP, as it can fully obscure the distribution of parameters. However, there are two **key challenges**:

- **Efficiency-Confidentiality dilemma:**
  Edge TEEs face significant resource constraints, particularly when protecting high-complexity computations as dense mask de-obfuscation, which can introduce a hundred-fold increase in inference latency.

- **Limited secure memory challenge:**
  Edge TEEs are constrained by limited secure memory, typically capped at 128 MB. However, to shield masks for 7B LLMs, the secure memory could reach several GB. The introduced memory paging issue leads to a unacceptable increase in inference latency and potential security risk.

Our Solution: Low Rank Obfuscation (LoRO)

- **Low-Rank Masks:** Use low-rank factors to generate dense masks and obfuscate LLM parameters, and recover results in TEE efficiently.
- **Factors Multiplexing:** Reuse several cornerstone factors to generate masks for all layers, reduce secure memory requirement from GB to MB level.



Our Contribution:

- Reveal statistical vulnerabilities in existing methods, and propose Model Stealing with Prior.
- Propose LoRO to defend against powerful Model Stealing attack.
- Experiments on both Intel SGX and ARM TrustZone demonstrate that:
  - MS attack accuracy is reduced to black-box level (0.94×) from existing 3.37×.
  - LoRO introduces only 1.49× inference latency, compared to the 112× of TEE-shielded inference.
  - LoRO introduces no accuracy loss, and requires no re-training or architecture modification.

Thank You!

- Acknowledgement
  - CCF-Phytium Fund (CCF-Phytium 202306)

- Project
  - https://github.com/D1aoBoomm/LoRO