STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Shan Huang
Degree:	Doctor of Philosophy
School/Department:	School of Computing
Date:	December 3, 2026
Time/Location:	11:00 AM EST / Gateway North 303
Title:	Optimizing Quantization for Private and Efficient Machine Learning
Chairperson:	Dr. Tegan Brennan, School of Computing
Committee Members:	Dr. Wendy Hui Wang, School of Computing Dr. Kai Li, School of Computing Dr. Steve Yang, School of Business

ABSTRACT

The ubiquitous deployment of Machine Learning (ML) Systems faces a dual challenge of resource efficiency and data privacy, In Federated Learning (FL), communication bottlenecks and privacy leakage from shared model updates remain critical concerns. Similarly, Large Language Models (LLMs) unlock transformative reasoning capabilities at the cost of immense memory and computational overhead, alongside emerging security and privacy risks during inference. Quantization addresses these resource constraints by compressing gradients in FL and reducing KV-cache memory footprint during LLM inference. However, its broader implications for privacy and security remain insufficiently understood. It is therefore critical to investigate how quantization complicates privacy-utility trade-offs when paired with Differential Privacy (DP) in FL, and whether it alters runtime execution behaviors that amplify side-channel leakage during LLM inference.

This dissertation systematically investigates the privacy and security implications of quantization across FL and LLM inference through three complementary studies. First, we explore quantization as a standalone heuristic defense in FL, proposing a novel quantizer that aligns quantization error with gradient eigenvectors and incorporates bounded randomization to mitigate reconstruction attacks. Empirical results demonstrate higher reduction on the visual fidelity of leaked images than standard quantization method, while preserving higher utility than conventional DP mechanisms. Second, we systematize the landscape of DP-quantization through a unified taxonomy and comprehensive replication study across privacy, utility, and compressibility. We uncover how noise injection timing impacts data entropy and sensitivity, proposing actionable optimization strategies. Third, we characterize KV-cache prefix-matching timing side-channels in LLM inference, demonstrating how KV-cache quantization amplifies their observability and exploitability through changes in runtime behavior and identifying potential mitigations. These contributions aim to establish a comprehensive understanding of the security and privacy dynamics of quantization, providing principled mechanisms to optimize modern ML systems to be private and efficient in practice.