Teng Li

PH.D. CANDIDATE IN ELECTRICAL AND COMPUTER ENGINEERING
University of Illinois Chicago · Chicago, Illinois
+1 (312) 358-0207 | ninmiao@gmail.com | tli81@uic.edu | linkedin.com/in/tengli95

Personal Website & Full Resume

LinkedIn

Professional Summary

Ph.D. candidate in **Edge AI** and **distributed learning and inference** for resilient, communication-efficient ML systems. Research focuses on wireless and decentralized network **optimization** and resource scheduling. Deployed **vision models and LLMs** with **Python/PyTorch** on **NVIDIA Jetson**, developed **RF signal detection** systems, and optimized GPU inference latency and throughput with **ONNX Runtime / TensorRT**. Wireless systems experience includes **O-RAN**, **OpenAirInterface (OAI)**, and **SDR (USRP)**.

Education

University of Illinois Chicago — *Chicago, Illinois*

PH.D. CANDIDATE, ELECTRICAL AND COMPUTER ENGINEERING · GPA: 4.0/4.0 — *Sept 2019 - **Expected December 2026***

- **Advisor:** Prof. Hulya Seferoglu
- **Research Interests:** Distributed Learning, Edge AI, Network Optimization, Resilient Systems, LLMs on Edge Devices.

M.S., ELECTRICAL AND COMPUTER ENGINEERING · GPA: 4.0/4.0 — *Aug 2017 - Apr 2019*

- **Concentration:** Computer Communication Networks, Deep Learning, Wireless Communications, Computer Architecture.

Northeastern University (China) — *Shenyang, China*

BACHELOR OF ENGINEERING IN AUTOMATION · GPA: 80.8/100 — *Aug 2013 - May 2017*

- **Concentration:** Automation Control Systems, Intelligent Optimization, Embedded Systems.

Research and Projects

Networking Research Lab, University of Illinois Chicago

Priority-Aware Model-Distributed Inference at Edge Networks

FIRST AUTHOR · IEEE ICMLCN 2025

- Designed PA-MDI for **distributed inference** of LLM and computer vision models in multi-source edge networks, optimizing model partitioning and priority-aware allocation with resilience to node mobility and dynamic network changes.
- Benchmarked **ResNet on NVIDIA Jetson** and **GPT-2 on Colosseum**, achieving **reduced inference time** versus baselines.

REDIT: Resilient Distributed Text-to-Speech at Edge Networks

FIRST AUTHOR · IEEE GLOBECOM 2022

- Designed and implemented REDIT for distributed **text-to-speech (TTS)**, using text summarization as redundancy for **straggler resilience** across resource-constrained edge devices.
- Evaluated REDIT on **NVIDIA Jetson**, demonstrating **reduced completion delay** versus distributed TTS without resiliency.

Communication-Aware Model-Distributed Inference via LRC

Co-author · ACM MobiHoc 2026 (Accepted)

- We reduce communication overhead in model-distributed inference through dynamic activation compression using Top-k sparsification, quantization, and LLM.int8.
- I contributed **activation compressor** exploration and implementation, plus **T5 (Flan-T5) and ResNet** inference implementation, testing, and cross-dataset evaluation on **NVIDIA Jetson** to the project.
- We met **throughput targets** missed by uncompressed inference and achieved **higher accuracy** than baseline compression policies; our Jetson tests also **reduced inference delay** at comparable accuracy.

YOLO-Based CBRS Detection and Distributed Edge Inference (under submission)

FIRST AUTHOR

- Built a **YOLO-based object detection** pipeline for CBRS to identify and localize **radar/LTE signals** in spectrograms. Developed CEDAR-YOLO to reduce model size and activation-transfer volume.
- Implemented CEDAR-MDI to optimize model partitioning, node assignment, and activation compression across **NVIDIA Jetson** devices, achieving **higher throughput** than local execution while preserving radar/LTE overlap-detection quality.
- Separately explored local NVIDIA GPU deployment with **ONNX Runtime / TensorRT** workflows across model configurations, evaluating FP32/FP16 and calibrated INT8 against PyTorch on **latency, throughput, and detection accuracy**.

Open6G+AI Hackathon: 5G Multi-Service Scheduler

Northeastern University · Mar 2026

- Co-designed and implemented a **5G MAC scheduler** via Lua in a six-member team on a physical **O-RAN** testbed, balancing **high eMBB throughput** with **low latency and packet loss for URLLC UEs**.

Technical Skills

Programming	Python, C/C++, MATLAB, Java, Bash
ML & Deployment	PyTorch, TensorFlow, Hugging Face Transformers, ONNX / ONNX Runtime, TensorRT
Platforms & Tools	NVIDIA Jetson, Linux, Docker, Git, LaTeX, sockets, HTTP, Wi-Fi, O-RAN, OAI, SDR (USRP)
Languages	Mandarin Chinese (Native), English (Proficient)

Academic Service

- **Reviewer:** *IEEE MILCOM*, *IEEE Globecom*, *IEEE Transactions on Mobile Computing*.