Chengze Du

ducz.Monickar@gmail.com| https://Monickar.github.io

EDUCATION

Beijing University of Posts and Telecommunications(BUPT) Beijing, China

Bachelor of Science in Information Security, School of Cyberspace Security *Sep. 2021 – Jun. 2025*

- Average Scores: 87/100 (GPA: 3.56/4)
- The Second Prize Scholarship in 2023 **(top 15%)**

REARCH INTERESTS

Network Tomography; GPU scheduling and optimization; AI for Network Security

PUBLICATIONS & PATENTS

[1]Chengze Du, Zhiwei Yu, Heng Xu, Bo Liu, Jialong Li (2025). Temporal-Aware GPU Resource Allocation for Distributed LLM Inference via Reinforcement Learning. *Preprint on arXiv.*

[2]Chengze Du, Guangzhen Yao et al. (2025). GuidedLatent: Defending VAEs against Membership Inference Attacks via Distribution-Guided Privacy. *International Joint Conference on Neural Networks (IJCNN 2025).*

[3]Chengze Du, Jibin Shi, Hui Xu, Guangzhen Yao (2024). SecureNT: Smart Topology Obfuscation for Privacy-Aware Network Monitoring. *International Conference on Intelligent Computing (ICIC 2025).*

[4]Chengze Du, Zhiwei Yu, Xiangyu Wang (2024). Identification of Path Congestion Status for Network Performance Tomography using Deep Spatial-Temporal Learning. *Computer Communications 2025.*

EXPERIENCE

Shenzhen University of Advanced Technology (SUAT) Shenzhen, China

Research Internship *Apr. 2025 – Present*

- Led the design of TORTA, a two-layer reinforcement learning framework for temporal-aware GPU scheduling in distributed LLM inference. Integrated optimal transport and RL to reduce response time and operational cost.
- Built a scalable simulator to evaluate inference scheduling across real-world topologies. Achieved 15% latency reduction and 10–20% power cost savings over state-of-the-art baselines.

Zhipu AI, AI Department Beijing, China

Engineering Internship *Oct. 2024 – Jan. 2025*

- Contributed to internal AutoGLM tooling, improving dataset annotation reliability and model retraining efficiency within production pipelines.
- Designed a privacy-preserving method to defend variational autoencoders against membership inference attacks. Introduced distribution-guided latent control and two-phase privacy-aware training (published at IJCNN 2025).

Beijing University of Posts and Telecommunications (BUPT) Beijing, China

Undergraduate Research Assistant *Nov. 2023 – Jun. 2024*

- Proposed a novel congestion inference method using adversarial autoencoders (AAE) and LSTM for end-to-end network measurements. Introduced the concept of Additive Congestion Status (ACS) to quantify bottleneck severity (published in *Computer Communications*).
- Developed SecureNT, a topology obfuscation system that balances privacy and utility in network monitoring. Combined adversarial graph perturbation with traffic-awareness (accepted at ICIC 2025, oral).

SKILLS

Languages: Python(Pytorch, Keras, Pandas, SciPy, SkLearn, etc.), C/C++, Bash, Latex

Technologies/Frameworks:: NS-3 Network Simulator, Docker, Server Maintenance (Linux)

MISC

Interests: Passionate about running, with personal bests of 21:21 for 5km and 1:45:28 for the half marathon.