Zhifeng Jiang

AI & System Enthusiast

zjiangaj@connect.ust.hk
samuelgong.github.io
SamuelGong
ZR-W8XsAAAAJ

Research Interests

Efficient and trustworthy NLP, *especially on speech separation, large language model (LLM) inference, retrieval-augmented generation (RAG), and computer agents.*

Embodied intelligence, *continuously extending to vision-language-action (VLA) models, world models, and reinforcement learning (RL).*

Education

Sep 2019–Jul 2024 **Ph.D. in Computer Science and Engineering**,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong
Dissertation: "Towards Private and Efficient Cross-Device Federated Learning"

Jul 2018 – Oct 2018 **Research Intern in Prof. Dean Tullsen's Research Group**,
University of California, San Diego, La Jolla, US
- Defense against Return-Oriented Programming with Context-Sensitive Decoding on x86-64.

Sep 2015–Jun 2019 **B.Eng. in Computer Science**,
Zhejiang University, Hangzhou, China
GPA: 3.97/4.0, Graduated with Outstanding Honor (Zhejiang Province)

Experiences

Nov 2024 – Present **Research Scientist in Central Software Institute**,
Huawei Technologies, Shenzhen, China
- On-Device Audio Separation
  - Speaker separation with and without voiceprint prior
  - Target speech hearing with model-based beamforming
- Efficient Inference for LLMs
  - Decode in parallel for autoregressive LLMs
  - Accelerate diffusion LLMs (dLLMs)
- (L)LM Applications: RAG and Agents
  - On-device anti-fraud detection by semantics for web browsing
  - Large-scale RAG-based question answering
  - A general-purpose, command-line agent compatible with MCP (Hey)

Sep 2019 – Jul 2024 **PhD Candidate in Computer Science and Engineering**,
The Hong Kong University of Science and Technology, Kowloon, Hong Kong
- Private Inference for LLMs
  - Protect system prompt with minimized impact on conversational quality (PromptKeeper)
  - Replicate membership inference and data reconstruction attacks against LLMs (Tutorial)
- Private and Efficient Federated Training
  - Secure participant selection against adversarial servers (Lotto)
  - Enable dropout-resilient differential privacy without loss in time performance (Dordis)
  - Train asynchronously with principles for end-to-end speedup (Pisces)

Publications

Conference and Journal Publications

2026 Na Lyu, Jiayi Zhang, Zhi Shen, Chen Chen, **Zhifeng Jiang**, Quan Chen, Minyi Guo. "Enabling Client-Autonomous Training Optimizations for Efficient Federated Learning" , *accepted to appear in IEEE TPDS (IF: 9.2, top journal in High-Performance Computing)*

2025 **Zhifeng Jiang**, Zhihua Jin, Guoliang He. "PromptKeeper: Safeguarding System Prompts for LLMs" , *in the Findings of the Association for Computational Linguistics: EMNLP 2025 (acceptance rate: 39%)*

2024 **Zhifeng Jiang**, Peng Ye, Shiqi He, Wei Wang, Ruichuan Chen, Bo Li. "Lotto: Secure Participant Selection against Adversarial Servers in Federated Learning" , *in the Proc. of USENIX Security 2024 (acceptance rate: 17%)*

2024 **Zhifeng Jiang**, Wei Wang, Ruichuan Chen. "Dordis: Efficient Federated Learning with Dropout-Resilient Differential Privacy" , *in the Proc. of ACM EuroSys 2024 (acceptance ratio: 15%)*

2024 Peng Ye, **Zhifeng Jiang**, Wei Wang, Bo Li, Baochun Li. "Feature Reconstruction Attacks and Countermeasures of DNN Training in Vertical Federated Learning" , *accepted to appear in IEEE TDSC (IF: 7.5, top journal in Computer Security)*

2024 Yongkang Zhang, Haoxuan Yu, Chenxia Han, Cheng Wang, Baotong Lu, Yunzhe Li, **Zhifeng Jiang**, Yang Li, Xiaowen Chu, Huaicheng Li. "SGDRC: Software-Defined Dynamic Resource Control for Concurrent DNN Inference on NVIDIA GPUs" , *accepted to appear in ACM PPoPP 2025 (acceptance rate: 20%)*

2024 Na Lv, Zhi Shen, Chen Chen, **Zhifeng Jiang**, Jiayi Zhang, Quan Chen, Minyi Guo. "FedCA: Efficient Federated Learning with Client Autonomy" , *in the Proc. of ICPP 2024 (acceptance rate: 29%)*

2023 **Zhifeng Jiang**, Wei Wang, Bo Li, Qiang Yang. "Towards Efficient Synchronous Federated Training: A Survey on System Optimization Strategies" , *in IEEE TBD, Volume 9, Issue 2 (IF: 5.7. top journal in Big Data)*

2022 **Zhifeng Jiang**, Wei Wang, Baochun Li, Bo Li. "Pisces: Efficient Federated Learning via Guided Asynchronous Training" , *in the Proc. of ACM SoCC 2022 (acceptance ratio: 25%)*

2021 Minchen Yu, **Zhifeng Jiang**, Hok Chun Ng, Wei Wang, Ruichuan Chen, Bo Li. "Gillis: Serving Large Neural Networks in Serverless Functions with Automatic Model Partitioning" , *in the Proc. of IEEE ICDCS 2021 (acceptance ratio: 20%;* ***Best Paper Runner-Up****, 3 out of 97 accepted submissions)*

Manuscripts

2021 **Zhifeng Jiang**, Wei Wang, Yang Liu. "FLASHE: Additively Symmetric Homomorphic Encryption for Cross-Silo Federated Learning" , *in arXiv preprint (Citation: 87+)*

Honors and Awards

2024, 2023 Redbird Academic Excellence Award, HKUST

2021 Best Paper Runner-Up Award (Top 3 out of 97 accepted papers), IEEE ICDCS

2019 Outstanding Graduate Award (Top 1%), Zhejiang Province

2017 He Zhijun Scholarship (Top 10 in Dept. of CS), ZJU

2017 National Scholarship (Top 0.1% nationwide), Ministry of Education, China

Professional Service

Invited Reviewer IEEE Transactions on Mobile Computing, IEEE Transactions on Big Data, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Computers

Program Committee Shadow ACM EuroSys 2023.

AEC Member USENIX OSDI 2022, USENIX ATC 2022, ACM SOSP 2021.

Journal Sub-Reviewer IEEE Transactions on Network Science and Engineering, IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Big Data.

Conference Sub-Reviewer IEEE INFOCOM 2020-2024, IEEE ICDCS 2024, 2023 and 2021, IEEE/ACM IWQoS 2020-2021, IEEE WoWMoM 2021, IEEE ICNP 2020.

Teaching

Teaching Assistant

HKUST COMP3511 Operating System: Fall 2022, Fall 2020.
HKUST COMP4651 Cloud Computing: Fall 2021.
HKUST COMP4521 Mobile Application Development: Spring 2020.
ZJU Operating System (Educational Reform Class): Fall 2018.