Dongqi Cai (蔡栋琪)

PhD Student (Final Year)

Beijing University of Posts and Telecommunications, China

Email: dc912@cam.ac.uk

Homepage: http://www.caidongqi.com/

Education

09/2024 – present **Visiting PhD**, St John's College, University of Cambridge

- Advisor: Nicholas D. Lane

09/2021 – present **PhD in Computer Science and Technology**, BUPT

- Advisor: Shangguang Wang, Mengwei Xu
- Remote Advisor: Felix Xiaozhu Lin

09/2017 – 07/2021 **BS in Communication Engineering**, BUPT

- Advisor: Lin Fan

Intership

07/2021 – 12/2021 **Research Intern**, WeBank

- Mentor: Lixin Fan

Honors & Awards

- MobiSys Rising Star, SigMobile, 2025
- Young Elite Scientists Sponsorship (PhD student Special Program), CAST, 2025
- National Scholarship, Ministry of Education, 2024
- Distinguished Artifact Nomination (~9 out of 494 submission, ~1.8%), MobiCom, 2024,
- St John's College Fellow-Sponsored Member, University of Cambridge, 2024
- CSC Scholarship, China Scholarship Council, 2024
- Travel Grant, NeurIPS'24/EuroSys'24/MobiSys'24/ATC'24/MobiSys'25/MobiUK'25
- National Scholarship, Ministry of Education, 2023
- Outstanding Graduate Student, BUPT, 2023
- Excellent Ph.D. Students Foundation, BUPT, 2023
- Outstanding Graduate Student, State Key Laboratory of Networking and Switching Technology, 2022/2023

Conference Publications (* = equal contributions; # = corresponding)

(full list at https://scholar.google.com/citations?user=dIimkboAAAAJ&hl=zh-CN)

[C12] "Demystifying Small Language Models for Edge Deployment"

Zhenyan Lu, Xiang Li, **Dongqi Cai**, Rongjie Yi, Fangming Liu, Wei Liu, Jian Luan, Xiwen Zhang, Nicholas D. Lane, Mengwei Xu, in *The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL, CCF-A**)*, 2025.

[C11] "SystemX: Federated LLM Pre-Training"

Lorenzo Sani, Alex Iacob, Zeyu Cao, Royson Lee, Bill Marino, Yan Gao, Wanru Zhao, **Dongqi Cai**, Zexi Li, Xinchi Qiu, Nicholas Donald Lane, in the *Eighth Annual Conference on Machine Learning and Systems (**MLSys**)*, 2025.

[C10] "DEPT: Decoupled Embeddings for Pre-training Language Models"

Alex Iacob, Lorenzo Sani, Meghdad Kurmanji, William F. Shen, Xinchi Qiu, **Dongqi Cai**, Yan Gao, Nicholas Donald Lane, in the *Thirteenth International Conference on Learning Representations (**ICLR, [Oral, top 1.8%]**)*, 2025.

[C9] "ShortcutsBench: A Large-Scale Real-world Benchmark for API-based Agents"

Haiyang Shen, Yue Li, Desong Meng, **Dongqi Cai**, Sheng Qi, Li Zhang, Mengwei Xu, Yun Ma, in the *Thirteenth International Conference on Learning Representations (**ICLR**)*, 2025.

[C8] "SILENCE: Protecting privacy in offloaded speech understanding on wimpy devices"

Dongqi Cai, Shangguang Wang, Zeling Zhang, Felix Xiaozhu Lin, Mengwei Xu, in *the Annual Conference on Neural Information Processing Systems (**NeurIPS, CCF-A**),* 2024.

[C7] "FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences"

Mengwei Xu (My advisor), **Dongqi Cai#**, Yaozong Wu, Xiang Li, Shangguang Wang, in *USENIX Annual Technical Conference (**USENIX ATC, CCF-A**),* 2024.

[C6] "Mobile Foundation Model as Firmware"

Jinliang Yuan*, Chen Yang*, **Dongqi Cai***, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, Shangguang Wang, Mengwei Xu, in *Proc. ACM Int. Conf. Mobile Computing and Networking (**MobiCom, CCF-A, [Distinguished Artifact Nomination, ~1.8%]**)*, 2024.

[C5] "Federated Few-shot Learning for Mobile NLP"

Dongqi Cai, Shangguang Wang, Yaozong Wu, Felix Xiaozhu Lin, Mengwei Xu, in *Proc. ACM Int. Conf. Mobile Computing and Networking (**MobiCom, CCF-A**)*, 2023.

[C4] "Efficient Federated Learning for Modern NLP"

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, Mengwei Xu, in *Proc. ACM Int. Conf. Mobile Computing and Networking (**MobiCom, CCF-A**)*, 2023.

[C3] "GPT4D: Automatic Cross-Version Linux Driver Upgrade Toolkit"

Borui Yang, Hongyu Li, **Dongqi Cai**, in *the 8th EAI International Conference on Machine Learning and Intelligent Communications (MLICOM)*, 2023.

[C2] “FedAdapter: Efficient Federated Learning for Mobile NLP”

Dongqi Cai, Shangguang Wang, Yaozong Wu, Mengwei Xu, in *Proceedings of the ACM Turing Award Celebration Conference (TURC)*, 2023.

[C1] “Mitigating App Collusion using Machine Learning”

Xuefei Duan, Hua Lu, Jinliang Yuan, Qiyang Zhang, **Dongqi Cai**, in *IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom)*, 2021.

Journal Publications (* = equal contributions)

[J5] “Ubiquitous Memory Augmentation via Mobile Multimodal Embedding System”

Dongqi Cai, Shangguang Wang, Chen Peng, Zeling Zhang, Zhenyan Lu, Tao Qi, Nicholas D. Lane, Mengwei Xu, accepted by ***Nature Communications***, 2025.

[J4] “Efficient and Privacy-Preserving Spoken Language Understanding for Resource-Constrained Microcontroller Unit”

Dongqi Cai, Shangguang Wang, Zeling Zhang, Xiao Ma, Mengwei Xu, accepted by ***Chinese Journal of Electronics (CCF-A Chinese Journal)***, 2025.

[J3] “Resource-efficient Algorithms and Systems of Foundation Models: A Survey”

Mengwei Xu* (My advisor), **Dongqi Cai***, Wangsong Yin*, Shangguang Wang, Xin Jin, Xuanzhe Liu, accepted in *ACM Computing Surveys (**ACM CSUR, Impact Factor: 23.8，ranked 1/143 in Computer Science Theory & Methods**)*, 2024.

[J2] “Accelerating Vertical Federated Learning”

Dongqi Cai, Tao Fan, Yan Kang, Lixin Fan, Mengwei XU, Shangguang Wang, Qiang Yang, e in *IEEE Transactions on Big Data (**IEEE TBD**)*, 2024.

[J1] “Implementation of an E-payment security evaluation system based on quantum blind computing”

Dongqi Cai, Xi Chen, Yuhong Han, Xin Yi, Jinping Jia, Cong Cao, Ling Fan, in *International Journal of Theoretical Physics (IJTP),* 2020.

Workshop Publications (* = equal contributions)

[W4] “Large Language Models on Mobile Devices: Measurements, Analysis, and Insights”

Xiang Li, Zhenyan Lu, **Dongqi Cai**, Xiao Ma, Mengwei Xu, in *Proceedings of the Workshop on Edge and Mobile Foundation Models (EdgeFM), co-located with ACM International Conference on Mobile Systems, Applications, and Services (**MobiSys, CCF-B**),* 2024.

[W3] “FedRDMA: Communication-Efficient Cross-Silo Federated LLM via Chunked RDMA Transmission”

Zeling Zhang*, **Dongqi Cai***, Yiran Zhang, Mengwei Xu, Shangguang Wang, Ao Zhou, in *Proceedings of the 4rd Workshop on Machine Learning and Systems (EuroMLSys), co-located with European Conference on Computer Systems (**EuroSys, CCF-A**),* 2024.

[W2] “Towards Practical Few-shot Federated NLP”

Dongqi Cai, Yaozong Wu, Haitao Yuan, Shangguang Wang, Felix Xiaozhu Lin, Mengwei Xu, in *Proceedings of the 3rd Workshop on Machine Learning and Systems (EuroMLSys), co-located with European Conference on Computer Systems (**EuroSys, CCF-A**),* 2023.

[W1] “Towards ubiquitous learning: A first measurement of on-device training performance”

Dongqi Cai, Qipeng Wang, Yuanqiang Liu, Yunxin Liu, Shangguang Wang, Mengwei Xu, in *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning* (*EMDL*), *co-located with ACM International Conference on Mobile Systems, Applications, and Services (**MobiSys, CCF-B**),* 2021.

Teaching Experience

- Teaching Assistant, Principles of Machine Learning Systems, University of Cambridge (Michaelmas Term 2024)

Academic Services

- **TPC Member**

MobiSys’24 AE, MobiCom’24 AE, NCSC-edge'22, TURC-SIGBED-China'23

- **Reviewer**

Scientific Reports, TSC, TMC, TKDE, TECS, IoTJ, SAGC'22, ICASSP’24, ICASSP’25.

- **External Reviewer**

MLSys’25, ICWS’24, IEEE EDGE’24, IEEE EDGE’23, ICWS'23, EIS'21

Patents

[P4] "A Federated Learning Method, System, and Apparatus Based on Forward Gradient"

Mengwei Xu; Yaozong Wu; **Dongqi Cai;** Shangguang Wang

[P3] "A Federated Few-Shot Learning Method, System, and Device for Natural Language Models"

Mengwei Xu; **Dongqi Cai**; Ao Zhou; Xiao Ma; Shangguang Wang

[P2] "A Federated Learning Method, Device, and System for Pre-trained Models"

Mengwei Xu; **Dongqi Cai**; Ao Zhou; Xiao Ma; Shangguang Wang,

[P1] "Vertical Federated Learning Modeling Optimization Method, Device, Medium, and Program”

Dongqi Cai; Lixin Fan; Qiang Yang

Projects

1. 校企合作（小米集团），端侧大模型的个性化高效微调关键技术研究，2024.09–2025.09，0.18M，在研，项目骨干（项目申报、技术研究）

2. 创新基金（北京邮电大学），面向复杂自然语言模型的联邦小样本学习方法研究，2023.4-2024.04，0.012M，已结题，项目负责人（独立 PI）
3. 校企合作（微众银行），可信联邦学习算法研究及应用 - 可信联邦大模型研究，2023.09-2024.09，0.2M，已结题，项目骨干（项目申报、技术研究、系统集成开发、验收结项）
4. 国家重点研发计划项目（科技部），面向大规模分布式人工智能应用的关键网络技术研究，2020.07-2024.01，20M，已结题，项目骨干（技术研究、系统集成开发、验收结项）
5. 国家重点研发计划项目（科技部），跨域异质分布式学习和推理系统，2021.08-2024.12，75M，已结题，项目骨干（项目申报、技术研究、系统集成开发、验收结项）

Invited Talk

- EMDL'21 (Co-located with MobiSys'21), Towards ubiquitous learning: A first measurement of on-device training performance, Online, 2021/06/25
- EuroMLSys'23 (Co-located with EuroSys'23), Towards Practical Few-shot Federated NLP Rome, Italy, 2023/05/08
- MobiCom'23, Efficient Federated Learning for Modern NLP, Madrid, Spain, 2023/10/05
- MobiCom'23, Federated Few-shot Learning for Mobile NLP, Madrid, Spain, 2023/10/05
- Northwestern Polytechnical University, PhD Research Methodology, Online, 2023/10/30
- BUPT 'Diligent Research, Academic Leadership' Academic Forum, Efficient Federated Learning for Modern NLP, Beijing, China, 2023/12/26
- EuroMLSys'24 (Co-located with EuroSys'24), FedRDMA: Communication-Efficient Cross-Silo Federated LLM via Chunked RDMA Transmission, Athens, Greece, 2024/04/22
- MoiSys'24 N2Women, Large Language Models on Mobile Devices: Measurements, Analysis, and Insights, Tokyo, Japan, 2024/06/03
- EdgeFM'24 (Co-located with MobiSys'24), Large Language Models on Mobile Devices: Measurements, Analysis, and Insights, Tokyo, Japan, 2024/06/07
- USENIX ATC'24, FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences, SANTA CLARA, CA, USA, 2024/07/11
- AI TIME NeurIPS 2024 Forum, SILENCE: Protecting Privacy in Offloaded Speech Understanding on Resource-constrained Devices, Online, 2024/11/20
- NeurIPS'24, SILENCE: Protecting Privacy in Offloaded Speech Understanding on Resource-constrained Devices, Vancouver, Canada, 2024/12/11
- CCF Talk, Efficient Federated Learning System for LLMs, Online, 2024/12/22
- Cambridge ML Systems Seminar Series, Efficient Machine Learning System, Cambridge, UK, 2025/1/28