蔡栋琪

博士研究生（预计毕业时间：2025 年）

北京邮电大学丨剑桥大学

联系方式：手机号/微信（13261808588）；邮箱（dc912@cam.ac.uk）

个人主页：http://www.caidongqi.com/

研究方向

端侧大模型系统优化：高效联邦大模型训练、多模态大模型推理加速等

教育经历

09/2024 至今　**联合培养博士生**，剑桥大学 圣约翰学院

- 合作导师：Nicholas D. Lane

09/2021 至今　**博士研究生** 计算机科学与技术专业，北京邮电大学

- 导师：王尚广
- 指导老师：徐梦炜
- 远程导师： Felix Xiaozhu Lin

09/2017 - 06/2021　**学士** 通信工程专业，北京邮电大学

实习经历

07/2021 - 12/2021.　**算法实习生**, 微众银行

- 企业导师：范力欣
- 部门负责人：杨强

奖项与荣誉

- 首届中国科协青年人才托举工程（博士生特别项目），2024
- Best Poster Award, MobiUK, 2025
- MobiSys Rising Star，SigMobile，2025
- 国家奖学金，教育部，2023/2024（连续两年）
- Distinguished Artifact 提名（494 篇投稿中约 9 篇入选，约 1.8%），MobiCom，2024
- 剑桥大学圣约翰学院 院士赞助学员，2024
- 国家留学基金委员会（CSC）奖学金，2024
- Travel Grant, NeurIPS'24/EuroSys'24/MobiSys'24/ATC'24/MobiSys'25/MobiUK'25
- 北京邮电大学优秀研究生，2023
- 网络与交换技术国家重点实验室优秀研究生，2022/2023

学术研究与评价

我的博士研究方向为面向移动终端的大模型高效个性化方法：首先，针对端侧多模态数据统一表征难题，设计了基于粗粒度早停机制的多模态大模型推理加速系统，在真实场景中最高提升吞吐率 45 倍、降低能耗 3 倍以上；其次，面向移动终端普遍存在的少标注场景，提出了基于自监督提示学习的高效小样本学习框架，在可控学习开销下，将标注需求降低三个数量级，并保持可用准确率；最后，针对跨终端分布式训练中的通信和隐私挑战，提出了一套基于参数高效适配器的联邦大模型微调系统，在七十亿级参数模型上实现分钟级微调时延，最高提升收敛效率 112 倍。

以（共同）第一作者/通讯作者发表或接收论文 15 篇，其中包括 1 篇 Nature Communications, 5 篇 CCF-A 类英文会议和 1 篇 CCF-A 类中文期刊。相关工作已被应用于剑桥大学 Flower 框架、微众银行 FATE 框架和烽火 RDMA 智能网卡，获谷歌学术引用超 600 次，被图灵奖得主 David Patterson 在其 Commun. ACM '24 论文中评价为“专注于移动端的资源效率问题，发现了移动端训练推理和数据中心内的巨大差异”。

期刊论文 (* = 同等贡献)

[J1] “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, ***Nature Communications (Nature 子刊)***, 2025.

[J2] “面向微控制单元的高效语音隐私保护编码器”

蔡栋琪, 王尚广, 张泽凌, 马骁, 徐梦炜, ***电子学报 (CCF-A 中文期刊)***, 2025.

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

Mengwei Xu* (指导老师), **Dongqi Cai***, Wangsong Yin*, Shangguang Wang, Xin Jin, Xuanzhe Liu, accepted in *ACM Computing Surveys (**ACM CSUR, 中科院一区**)*, 2024.

[J4] “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，中科院二区**)*, 2024.

[J5] “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, SCI),* 2020.

部分会议论文 (* = 同等贡献; # = 通讯作者)

[C1] “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.

[C2] “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.

[C3] “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.

[C4] “FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences”

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

[C5] “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.

[C6] “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.

[C7] “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.

[C8] “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.

[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.

Workshop 论文(* = 同等贡献)

[W1] “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.

[W2] “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.

[W3] “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.

[W4] “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.

专利

[P1] 纵向联邦学习建模优化方法、设备、介质及程序产品。**蔡栋琪**，范力欣，杨强

[P2] 一种面向预训练模型的联邦学习方法、装置及系统。徐梦炜，**蔡栋琪**，周傲，马骁，王尚广

[P3] 面向自然语言模型的联邦小样本学习方法、系统及设备。徐梦炜，**蔡栋琪**，周傲，马骁，王尚广

[P4] 纵向联邦学习建模优化方法、设备、介质及程序产品。徐梦炜，武耀宗，**蔡栋琪**，王尚广

学术服务

- **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

教学经历

- 助教，机器学习系统原理，剑桥大学，2024

参与项目

1. 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）

2. 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）
3. 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）
4. 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）
5. 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）

受邀汇报/讲座

- 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, Training LLMs Anywhere: Enabling Large-Scale Decentralized Learning on Your Mobiles Devices, Cambridge, UK, 2025/1/28
- Department of Computer Science and Technology, Efficient Machine Learning System for Mobile Devices, Soochow University, China, 2025/04/22