蔡栋琪 (Dongqi Cai)

博士生（毕业年）

北京邮电大学，中国

邮箱: dc912@cam.ac.uk

主页: http://www.caidongqi.com/

研究兴趣

- 联邦学习
- 高效 NLP 系统
- 语音隐私

教育背景

2024 年 9 月 - 至今

访问博士研究生，剑桥大学 (University of Cambridge)

- 导师：Nicholas D. Lane
- St John’s College

2021 年 9 月 - 至今

博士研究生，计算机科学与技术专业，北京邮电大学 (BUPT)

- 导师：王尚广 (Shangguang Wang)，许梦薇 (Mengwei Xu)
- 远程导师：Felix Xiaozhu Lin（弗吉尼亚大学）

2017 年 9 月 - 2021 年 7 月

学士，通信工程，北京邮电大学 (BUPT)

- 导师：樊林 (Lin Fan)

实习经历

2021 年 7 月 - 2021 年 12 月

研究实习生，微众银行 (WeBank)

- 导师：樊力新 (Lixin Fan)

荣誉与奖励

- 2025 年 中国科协青年人才托举工程（博士研究生专项）
- 2024 年 国家奖学金（教育部）
- 2024 年 MobiCom (CCF-A) “杰出 Artifact 提名” （494 篇投稿仅约 9 篇，~1.8%）
- 2024 年 剑桥大学 St John’s College Fellow-Sponsored Member
- 2024 年 NeurIPS (CCF-A) 奖学金
- 2024 年 中国国家留学基金委奖学金（CSC Scholarship）
- 2024 年 EuroSys / MobiSys / ATC 旅费资助
- 2023 年 国家奖学金（教育部）
- 2023 年 北京邮电大学“优秀研究生”
- 2023 年 北京邮电大学“优秀博士生专项基金”
- 2022/2023 年 国家重点实验室（网络与交换技术）“优秀研究生”

学术服务

• **TPC 委员**

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

• **审稿人**

• 期刊：Scientific Reports, TSC, TMC, TKDE, TECS, IoTJ

• 会议/研讨会：SAGC’22, ICASSP’24, ICASSP’25

• **外部审稿人**

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

会议论文（* 表示共同贡献；# 表示通讯作者）

完整列表详见 Google Scholar

1. **[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,

发表在 第八届机器学习与系统大会（MLSys 2025，仍在 shepherding 阶段）。

2. **[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,

发表在 第十三届国际学习表征大会（ICLR 2025，[Oral, 前 1.8%]）。

3. **[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,

发表在 第十三届国际学习表征大会（ICLR 2025）。

4. [C8] “SILENCE: Protecting privacy in offloaded speech understanding on wimpy devices”

Dongqi Cai, Shangguang Wang, Zeling Zhang, Felix Xiaozhu Lin, Mengwei Xu,

发表在 神经信息处理系统大会（NeurIPS 2024，CCF-A）。

5. [C7] “FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences”

Mengwei Xu (导师), **Dongqi Cai#**, Yaozong Wu, Xiang Li, Shangguang Wang,

发表在 USENIX Annual Technical Conference (USENIX ATC 2024，CCF-A)。

6. [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,

发表在 ACM 国际移动计算与网络大会（MobiCom 2024，CCF-A），获“Distinguished Artifact Nomination”（~1.8%）。

7. [C5] “Federated Few-shot Learning for Mobile NLP”

Dongqi Cai, Shangguang Wang, Yaozong Wu, Felix Xiaozhu Lin, Mengwei Xu,

发表在 ACM 国际移动计算与网络大会（MobiCom 2023，CCF-A）。

8. [C4] “Efficient Federated Learning for Modern NLP”

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, Mengwei Xu,

发表在 ACM 国际移动计算与网络大会（MobiCom 2023，CCF-A）。

9. [C3] “GPT4D: Automatic Cross-Version Linux Driver Upgrade Toolkit”

Borui Yang, Hongyu Li, **Dongqi Cai**,

发表在 第 8 届 EAI 机器学习与智能通信国际会议（MLICOM 2023）。

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

Dongqi Cai, Shangguang Wang, Yaozong Wu, Mengwei Xu,

发表在 ACM 图灵大会(TURC) 2023。

11. [C1] “Mitigating App Collusion using Machine Learning”

Xuefei Duan, Hua Lu, Jinliang Yuan, Qiyang Zhang, **Dongqi Cai**,

发表在 IEEE 第 7 届大数据智能与计算国际会议（DataCom 2021）。

期刊论文（* 表示共同贡献）

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

Mengwei Xu* (导师), **Dongqi Cai***, Wangsong Yin*, Shangguang Wang, Xin Jin, Xuanzhe Liu,

录用于 ACM Computing Surveys (ACM CSUR，影响因子 23.8，计算机科学理论与方法方向排名 1/143)，2024。

2. [J2] “Accelerating Vertical Federated Learning”

Dongqi Cai, Tao Fan, Yan Kang, Lixin Fan, Mengwei Xu, Shangguang Wang, Qiang Yang,

发表在 IEEE Transactions on Big Data (IEEE TBD), 2024。

3. [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,

发表在 International Journal of Theoretical Physics (IJTP), 2020。

研讨会论文（* 表示共同贡献）

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

Xiang Li, Zhenyan Lu, **Dongqi Cai**, Xiao Ma, Mengwei Xu,

发表在 EdgeFM（与 MobiSys 2024 合办的 Workshop），CCF-B，2024。

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

Zeling Zhang*, **Dongqi Cai***, Yiran Zhang, Mengwei Xu, Shangguang Wang, Ao Zhou,

发表在第 4 届机器学习与系统研讨会（EuroMLSys 2024，和 EuroSys 2024 合办，CCF-A）。

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

Dongqi Cai, Yaozong Wu, Haitao Yuan, Shangguang Wang, Felix Xiaozhu Lin, Mengwei Xu,

发表在第 3 届机器学习与系统研讨会（EuroMLSys 2023，和 EuroSys 2023 合办，CCF-A）。

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

Dongqi Cai, Qipeng Wang, Yuanqiang Liu, Yunxin Liu, Shangguang Wang, Mengwei Xu,

发表在第 5 届嵌入式与移动深度学习国际研讨会（EMDL 2021，和 MobiSys 2021 合办，CCF-B）。

专利

1. [P4] “A Federated Learning Method, System, and Apparatus Based on Forward Gradient”

徐梦薇，吴耀宗，**蔡栋琪**，王尚广

2. [P3] “A Federated Few-Shot Learning Method, System, and Device for Natural Language Models”

徐梦薇，**蔡栋琪**，周鳌，马潇，王尚广

3. [P2] “A Federated Learning Method, Device, and System for Pre-trained Models”

徐梦薇, **蔡栋琪**, 周鳌, 马潇, 王尚广

4. [P1] “Vertical Federated Learning Modeling Optimization Method, Device, Medium, and Program”

蔡栋琪, 樊力新, 杨强

教学经历

• 2024 年 剑桥大学 机器学习系统原理（Michaelmas 学期），助教

受邀报告

• EMDL’21 (MobiSys’21 附属研讨会)

“Towards ubiquitous learning: A first measurement of on-device training performance,” 在线，2021/06/25

• EuroMLSys’23 (EuroSys’23 附属研讨会)

“Towards Practical Few-shot Federated NLP,” 罗马，意大利，2023/05/08

• MobiCom’23

• “Efficient Federated Learning for Modern NLP,” 马德里，西班牙，2023/10/05

• “Federated Few-shot Learning for Mobile NLP,” 马德里，西班牙，2023/10/05

• 西北工业大学博士研究方法课程

在线，2023/10/30

• 北京邮电大学 ‘勤奋治学·学术引领’ 学术论坛

“Efficient Federated Learning for Modern NLP,” 北京，2023/12/26

• EuroMLSys’24 (EuroSys’24 附属研讨会)

“FedRDMA: Communication-Efficient Cross-Silo Federated LLM via Chunked RDMA Transmission,” 雅典，希腊，2024/04/22

• MoiSys’24 N2Women

“Large Language Models on Mobile Devices: Measurements, Analysis, and Insights,” 东京，日本，2024/06/03

• EdgeFM’24 (MobiSys’24 附属研讨会)

“Large Language Models on Mobile Devices: Measurements, Analysis, and Insights,” 东京，日本，2024/06/07

• USENIX ATC’24

“FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences,” 圣克拉拉，美国，2024/07/11

• AI TIME NeurIPS 2024 Forum

“SILENCE: Protecting Privacy in Offloaded Speech Understanding on Resource-constrained Devices,” 在线，2024/11/20

• NeurIPS’24

“SILENCE: Protecting Privacy in Offloaded Speech Understanding on Resource-constrained Devices,” 温哥华，加拿大，2024/12/11

• CCF Talk

“Efficient Federated Learning System for LLMs,” 在线，2024/12/22

• 剑桥大学 机器学习系统研讨会

“Efficient Machine Learning System,” 剑桥，英国，2025/1/28