WENZHI FANG

West Lafayette, IN 765-694-5334 fang375@purdue.edu https://wenzhifang.github.io Google Scholar

RESEARCH BACKGROUND

Ph.D. researcher in large language models (LLMs), reinforcement learning, and distributed optimization. My work focuses on LLM post-training and reasoning, with an emphasis on RL-based alignment, agent collaboration, and communication-efficient fine-tuning for distributed on-device LLM systems.

EDUCATION

Purdue University *West Lafayette, IN, US*
Elmore Family School of Electrical and Computer Engineering
Ph. D. Candidate in Electrical and Computer Engineering *Aug. 2023 – 2027*

ShanghaiTech University *Shanghai, China*
School of Information and Science Technology
M.S. in Communication and Information Systems *Sept. 2020 – Jul. 2023*

Shanghai University *Shanghai, China*
School of Communication and Information Engineering
B.S. in Communication Engineering *Sept. 2016 – Jul. 2020*

MAJOR COURSES

- Machine Learning, Deep Learning, Reinforcement Learning
- Bayesian Data Analysis, Convex Optimization, Matrix Computation

RESEARCH INTERESTS

- **Large Language Model.** Efficient Fine-Tuning of LLM & RL-based Post-training, LLM Agents
- **Machine Learning and Optimization.** Federated Learning & Distributed Optimization

WORKING EXPERIENCE

ION Lab ***Aug., 2023 - Present***

PhD Research Assistant *Advisor: Prof. Christopher G. Brinton* Purdue University

- **Collaborative Device-cloud LLMs Reasoning**, *May. 2025 – Sept. 2025*
  We propose a unified reinforcement learning framework that empowers on-device LLMs to autonomously learn both problem-solving reasoning and routing policies, deciding when to process queries locally or to offload them to cloud LLMs. Our method removes the need for separate routing classifiers.
- **Mitigating Catastrophic Forgetting of On-device LLM via Cloud Offloading**, *Oct. 2025 – Feb. 2026*
  We propose a constraint-aware RL method that embeds cloud-usage regulation into advantage computation, enabling stable, budget-controlled local–cloud LLM collaboration and reduced forgetting under continual learning without per-task reward tuning.
- **Multi-modal LLM Fine-tuning**, *May. 2025 – Sept. 2025*
  we propose two-stage adaptive personalization (TAP), a federated learning framework that personalizes multi-task and multi-modal LLMs by integrating selective local adaptation with post-FL knowledge distillation to balance task-specific performance and generalization.
- **On-device LLM Fine-tuning**, *Oct. 2024 – Feb. 2025*
  We propose sketching LoRA, a theoretically-grounded parameter efficient fine-tuning methodology that retains LoRA's flexibility while adapting to the communication and computational abilities of individual devices.

- **Distributed Optimization for Fog Learning**, *Jan. 2024 – June. 2024*
  We proposed an algorithm to address multi-level data heterogeneity in hierarchical federated learning, a.k.a. fog learning, deriving strong theoretical results without relying on additional data heterogeneity assumptions. This work fills a critical gap in the existing federated learning literature.

Machine Learning and Optimization Lab ***Aug., 2022 - Feb. 2023***

Summer Intern *Advisor: Prof. Peter Richtarik* KAUST

- **Primal-Dual Optimization**, *Aug., 2022 - Feb. 2023*
  We studied a primal-dual hybrid gradient descent algorithm and explored variance reduced technique for primal dual optimization.

Optimization for Communication Lab ***Sept., 2020 - June. 2022***

Master Research Assistant *Advisor: Prof. Yong Zhou* ShanghaiTech

- **Zeroth-order Optimization for Federated Learning**, *Jul. 2021 – Jan. 2022*
  We proposed a federated zeroth-order algorithm (FedZO) with a convergence guarantee. This algorithm makes the training process forward-only, eliminating the memory overhead of backward propagation, which has been widely applied in LLM related research, e.g., LLM attack, prompt tuning, and efficient model fine-tuning.

ACEDEMIC SERVICE

Reviewer of top tier conferences including NeurIPS, ICML, ICLR, AISTAT, AAAI

Reviewer of top tier journals including IEEE/ACM Transactions On Networking; IEEE Transactions on Signal Processing (TSP); IEEE Transactions On Mobile Computing (TMC); Transactions on Machine Learning Research (TMLR)

SELECTED PUBLICATIONS

- Evan Chen*, **W. Fang***, S. Wang, L. Yuan, and C. G. Brinton, "Joint Continual Learning of Local Language Models and Cloud Offloading Decisions with Budget Constraints", 2026, *Under Review* [pdf]
- **W. Fang**, D-J. Han, L. Yuan, and C. G. Brinton, "Bridging On-Device and Cloud LLMs for Collaborative Reasoning", 2025, *Under Review* [pdf]
- **W. Fang**, D-J. Han, L. Yuan, S. Hosseinalipour, and C. G. Brinton, "Federated Sketching LoRA: On-Device Collaborative Fine-Tuning of Large Language Models", 2025, *Under Review* [pdf]
- J. Lee, **W. Fang**, D-J. Han, S. Hosseinalipour, and C. G. Brinton, "TAP: Two-Stage Adaptive Personalization of Multi-task and Multi-Modal Foundation Models in Federated Learning", 2025, *Under Review* [pdf]
- **W. Fang**, D-J. Han, and C. G. Brinton, "Federated Learning over Hierarchical Wireless Networks: Training Latency Minimization via Submodel Partitioning", *IEEE/ACM Transactions On Networking (ToN), 2025* [pdf]
- Y. Zhou, **W. Fang**, Y. Shi, and K. B. Letaief "Federated Edge Learning: Algorithms, Architectures and Trustworthiness", *Springer Nature, 2025* [pdf]
- **W. Fang**, D-J. Han, E, Chen, S. Wang, and C. G. Brinton, "Hierarchical Federated Learning with Multi-Timescale Gradient Correction", *Neural Information Processing Systems (NeurIPS), 2024.* [pdf]
- **W. Fang**, Z. Yu, Y. Jiang, Y. Shi, C. Jones, and Y. Zhou, "Communication-Efficient Stochastic Zeroth-Order Optimization for Federated Learning", *IEEE Transactions on Signal Processing (TSP), 2022* [pdf]

TECHNICAL SKILLS

Languages	Python, C, C++, Git, Matlab, LaTex
Tools	PyTorch, Tensorflow, vLLM, VERL