Yiping Li

Email: yipingli@ucmerced.edu; yiping11024@gmail.com | Phone: 206-302-8770 | GitHub | Google Scholar

RESEARCH INTERESTS

I study how knowledge is represented inside models, and how it moves and combines across them. The same knowledge takes different carriers in different settings, appearing as gradients in federated learning, as KV caches in multi-agent LLM systems, and as model weights when specialized models are merged. Across these settings the recurring obstacle is conflict, where knowledge from different sources interferes once combined. My work so far spans KV-cache compression for latent multi-agent LLMs, knowledge distillation with curriculum learning, and conflict-aware federated learning under non-IID data. I am currently working on cross-family on-policy distillation, where a student must re-express what the teacher holds in its own internal states, and on conflict in model steering, merging, and fine-tuning.

EDUCATION

University of California, Merced, Merced, CA 08/2025 – Expected 05/2029

- Ph.D. in **Electrical Engineering and Computer Science** (Advisor: Dr. Wan Du)

University of Illinois Urbana-Champaign, Champaign, IL 08/2022 – 12/2023

- Master of **Computer Science**

University of Washington, Seattle, WA 10/2017 – 06/2022

- B.S. in **Applied & Computational Mathematical Sciences (ACMS)** & **Psychology**

SKILLS

Research areas: on-policy distillation (OPD), KV-cache compression, LLM inference efficiency, model merging and steering, LoRA/PEFT, RL post-training (DPO, GRPO), RAG, federated learning, knowledge distillation

Languages & tools: Python, Java, SQL; PyTorch (DDP), Hugging Face Transformers / Accelerate, TensorFlow, verl, Unsloth, NumPy, Pandas, scikit-learn; AWS, Linux, Git

PUBLICATIONS

- **Li Y**, An Z, Du W. When Less Latent Leads to Better Relay: Information-Preserving Compression for Latent Multi-Agent LLM Collaboration. *Under review, NeurIPS 2026. arXiv: arxiv.org/abs/2604.13349.* Code: github.com/markli404/When-Less-Latent-Leads-to-Better-Relay
- Li J, Zhang Y, **Li Y**, Gong X, Wang W. FedSparse: A Communication-Efficient Federated Learning Framework Based on Sparse Updates. *Electronics*. 2024; 13(24):5042.
- Li, J., Zhang, Y., **Li, Y.**, Gong, X., & Wang, W. (2024, June). HybridCom: Improve Federated Learning Efficiency on Unstable Data. In *ICC 2024-IEEE International Conference on Communications* (pp. 2622-2627). IEEE.
- Li, J., Zhang, Y., **Li, Y.**, Gong, X., & Wang, W. (2023, October). Gradient Calibration for Non-IID Federated Learning. In *Proceedings of the 2nd ACM Workshop on Data Privacy and Federated Learning Technologies for Mobile Edge Network* (pp. 119-124).
- Li, J., Zhang, Y., **Li, Y.**, Gong, X., & Wang, W. (2022, December). DCD: A New Framework for Distillation Learning with Dynamic Curriculum. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)* (pp. 928-935). IEEE.

RESEARCH EXPERIENCE

University of California, Merced 08/2025 – Present

Research Assistant, Advised by Dr. Wan Du

Efficient Communication in Latent Multi-Agent LLM Systems

- Observed that as latent multi-agent LLMs move from text to embeddings to full KV caches, bandwidth grows while information density falls, so compressing the relayed KV is both necessary and effective.
- Introduced the first adaptation of KV-cache eviction to multi-agent latent communication, and defined a role decomposition of the relayed cache that makes eviction interpretable.

- Proposed Orthogonal BackFill (OBF): inject a low-rank component of the discarded value states, orthogonal to the retained ones, back into the retained KV, recovering complementary information otherwise lost to hard KV eviction.
- Showed across nine benchmarks that compressed relay matches or beats full KV relay on five (math, expert QA, coding) while keeping only 9.9–20.2% of prompt KV; first-author, under review at NeurIPS 2026.

Joint Laboratory of BUPT and ChuangCache 08/2021 – 03/2025

Research Assistant, Advised by Dr. Yuchao Zhang

Across the four publications below, designed the core algorithms, wrote nearly all the code, and ran every experiment; also built the lab's shared FL testing framework that all four projects ran on.

Communication Efficiency in Federated Learning

- Designed a Resource Optimization Proximal (ROP) loss enforcing sparse client updates, cutting communication up to 24% at near-unchanged accuracy; published as FedSparse (Electronics, 2024).
- Added a backpropagation-time node-pruning scheme to further sparsify updates, cutting compute and communication.

Gradient Calibration in Federated Learning 09/2022 – 03/2024

- Built the Federated Gradient Tailor (FGT) plug-in that detects and corrects conflicting client gradients before aggregation, improving accuracy by 5% with 6× faster convergence on non-IID data; published as Gradient Calibration for Non-IID FL (ACM workshop, 2023).
- Authored the formal convergence proof for FGT.

Federated Learning Efficiency on Dynamic Non-IID Data 09/2022 – 06/2024

- Designed an FL framework that adapts client participation to distribution drift via a hybrid client/server contribution indicator, improving accuracy ~1.3% over baselines on an already near-saturated benchmark; published as HybridCom at IEEE ICC 2024.

Integration of Knowledge Distillation and Curriculum Learning 08/2021 – 02/2022

- Built DCD, a dynamic curriculum-based knowledge-distillation framework that mitigates harm from heterogeneous sampling during distillation.
- Designed a curriculum generator that adapts training batches to teacher–student interactions and learning signals.
- Benchmarked DCD against vanilla KD and MEKD across multiple datasets, validating consistent gains; published as DCD (IEEE HPCC, 2022).

PROFESSIONAL EXPERIENCE

Brix AI, Irvine, CA 07/2024 – 05/2025

Machine Learning Engineer

- Built backend infrastructure for OpenAI API integration (assistant, chat, function-calling, and file search), reducing token usage by 20%.
- Improved job-description generation and resume parsing through prompt engineering and task decomposition, substantially raising content-generation accuracy.

Orbex Labs, Remote 03/2024 – 07/2024

Machine Learning Intern

- Ran a hyperparameter search on YOLOv8 for runtime/cost trade-offs, then introduced a word-frequency document-classification method that outperformed the YOLOv8 baseline at lower cost.
- Replaced the Oriented Bounding Box (OBB) model with the random sample consensus (RANSAC) algorithm to fix image rotation and scaling, raising text-extraction accuracy from 33% to 99%.

Baidu, Inc., Beijing, China 08/2020 – 03/2021

Machine Learning Intern

- Built an evaluation pipeline for Deepfake-detection models with dataset balancing and automated ROC/F1 reporting, contributing to a 2% accuracy gain.
- Built a rule-based model on adversarial text that blocked 600,000+ harmful comments slipping past the BERT filter.