

Binchi Zhang

 GitHub |  LinkedIn |  Homepage |  epb6gw@virginia.edu |  +1.434.257.8089

RESEARCH INTERESTS

Binchi Zhang is a final-year ECE Ph.D. candidate at the University of Virginia. His research focuses on parametric knowledge management for AI, where knowledge is modularized in external model parameters, allowing flexible and lifelong knowledge update, removal, and fusion. He studies both fundamental problems (certification and verification of knowledge removal, parameter space symmetry) and practical challenges (knowledge conflicts, agentic memory, personalization, and LLM safety and cultural alignment).

EDUCATION

University of Virginia

Ph.D. in Computer Engineering

Advisor: Prof. Jundong Li

Charlottesville, VA, USA

Sep 2022 – Present

Xi'an Jiaotong University

B.E. in Electrical Engineering

GPA: 90.88/100, Special Class for the Gifted Young

Xi'an, Shaanxi, China

Sep 2018 – Jun 2022

WORK EXPERIENCE

LinkedIn, Mentor: Eric Qasemi, Manager: Shuang Song, Bo Long

May 2026 – Present

- **LLM Reasoning and Generative Recommendation:** Investigate the faithfulness of chain-of-thought and implicit reasoning in LLM-based generative recommenders.
- **Counterfactual Evaluation:** Design a counterfactual cross-injection framework and new metrics to quantify reasoning faithfulness and consistency beyond ranking accuracy and surface plausibility.
- **Reasoning Model Training:** Build reproducible SFT and RL pipelines for LLM-based generative recommenders, benchmarking multiple state-of-the-art recommenders and systematically studying reasoning format, curriculum, quality, self-distillation, and reward.

Visa Research, Mentor: Hadi Abdullah, Manager: Yiwei Cai

May 2025 – Aug 2025

- **LLM Safety and Alignment:** Diagnose a structural failure mode in safety alignment, modeling refusal-based safeguards as low-entropy outlier behaviors vulnerable to continual fine-tuning.
- **SFT and RL Techniques:** Develop natural and semantically grounded targets to improve safety training pipelines across SFT and DPO backbones, balancing safety, helpfulness, and naturalness.
- **Evaluation and Experimental Design:** Introduce a perplexity-based naturalness metric to quantify safety retention and distributional shift under downstream adaptation.

NEC Labs, Mentor: Zhengzhang Chen, Manager: Haifeng Chen

Oct 2024 – Mar 2025

- **Continual Learning with Knowledge Memory:** Design a memory-based LLM knowledge codebook to flexibly learn new knowledge and unlearn unwanted knowledge without catastrophic forgetting.
- **Knowledge Conflict Resolution:** Analyze gradient conflicts between learning and unlearning objectives and introduced a memory-based resolution.
- **Cultural Alignment and Low-Resource Data Synthesis:** Align LLMs with both task structures and cultural norms. Design an automatic pipeline for culture-grounded data synthesis from web sources.
- **Benchmarking and Evaluation:** Construct and evaluate knowledge update benchmarks across LLM updating, unlearning, and editing. Cultural evaluation across 10 cultures and culture-sensitive tasks.

SELECTED PUBLICATIONS

Teaching LLMs to Remember: Parametric Memory via Personalized Adapter Weights.

Binchi Zhang, Jundong Li.

In submission.

When Safety Becomes An Outlier: Understanding the Retention of LLM Safety Behaviors.

Binchi Zhang, Hadi Abdullah, Yiwei Cai, Jundong Li.

In submission.

Exploiting Symmetry in Low-Rank Decomposition for LoRA Fusion.

Zaiyi Zheng*, **Binchi Zhang***, Liang Wu, Liangjie Hong, Jundong Li.

In submission.

LOKA: Conflict-Aware LLM Knowledge Update with Adaptive Knowledge Memory.

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, Haifeng Chen.

Annual Meeting of the Association for Computational Linguistics (ACL), 2026.

Mind the Gap in Cultural Alignment: Task-Aware Culture Management for Large Language Models.

Binchi Zhang, Xujiang Zhao, Jundong Li, Haifeng Chen, Zhengzhang Chen.

Annual Meeting of the Association for Computational Linguistics (ACL), 2026.

Beyond the Permutation Symmetry of Transformers: The Role of Rotation for Model Fusion.

Binchi Zhang*, Zaiyi Zheng*, Zhengzhang Chen, Jundong Li.

International Conference on Machine Learning (ICML), 2025 (Spotlight).

Verification of Machine Unlearning is Fragile.

Binchi Zhang, Zihan Chen, Cong Shen, Jundong Li.

International Conference on Machine Learning (ICML), 2024.

Towards Certified Unlearning for Deep Neural Networks.

Binchi Zhang, Yushun Dong, Tianhao Wang, Jundong Li.

International Conference on Machine Learning (ICML), 2024.

Adversarial Attacks on Fairness of Graph Neural Networks.

Binchi Zhang, Yushun Dong, Chen Chen, Yada Zhu, Minnan Luo, Jundong Li.

International Conference on Learning Representations (ICLR), 2024.

HONORS

OpenAI Cybersecurity Grant Program, \$2,500 in API credits	2025
--	------

Wilson Bicentennial Grad Fellowship, Department of ECE, UVA	2024
---	------

NSF Graduate Student Travel Award, SDM	2025, 2023
--	------------

Meritorious Winner, Mathematical Contest in Modeling	2021
--	------

Dean's List, Xi'an Jiaotong University	2020
--	------

SKILLS

Python, C++, PyTorch, NumPy, Huggingface, WandB, LangChain, TRL, Verl, Torch-Geometric