Weiyi He

619 Red Cedar Road, Michigan State University, East Lansing, MI 48824

☎ +1 (517) 512-0022 ✉ heweiyi@msu.edu www.linkedin.com/in/weiyi-he github.com/hwyii 🌐 hwyii.github.io

EDUCATION

Michigan State University **Aug. 2024 – Now**
Ph.D. in Statistics & Probability *East Lansing, MI*
Ph.D. in Computer Science & Engineering

University of Science and Technology of China (USTC) **Sep. 2020 – Jul. 2024**
Bachelor of Science in Statistics (Outstanding Graduate Award) *Hefei, China*

WORK EXPERIENCE

Articuler AI, Inc. **Jun. 2024 – Aug. 2024**
Generative AI Development Intern (Remote) *San Francisco, CA*

- Spearheaded the development of the **Search Agent** product — a scalable Python pipeline that automates professional background research for end users, from data collection to structured insight generation. (Deployed to production)
- Designed multi-step LLM workflows to parse user profiles, generate targeted search queries, extract and validate information from industry and news sources via Google Search API and custom web-scraping logic, improving retrieval accuracy and end-user satisfaction.

Elven Technologies Pte. Ltd. **Jan. 2024 – May. 2024**
Data Engineering Intern (Remote) *Singapore*

- Built Python API integrations with major cryptocurrency exchanges (e.g., Coinbase, Binance, OKX) to automate transaction extraction and synchronization for Elven's Web3 financial reporting platform.
- Parsed and formatted transaction logs, enabling reliable downstream reporting and visualization in the platform.

RESEARCH AND OTHER PROJECTS

ReFT-based latent adversarial training on low-dimensional manifolds **Nov. 2025 – Now**

- Implemented an algorithm combining **Representation Fine-tuning (ReFT)** with **Latent Adversarial Training (LAT)** to enhance the efficiency and robustness of Large Language Models.
- Designed and implemented an efficient adversarial attack based on Greedy Coordinate Gradient (GCG) in the latent space, where perturbations are constrained within a low-rank subspace to ensure attack precision and relevance.
- Developed a **Circuit-Aware** training framework that identifies critical attention heads via **Attribution Patching** and dynamically prunes non-essential gradient paths during the inner attack loop, reducing backward pass FLOPs by **30%**.

Impact of Positional Encoding on Transformer Generalization **May. 2025 – Oct. 2025**

- Presented the first **generalization and robustness theoretical analysis** of single-layer Transformers with trainable positional encoding (PE) under the in-context regression, bridging theoretical understanding with empirical behavior.
- Derived both clean and adversarial Rademacher complexity bounds, demonstrating that trainable PE systematically enlarges the generalization gap and increases the model's sensitivity and vulnerability under adversarial perturbations.

Privacy Risks in LLM Agent Memory **Oct. 2024 – Feb. 2025**

- Studied how LLM agents storing user-agent interations can expose private information through memory modules.
- Proposed **MEXTRA (Memory EXTRaction Attack)**, a black-box framework combining targeted attacking prompts with automated prompt generation under varying attacker knowledge levels.
- Analyzed architectural and deployment factors influencing memory vulnerability; demonstrated MEXTRA succeeded in extracting at least one private item in **83%–90%** of targeted trials across two real agents, highlighting the urgent need for effective memory safeguards in LLM agent designs.

PUBLICATIONS & MANUSCRIPTS

- **Weiyi He**, Yue Xing. "Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers Under In-Context Regression." *arXiv preprint arXiv:2512.09275*, 2025. (Under review.)
- Bo Wang, **Weiyi He**, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. "Unveiling Privacy Risks in LLM Agent Memory." *Proceedings of the Association for Computational Linguistics (ACL)*, 2025.

TECHNICAL SKILLS

Programming: Python, R, C, Sklearn, Numpy, Pandas, PyTorch, Tensorflow, Huggingface
Tools: Git, Linux, Jupyter, VS Code
Skills: Large Language Models, Trustworthy AI, Data Analysis, Statistical Learning Theory