ZI LI

zili@smail.nju.edu.cn ◇ zil2@andrew.cmu.edu

EDUCATION

Nanjing University *Aug. 2023 – Expected Jun. 2027*
B.S. in Computer Science (Information and Computing Science) Average: **91.2/100**

University of California, Berkeley *Summer 2025*
Visiting Student, Summer Session
Coursework: User Interface Design and Development (CS 160, **A+**); Computer Security (CS 161, **A+**)

Carnegie Mellon University *Mar. 2026 – Sep. 2026*
Visiting Research Student, Language Technologies Institute
Conducted research on robust AI-generated text detection under Prof. Lei Li's supervision.

PUBLICATIONS

1. **Zi Li**, Andre Vicente Duarte, and Lei Li. *PeerReveal: Learning to Generalize in AI Peer Review Detection*. Submitted to **ICLR 2027**.
2. **Zi Li**, Tian Zhou, Wenze Li, Jingyu Hua, Yunlong Mao, and Sheng Zhong. *Secret Stealing Attacks on Local LLM Fine-Tuning through Supply-Chain Model Code Backdoors*. Accepted to **EMNLP 2026 (Main Conference)**. **[arXiv]**
3. Wenjing Li, **Zi Li**, Yuan Zhang, and Sheng Zhong. *Baker: A Privacy-Preserving, NIZK-free and Efficient Payment Channel Hub Supporting Bidirectional Channels*. Accepted to **SRDS 2026**. **[ePrint]**

RESEARCH EXPERIENCE

Team Lead, Robust AI-Generated Text Detection *Mar. 2026 – Sep. 2026*
Language Technologies Institute, Carnegie Mellon University *Advisor: Prof. Lei Li*

- Designed and built PeerReveal, a three-class detector for human-written, AI-rewritten, and fully AI-generated reviews via **directed claim-level attention** over target and same-paper AI reference reviews; improved Macro-F1 over Sem-Detect by **21.55 points** under unseen-model shifts and **18.84 points** when both models and prompts are unseen.
- Developed a **quality-constrained search method** to automatically discover diverse, challenging review-generation and rewriting workflows; constructed **PeerShift**, a balanced **5,430-review benchmark** reflecting real-world complexity across frontier models and varied reviewing strategies in machine learning, life sciences, and social sciences.
- Built an **efficient adversarial training framework** that optimizes small local **9B generators** with **GRPO**, multi-detector feedback, and quality control, then retrains the detector; improved detector generalization to **unseen frontier models** and diverse review-generation workflows, raising pooled Macro-F1 on PeerShift from **70.52% to 82.40%**.

Team Lead, Supply-Chain Attacks on Local LLM Fine-Tuning *Sep. 2024 – May 2026*
Nanjing University *Advisor: Prof. Jingyu Hua*

- Led the project from ideation through submission, formulating a **code-level supply-chain threat model** in which compromised model code memorizes **sparse high-entropy secrets** during local fine-tuning for post-deployment black-box recovery.
- Designed and implemented the end-to-end **Active Execution Hijacking** framework, combining online tensor-rule matching and value-gradient decoupling for secret discovery and stealthy memorization with **CheckCode** to distinguish true leakage from hallucination under black-box access.
- Evaluated the attack on Llama-3.2 across summarization, medical question answering, and code generation, with **>98% Strict ASR** and no degradation in **primary-task utility** while evading semantic safety filtering, code auditing, and post-training backdoor detectors.

Second Author, Baker Payment Channel Hub *Jan. 2025 – Aug. 2025*
State Key Laboratory for Novel Software Technology, Nanjing University *Advisor: Prof. Yuan Zhang*

- Contributed to the design of Baker, the first bidirectional, privacy-preserving, and NIZK-free payment channel hub, integrating **Aggregatable Adaptor Signatures (AAS)** to enable atomic signature exchange and aggregation.

- Led **protocol optimization and experimental evaluation**, implementing the complete system and conducting large-scale benchmarks against state-of-the-art payment channel hubs.
- Demonstrated that Baker reduced off-chain communication and computation overhead to **0.0036% and 7%** of BlindHub's, respectively, while incurring **80%** of Accio's on-chain cost and achieving a **25% higher** average success rate than Accio.

SELECTED PROJECTS

Web3.0 Private Account Trading Platform GitHub

- Built a decentralized trading platform around a **novel dual-escrow model**, combining blockchain-based fund escrow with account-credential escrow to resolve the "who pays first" problem.
- Invented and implemented a **dynamic one-time 2FA mechanism** that allows buyers to verify account authenticity without being able to alter the credentials—**an approach not seen in existing platforms**.
- Added real-time trade tracking and flexible cancellation to improve transparency and protect users during disputes.
- Deployed on **Arbitrum Layer 2**, enabling secure, low-cost, end-to-end trades leveraging Ethereum scalability.

ICS Programming Assignment (NJU ICS-PA, RISC-V Track) Project Guide

- Built a teaching-oriented **full-system RISC-V emulator (NEMU)** in C as the execution foundation for a complete RISC-V software stack.
- Implemented instruction execution, device I/O models, and an interactive monitor with watchpoints, enabling realistic program execution and debugging.
- Developed the **Abstract Machine (AM)** runtime and a compact **operating system** with program loading, system calls, and a simple file system, successfully running real applications on the emulator.
- Added **preemptive time-sharing and virtual memory** for multiple user programs, completing the end-to-end stack from bare-metal execution to a multitasking OS.

Secure File Sharing System (UC Berkeley CS 161) Project Specification

- Designed and implemented a Go system with user authentication and **encrypted storage, secure sharing, and access revocation** over an untrusted remote datastore.
- Designed a **cryptographic data model** using symmetric encryption, HMACs, RSA encryption, and digital signatures to preserve confidentiality and integrity against datastore tampering and prevent revoked users from regaining access.
- Authored a detailed design document specifying data structures, core operations, and security guarantees.

HONORS AND AWARDS

National Scholarship, Ministry of Education of China *2025*
Awarded to top 0.2% of undergraduate students nationwide for outstanding academic and overall performance.
Zheng Gang Overseas Study Scholarship, Nanjing University *2025*
Awarded to support visiting research at Carnegie Mellon University.
Best Use of Artificial Intelligence, UC Berkeley CS 160 Course Award *2025*
Excellence Award, Undergraduate Basic Disciplines Scholarship, Nanjing University *2024*

TECHNICAL SKILLS

Programming & Web: Python, C/C++, TypeScript/JavaScript, Solidity, React
ML Frameworks: PyTorch, Hugging Face Transformers, SGLang, Slime, Megatron-LM
LLM Training: LLM fine-tuning (PEFT/LoRA), reinforcement learning (GRPO)
Security & Blockchain: applied cryptography, smart contract development, Ethereum
Tools: Linux, Git, LaTeX, Codex, Claude Code, OpenClaw