Haoyu Zhai

Thomas M. Siebel Center, 201 North Goodwin Avenue, Urbana, IL 61801-2302
zhai11@illinois.edu ⋄ (+1) 217-850-7555 ⋄ https://www.linkedin.com/in/haoyu-zhai-jeffrey

EDUCATION

- **University of Illinois at Urbana-Champaign** *Aug.2024 - May 2026*
  M.S. in Computer Science, GPA: 3.76/4.00
  Advisor: Gang Wang
- **University of Illinois at Urbana-Champaign** *Aug.2020 - May.2024*
  B.S. in Mathematics and Computer Science, GPA: **3.97**/4.00
  B.S. in Statistics

RESEARCH INTERESTS

AI for Security, Data-driven Security, AI Agent, Machine Learning

PUBLICATIONS

(*The authors contribute equally to this paper (co-first authors))

- **[IEEE SP 2026] H. Zhai***, S. Wang*, Q. Hao, P. Naghavi, G. Wang. **Revelio: Blurred Images Can Still Disclose Your Identity**. Proceedings of *The 47th IEEE Symposium on Security and Privacy*, San Francisco, CA, May 2026.
- [**SOUPS 2025**] Y. Wang, **H. Zhai**, C. Wang, Q. Hao, N. A. Cohen, R. Foulger, J. A. Handler, G. Wang. **Can You Walk Me Through It? Explainable SMS Phishing Detection using LLM-based Agents**. *Proceedings of the 21st Symposium on Usable Privacy and Security*, Seattle, WA, August 2025
- [**NeurIPS 2025**] J. Liu, N. Diwan, Z. Wang, M. Wahed, **H. Zhai**, X. Zhou, K. A. Nguyen, T. Yu, M. Wahed, Y. Deng, H. Benkraouda, Y. Wei, L. Zhang, I. Lourentzou, G. Wang. **PurpCode-R1: Reasoning for Safer Code Generation**. *Proceedings of the 39th Annual Conference on Neural Information Processing Systems*, San Diego, CA, Dec 2025

Pre-Prints

- **[CHI 2026, Under Review] H. Zhai***, Y. Wang*, N. A. Cohen, R. Foulger, J. A. Handler, G. Wang. **Human Decision Model in AI-assisted Phishing Detection**.

RESEARCH EXPERIENCE

- **Understanding User Perception of Deepfake Video Conference Calls**, UIUC *Aug.2025 - Present*
  - Develop real-time deepfake video call prototypes by integrating face-swapping model with voice-conversion tool
  - Design and conduct a human-subject study to evaluate participants' ability to detect and trust deepfake presenters during simulated Zoom meetings.
- **PurpCode-R1: Reasoning for Safer Code Generation**, UIUC *Jan.2025 - Aug.2025*
  - Contribute to the Amazon Nova AI Challenge (blue team) to develop reliable **LLM-based coding assistants**. Design high-coverage adversarial prompts as internal red team, simulating real-world unsafe coding scenarios.
  - Lead the end-to-end data curation pipeline: aggregated jailbreak prompts and templates from 10+ public safety datasets, apply LLM-based filtering to identify high-quality prompts, and generate aligned targets.
  - Benchmark model robustness against advanced search-based jailbreak methods (e.g., AutoDAN, GCG-Transfer), demonstrating superior performance over Qwen-series models across all evaluated attacks.
- **Face Image Deblur**, UIUC *Feb.2023 - May.2025*
  - Lead research on reconstructing intentionally blurred face images posted on social platforms, assessing the potential privacy leakage risks associated with existing blurring techniques.
  - Develop a multi-step approach integrating a conditional **diffusion model** for preliminary face restoration and an identity retrieval model to enhance fidelity using similar images.

  - Implement models in Python (**PyTorch**) and conducted large-scale experiments on public facial datasets, achieving **95.9%** recognition accuracy and outperforming state-of-the-art restoration methods.

- **LLM Phishing Agent**, UIUC *Aug.2024 - May.2025*
  - Design and implemented a robust **multi-agent LLM system** to detect SMS phishing, incorporating external knowledge (e.g., domain intelligence, webpage screenshots) to enhance reasoning and explainability.
  - Develop a **user-centric LLM agent** capable of delivering clear, actionable security advice, with tailored explanations optimized for elderly users through chain-of-thought prompting.
  - Achieve **98.8%** accuracy on real-world SMS phishing datasets; conduct a user study with 175 participants, earning a top-tier usability rating (SUS score: **82.6**).

PROJECT EXPERIENCE

- **Splitter Web Application** *Jan.2023 - May.2023*
  - Lead the full-stack development of a web-based bill splitting platform using **React** (frontend) and **Flask** (backend)
  - Integrate Google Authentication API for user login and create back-end APIs for group management, bill creation, and automatic split calculation.
  - Design, manage **SQLite3** database for Apps, and connect with backend API for long-term user data storage.

WORK EXPERIENCE

- **Computer Network Information Center, Chinese Academy of Science** *May.2023 - Aug.2023*
  Big Data Developer Intern
  - Develop Lynx, a customizable Cypher query execution framework suitable for any type of database, allowing developers to query any type of database with graph database query syntax, avoiding costly database migrations.
  - Apply Lynx to **MySQL** and **MongoDB** using **Scala** and implement necessary APIs to connect mySQL database to Lynx, enabling graph query searching on relational databases and NoSQL databases.
  - Test the query performance using LDBC benchmark and optimize the framework by manipulating database indexes and improving code logic, reducing single record query time **from seconds to tens of milliseconds**.

HONORS AND AWARDS

- **Amazon Nova AI Challenge**, 1st Place Winner ($250K Prize) *Jul.2025*
- **Illinois Statistics Datathon**, Best Data Visualization & Top 10 Model Accuracy *Dec.2023*
- **Highest Distinction**, B.S. in Mathematics & Computer Science, UIUC *May.2024*
- **Highest Distinction**, B.S. in Statistics, UIUC *May.2024*
- **Dean's List**, College of Liberal Arts & Sciences, UIUC *2020-2023*

SKILLS

- **Programming Languages:** Python, R, C++, Java, SQL, Scala, JavaScript, HTML, Assembly, Shell
- **Technologies/Frameworks:** LLM Frameworks (vLLM, VerL), Python Libraries (PyTorch, Pandas, Matplotlib, OpenCV), AWS, NoSQL, Flask, Unreal Engine, Git, Docker, Linux