

# ZEHUA ZHANG

Tempe, AZ, 85281 | 617-893-9077 | zzhan645@asu.edu | [Google Scholar](#)

---

## EDUCATION

### ARIZONA STATE UNIVERSITY

PhD Student in Computer Science | GPA: 3.67 / 4.0

Co-advised by Dr. Chitta Baral and Dr. Ruoyu Wang

Expected completion: Spring 2028

Tempe, AZ

Aug 2024 – Present

### CARNEGIE MELLON UNIVERSITY

M.S. in AI Engineering – Information Security | GPA: 3.7 / 4.0

Advised by Dr. Amir Barati Farimani and Dr. David Varodayan

Pittsburgh, PA

May 2024

### BOSTON COLLEGE

B.A. in Computer Science and Economics | GPA: 3.65 / 4.0

Dean's List (First Honor): 2020-2022; Highlighted Schiller Student Challenge Projects

Chestnut Hill, MA

May 2022

## RESEARCH EXPERIENCE

### ARIZONA STATE UNIVERSITY

Graduate Research Assistant

Tempe, AZ

Aug 2024 – Present

Conduct research at the intersection of LLM agents, software security, and AI security, with a focus on building reliable and secure agentic systems.

Develop benchmarking, reasoning, and post-training methods for complex agentic tasks, including open-source software compilation and long-horizon tool use.

Investigate security vulnerabilities in LLM-based systems, including jailbreaks and prompt injection, and develop methods for prompt injection vulnerability analysis using data-flow centric approach and defense by conducting LLM post-training.

### CARNIGIE MELLON UNIVERSITY

Graduate Research Assistant

Pittsburgh, PA

May 2023 – May 2024

Utilized Deep Learning methods to surrogate molecular dynamics; proposed a new masking pretraining method for Force and Energy-Centric Graph Neural Networks.

Pretrained and finetuned GNNs to predict atomic force and potential energy of water molecule datasets (Tip3p and RPBE) and small organic molecule dataset (MD17), reducing respective RMSE for water datasets by 38.09% and 25.4%.

PEKING UNIVERSITY  
Research Assistant

Beijing, China  
May 2021 – June 2021

Evaluated and tested multiple protocols of Blockchains and Smart Contract codes, reporting on best practices of Smart Contract Development.

Constructed Test net based on Morden and Private blockchain to reproduce 4 attacks targeting Smart Contracts and qualitatively analyzed the significance of Governance Token on restricting Flash Loan Attack.

Quantitatively assessed suppressing effecting of regulations on Cryptocurrency transactions and major mining pools based on on-chain data.

## RESEARCH INTERESTS

General Interests: Natural Language Processing and its applications on Security/SE  
Recent Focuses: LLM Post-training, Multi-Agent System, AI for Software Security

## PROFESSIONAL EXPERIENCE

SAMSUNG RESEARCH AMERICA  
Research Intern

Mountain View, CA  
May 2026 – Aug 2026

Conducted research on indirect prompt injection (IPI) vulnerabilities in LLM-based agentic systems, with a focus on improving robustness against adversarial instructions embedded in untrusted data.

Developed a gym-like training environment for scalable IPI training data sampling and generation, supporting both online and offline post-training of assistant models and standalone safety verification models.

Trained LLMs of different sizes and families using SFT and RL (DPO and GRPO) and achieved state-of-the-art performance on established benchmarks, including AgentDojo and AgentDyn.

SOHU.INC  
Algorithm Development Intern

Beijing, China  
June 2021 – Sept 2021

Designed and deployed a real-time feed deduplication system to enhance user experience and content diversity for social media app.

Leveraged large-scale data streaming pipelines (Kafka, Spark) and advanced similarity detection methods (e.g., simhash, Bloom filters) to filter redundant multi-media data.

Achieved substantial efficiency gains, reducing repetitive text exposure by ~90% and duplicate video content by ~30%.

## TECHNICAL SKILLS

**Programming:** Python, C, Scala, Java, R, Swift, JavaScript, HTML, CSS

**ML/DL Frameworks:** PyTorch, TensorFlow, PyG, Verl, vLLM

**Data & Distributed Systems:** Spark, Kafka, Kubernetes

**Cloud & Databases:** MySQL, MongoDB, Hive, HBase; GCP, AWS

**Software Development:** React, IOS, Docker, MCP

## PUBLICATIONS

A. Saeidi, **Z. Zhang**, R. K. Singh, et al., “CAST: Critique-Aware Supervision for Training Reliable Long-Horizon Tool-Calling Agents.”  
<https://arxiv.org/abs/2608.30147v1>. Accepted to EMNLP 2026 Main Conference. (*Co-First Author*)

**Z. Zhang**, A. Bajaj, D. Handa, et al., “Build-Bench: Benchmarking LLM Agents on Compiling Real-World Open-Source Software.” <https://openreview.net/forum?id=BxJnz9EqO5>. Accepted in TMLR. (*First Author*)

D. Handa, **Z. Zhang**, A. Saeidi, S. Kumbhar, and C. Baral, “When ‘Competency’ in reasoning opens the door to vulnerability: jailbreaking LLMs via novel complex ciphers,” arXiv.org, Feb. 16, 2024. <https://arxiv.org/abs/2402.10601>. Accepted for presentation at NeurIPS 2025 Reliable ML from Unreliable Data workshop.

**Z. Zhang**, Z. Li, and A. B. Farimani, “Masked pretraining strategy for neural potentials,” AIP Advances, vol. 14, no. 9, Sep. 2024, doi: 10.1063/5.0202647. (*First Author*)

## PREPRINTS

One paper submitted to NDSS 2027. (*First Author*)

J. Li, Y. Yuan, and **Z. Zhang**, “Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases,” arXiv.org, Mar. 15, 2024. <https://arxiv.org/abs/2403.10446> (*Co-First Author*)

## **TEACHING EXPERIENCE**

ARIZONA STATE UNIVERSITY  
Graduate Teaching Assistant

Tempe, AZ  
Aug 2026 – Dec 2026

Courses: Information Assurance (CSE365)

Responsibilities: grading, office hours, student support

BOSTON COLLEGE  
Undergraduate Teaching Assistant

Chestnut Hill, MA  
Sept 2021 – May 2022

Courses: Randomness and Computation (CSCI2244), Web Application Development (CSCI2254)

Responsibilities: grading, office hours, student support