Jingxuan He

jxhe.info
jingxuan.he@berkeley.edu
+1 (341) 766-1883

Research Summary

I study how AI reshapes the cybersecurity landscape, a core pillar of broader AI safety. My goal is to transform AI from a potential source of vulnerabilities into a security enabler for future software systems. To this end, I build systematic benchmarks to measure AI's cybersecurity capabilities and risks, and I develop secure-by-design techniques that prevent AI from generating vulnerable code. My research draws on and contributes to AI, security, programming languages, and software engineering.

Academic Background

Postdoctoral Scholar, **University of California, Berkeley**, USA — 2024.10-Present
Advisor: Prof. Dawn Song

PhD in Computer Science, **ETH Zurich**, Switzerland — 2018.09-2024.09
Thesis: *Machine Learning for Code: Security and Reliability*
Awarded the **ETH Medal for Outstanding Doctoral Thesis**
Advisor: Prof. Martin Vechev

MS in Computer Science, ETH Zurich, Switzerland — 2016-2018

BE in Computer Science and Technology, Zhejiang University, China — 2012-2016

Representative Papers

Preprint 2025a — CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale.
Zhun Wang*, Tianneng Shi*, Jingxuan He, Matthew Cai, Jialin Zhang, Dawn Song.
Adopted in Anthropic's Claude Sonnet 4.5 System Card.
Top 0.6% of submissions at ICLR 2026 (under review).

ICML 2025 — BaxBench: Can LLMs Generate Secure and Correct Backends?
Mark Vero, Niels Mündler, Victor Chibotaru, Veselin Raychev, Maximilian Baader, Nikola Jovanović, Jingxuan He, Martin Vechev.
Spotlight Paper.

CCS 2023 — Large Language Models for Code: Security Hardening and Adversarial Testing.
Jingxuan He, Martin Vechev.
Distinguished Paper Award.

PLDI 2025 — Type-Constrained Code Generation with Language Models.
Niels Mündler*, Jingxuan He*, Hao Wang, Koushik Sen, Dawn Song, Martin Vechev.
Featured as #1 on Hacker News.

Honors and Awards

Two ICML Spotlight Papers — 2025
ETH Medal for Outstanding Doctoral Thesis — 2024
ACM CCS Distinguished Paper — 2023
NeurIPS Top Reviewer — 2023

Grants

OpenAI Cybersecurity Grant, **Google-BAIR** Commons ~$225,000
CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale
PI: Prof. Dawn Song
My role: Led end-to-end development, including core idea formulation, proposal writing, advising student researchers, and driving the project to publication and external impact.

DARPA (TRACTOR: Translating All C to Rust) $5,000,000
Improved Decoding and Equivalence Automated Testing at Scale (IDEAS)
Co-PIs: Intel, Prof. Dawn Song, Prof. Koushik Sen
My role: Contributed core research ideas, co-wrote the proposal, supported technical development, and participated in PI-level coordination meetings.

Open Philanthropy (Navigating Transformative AI) $3,390,000
Benchmark AI Cyberoffense Capabilities across the Cyber Kill Chain
PI: Prof. Dawn Song
My role: Contributed key ideas and technical components to the proposal writing process.

Full Paper List

Preprint 2025a CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale.
Zhun Wang*, Tianneng Shi*, Jingxuan He, Matthew Cai, Jialin Zhang, Dawn Song.
Adopted in Anthropic's Claude Sonnet 4.5 System Card.
Top 0.6% of submissions at ICLR 2026 (under review).

Preprint 2025b VERINA: Benchmarking Verifiable Code Generation.
Zhe Ye, Zhengxu Yan, Jingxuan He, Timothe Kasriel, Kaiyu Yang, Dawn Song.

PLDI 2025 Type-Constrained Code Generation with Language Models.
Niels Mündler*, Jingxuan He*, Hao Wang, Koushik Sen, Dawn Song, Martin Vechev.
Featured as #1 on Hacker News.

ICML 2025a BaxBench: Can LLMs Generate Secure and Correct Backends?
Mark Vero, Niels Mündler, Victor Chibotaru, Veselin Raychev, Maximilian Baader,
Nikola Jovanović, Jingxuan He, Martin Vechev.
Spotlight Paper.

ICML 2025b Formal Mathematical Reasoning: A New Frontier in AI.
Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter,
Swarat Chaudhuri, Dawn Song.
Spotlight Paper.

ICML 2025c Mind the Gap: A Practical Attack on GGUF Quantization.
Kazuki Egashira, Robin Staab, Mark Vero, Jingxuan He, Martin Vechev.
Oral Presentation at ICLR 2025 Workshop on Trustworthy LLM.

ICML 2025d Black-Box Adversarial Attacks on LLM-Based Code Completion.
Slobodan Jenko*, Niels Mündler*, Jingxuan He, Mark Vero, Martin Vechev.

NeurIPS 2024a Exploiting LLM Quantization.
Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, Martin Vechev.
Oral Presentation at ICML 2024 Workshop on Next Generation of AI Safety.

NeurIPS 2024b — SWT-Bench: Testing and Validating Real-World Bug-Fixes with Code Agents.
Niels Mündler, Mark Niklas Müller, Jingxuan He, Martin Vechev.

ICML 2024 — Instruction Tuning for Secure Code Generation.
Jingxuan He*, Mark Vero*, Gabriela Krasnopolska, Martin Vechev.

ICLR 2024 — Self-contradictory Hallucinations of LLMs: Evaluation, Detection and Mitigation.
Niels Mündler, Jingxuan He, Slobodan Jenko, Martin Vechev.

CCS 2023 — Large Language Models for Code: Security Hardening and Adversarial Testing.
Jingxuan He, Martin Vechev.
Distinguished Paper Award.

ICML 2022 — On Distribution Shift in Learning-based Bug Detectors.
Jingxuan He, Luca Beurer-Kellner, Martin Vechev.

ICML 2021 — TFix: Learning to Fix Coding Errors with a Text-to-Text Transformer.
Berkay Berabi, Jingxuan He, Veselin Raychev, Martin Vechev.

PLDI 2021 — Learning to Find Naming Issues with Big Code and Small Supervision.
Jingxuan He, Cheng-Chun Lee, Veselin Raychev, Martin Vechev.

CCS 2021 — Learning to Explore Paths for Symbolic Execution.
Jingxuan He, Gishor Sivanrupan, Petar Tsankov, Martin Vechev.

PLDI 2020 — Learning Fast and Precise Numerical Analysis.
Jingxuan He, Gagandeep Singh, Markus Püschel, Martin Vechev

CCS 2019 — Learning to Fuzz from Symbolic Execution with Application to Smart Contracts.
Jingxuan He, Mislav Balunović, Nodar Ambroladze, Petar Tsankov, Martin Vechev.

CCS 2018 — DeBin: Predicting Debug Information in Stripped Binaries.
Jingxuan He, Pesho Ivanov, Petar Tsankov, Veselin Raychev, Martin Vechev.

Teaching

(Advanced) Large Language Model Agents, UC Berkeley 2024-2025
Advised student research projects, hosted guest lecturers.

Program Analysis for System Security and Reliability, ETH Zurich 2020-2022
Gave guest lectures, organized course projects, taught exercises, and designed exam questions.

Reliable and Interpretable Artificial Intelligence, ETH Zurich 2019-2022
Organized course projects.

Rigorous Software Engineering, ETH Zurich 2019-2023
Gave guest lectures, taught exercises, and designed exam questions.

Seminars at ETH Zurich: **ML for Code**, **Software Engineering**, and **Blockchain Security** 2018-2023
Co-organized the entire course, co-examined students, and advised student presentations.

Mentoring

UC Berkeley 2024.10-Present

- **Zhun Wang** and **Tianneng Shi**: Ongoing PhD research on AI for cybersecurity, e.g., [Preprint 2025a].
- **Hao Wang**: Ongoing PhD research on secure code generation, e.g., [PLDI 2025].

- **Zhe Ye**: Ongoing PhD research on AI and formal verification, e.g., [Preprint 2025b].
- Additionally mentored 4 BS students to help with the above projects.

ETH Zurich 2018.9-2024.9 (with ongoing collaborations)

- **Niels Mündler**: Ongoing PhD research on secure code generation, e.g., [PLDI 2025, ICML 2025d]. Predoctoral project [NeurIPS 2024b] and MS thesis [ICLR 2024].
- **Mark Vero**: Ongoing PhD research on secure code generation, e.g., [ICML 2025a, ICML 2024].
- **Kazuki Egashira**: Co-mentored MS projects on LLM quantization security [ICML 2025c, NeurIPS 2024a].
- **Luca Beurer-Kellner**: PhD project [ICML 2022].
- **Slobodan Jenko**: MS semester project and thesis [ICML 2025d].
- **Gabriela Krasnopolska**: MS semester project and thesis [ICML 2024].
- **Berkay Berabi**: Co-mentored MS thesis [ICML 2021].
- **Gishor Sivanrupan**: Co-mentored MS semester project [CCS 2021].
- Additionally mentored 7 other MS students and 1 BS student on research projects and theses.

Service

Program Committees

- IEEE S&P, USENIX Security 2026
- ACM CCS, LMPL Workshop at SPLASH, AgentSE Workshop at ASE 2025
- PLDI Artifact Evaluation 2022

Reviewers

- ICLR 2026
- NeurIPS, ICLR, ICML, ACM TOSEM, IEEE TSE 2025
- NeurIPS, ACM TOSEM 2024
- NeurIPS (Top Reviewer), AISTATS, IEEE TSE 2023
- ICML, ACM TOSEM 2022

Selected Talks

Security: A Next Frontier in AI Coding 2025
Stanford Security Seminar, Berkeley Security Seminar, UIUC iSE Reading Group, Berkeley Undergraduate Cybersecurity Club, OpenAI Security Research Conference, Machine Learning at Berkeley

Securing AI Code Generation 2024
LLMs and Cognitive Systems Workshop at UC Berkeley

Large Language Models for Code: Security Hardening and Adversarial Testing 2023
Deep Learning-aided Verification Workshop at CAV 2023, National University of Singapore, Peking University, Zhejiang University, LLMs for Code Seminar, Privacy and Security in ML Seminar, Dagstuhl Seminar on Programming Language Processing

Learning to Explore Paths for Symbolic Execution 2022
KLEE Workshop 2022

Machine Learning for Program Analysis 2020-2022
Huawei Research Munich, University of Paris-Saclay, Peking University, Facebook, Democratizing Software Verification Workshop at CAV 2020