Hongwei Li

Ph.D. Student, Department of CS, UCSB
765-543-8337, lihongweiandre@gmail.com

1 Research Interests

My research centers around **AI agents for Security and SE and agentic Fine-tuning**. My current research focuses on developing customized *AI agents systems* and *reasoning LLMs* for software security and engineering tasks such as patch generation, vulnerability detection, and PoC generation. I also enjoy playing CTF as a core member of Shellphish, a renowned CTF team.

2 Highlighted Projects & Skills

Advanced LLM agents for software engineering and security

- Design and develop pioneer LLM agents with *domain-specific tools, advanced workflows, and memory managements* for SE and security
- Outcomes:
  - **PatchPilot** — A cost-efficient issue resolving agent, ranked Top-5 tools on SWE-bench-Verified;
  - **Artiphishell** — A multi-agent system for vulnerability detection and patching, ranked Top-5 in DARPA AIxCC competition;
  - **AigiSE** (Under Development) - A comprehensive cybersecurity agent framework with advanced dynamic agent creation and memory management, domain-specific tools, and sandboxed environments on Docker and Kubernetes suitable for RL-based training, achieved a 58% success rate on the CyberGym benchmark subset;
- Skills: Google Agent Development Kit, Agent design, MCP protocols, Fuzzing (Generic/Directed /Kernel/Web3), Static Analysis(CodeQL/Joern), Concolic Execution(Angr), Reverse Engineering, Vulnerability Exploitation

LLM post-training and agentic training

- Design *novel training recipes for distillation with agentic data*; data selection strategies, running multiple teacher models in our domain-specific agents, unique data filtering and constitution-based data rationalization and correction, and SFT-based fine-tuning; the obtained small models outperform large teacher models in our target tasks
- Outcomes:
  - **VulnLLM-R** - The first reasoning-based large language model for vulnerability detection, equipped with agent scaffolding and agentic fine-tuning, outperforms SOTA tools;
  - **Co-PatcheR** — A set of small reasoning models for issue resolving, ranked Top-2 open-weight models on SWE-bench-Verified;
- Skills: LLaMA-Factory, SFT, vLLM, RL, PyTorch, HuggingFace, verl, Data/model/pipeline parallelism (e.g., FSDP, DDP)

3 Education

University of California, Santa Barbara, Goleta, CA Sep. 2024 – Present
Ph.D. Student, Department of Computer Science, advised by Prof. Wenbo Guo

Purdue University, West Lafayette, IN Sep. 2023 - May. 2024
Ph.D. Student, Department of Computer Science, advised by Prof. Wenbo Guo

Shanghai Jiao Tong University, Shanghai, China Aug. 2020 - Mar. 2023
M.Eng. in Cybersecurity, advised by Prof. Yue Wu

Shanghai Jiao Tong University, Shanghai, China Sep. 2016 - Jun. 2020
B.Eng. in Information Engineering; B.A. in French (Dual Degree)

4 Publications

LLM post-training and agents

1. Yuzhou Nie, **Hongwei Li**, Ruizhe Jiang, Chengquan Guo, Zhun Wang, Bo Li, Dawn Song, Wenbo Guo, "VulnLLM-R: Specialized Reasoning LLM with Agent Scaffold for Vulnerability Detection", Submitted to *The 47th IEEE Symposium on Security and Privacy*, Rio de Janeiro, Brazil, 2026

2. Shellphish, "An AI-powered Cyber Reasoning System for Automatic Vulnerability Identification and Patching", Submitted to *Phrack Papers*, 2026

3. Yuheng Tang, Kaijie Zhu, Bonan Ruan, Chuqi Zhang, Michael Yang, **Hongwei Li**, Suyue Guo, Tianneng Shi, Zekun Li, Christopher Kruegel, Giovanni Vigna, Dawn Song, William Yang Wang, Lun Wang, Yangruibo Ding, Zhenkai Liang, Wenbo Guo, "DevOps-Gym: Benchmarking AI Agents in Software DevOps Cycle", Submitted to *The 14th International Conference on Learning Representations*, 2026

4. Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, **Hongwei Li**, Wenbo Guo, Dawn Song, "Progent: Programmable Privilege Control for LLM Agents", Submitted to *The 47th IEEE Symposium on Security and Privacy*, Rio de Janeiro, Brazil, 2026

5. Yujin Potter, Wenbo Guo, Zhun Wang, Tianneng Shi, **Hongwei Li**, Andy Zhang, Patrick Gage Kelley, Kurt Thomas, Dawn Song, "SoK: Frontier AI's Impact on the Cybersecurity Landscape", Submitted to *The 47th IEEE Symposium on Security and Privacy*, Rio de Janeiro, Brazil, 2026

6. **[ICML'25] Hongwei Li**, Yuheng Tang, Shiqi Wang, Wenbo Guo, "PatchPilot: A Verifiable and Cost-Efficient Agentic Patching Framework", In Proceedings of *The 42nd International Conference on Machine Learning*, Vancouver, CA, 2025

7. **[NeurIPS'25]** Yuheng Tang, **Hongwei Li**, Kaijie Zhu, Michael Yang, Yangruibo Ding, Wenbo Guo, "Co-PatcheR: Collaborative Software Patching with Component-specific Small Reasoning Models", In Proceedings of *The 39th Conference on Neural Information Processing Systems*, San Diego, CA, 2025

RL and DNN for Security

1. Wenxuan Shi, **Hongwei Li**, Jiahao Yu, Xinqian Sun, Wenbo Guo, Xinyu Xing, "BandFuzz: An ML-powered Collaborative Fuzzing Framework.", arXiv preprint arXiv:2507.10845.

2. **[ICSE SBFT'24]** Wenxuan Shi, **Hongwei Li**, Jiahao Yu, **Wenbo Guo** Xinyu Xing, "BandFuzz: A Practical Framework for Collaborative Fuzzing with Reinforcement Learning", In Proceedings of *In International Workshop on Search-Based and Fuzz Testing*, Lisbon, Portugal, 2024

3. **[Computers & Security'22]** Jingcheng Yang, **Hongwei Li**, Shuo Shao, Futai Zou, Yue Wu, "FS-IDS: A framework for intrusion detection based on few-shot learning", *Computers & Security*, 122: 102899, 2022.

5 Selected Competitions

- AIxCC, *DARPA*, 2023-2025 (**5th Place; Awarded $3M**)
- SWE-Bench patching leaderboard, 2025 (**Top 5 tools; Top 2 Open-weight models**)
- Google SBFT Fuzzing Tool Competition, 2024 (**Top 1, Awarded $23K**)
- DEF CON CTF Final, 2025 (**11th Place**)
- SunshineCTF, Sep 2021 (**11th Place**) (*Solo Team in a Team-based CTF*)