Peiran Wang

mobile: (+1) 323-986-8265 · email: whilebug@gmail.com · https://whilebug.github.io/

Education Experience

University of California, Los Angeles , Computer Science, *PhD* 2025.09-2031.06(Expected)

Tsinghua University, Cybersecurity, *Master* 2022.09-2025.06

Sichuan University, Cybersecurity, Talented Class, National Scholarship, Bachelor 2018.09-2022.06

Awards & Scholarship

[1] 2019 **National Scholarship** by Ministry of Education

[2] 2021 IEEE Symposium on Computers and Communications (ISCC) **Best Paper Award**

[3] 2022 Microsoft Research Asia (MSRA) Star of Tomorrow Award

[4] 2024 ACM Conference on Computer and Communications Security (CCS) **Distinguished Paper Award**

Career Experience

ByteDance, Research Intern 2025.03-2025.08

Work in *Security Research Group* on system security for LLM agent

ByteDance, Research Intern 2024.07-2025.01

Work in *Security Research Group* on LLM hallucination detection

Baidu, Research Intern 2024.05-2024.07

Work in *ACG (Intelligent Cloud Business Group) Summer Camp*, LLM training acceleration.

Future Capital, Investment Intern 2023.06-2023.08

Work in *computer industry investment group*.

Microsoft Research Asia, Research Intern 2021.09-2023.03

Work in the *system research group* on LLM training acceleration.

Research Experience

University of California, Los Angeles (UCLA), Research Assistant 2025.09-Present

Pursue PhD position under the supervison of *Prof. Yuan Tian*

University of Illinois at Urbana-Champaign (UIUC), Research Intern 2024.01-2025.03

Work with *Prof. Haohan Wang* on Trustworthy AI.

University of California, San Diego (UCSD), Research Intern 2023.05-2024.02

Work with *Prof. Haojian Jin* on TTI model moderation.

University of Wisconsin-Madison (UWM), Research Intern 2023.11-2024.10

Work with *Prof. Chaowei Xiao* on LLM security.

Tsinghua University (THU), Research Assistant 2022.09-2023.05

Work with *Prof. Jessie Hui Wang* on networking.

Sichuan University (SCU), Research Assistant 2020.04-2022.06

Work with *Prof. Beibei Li* on federated learning and *Prof. Cheng Huang* on mining hijacking.

Papers

[1] Liu, H., **Wang, P.**, Xing, T., Li, Y., Dalal, V., Li, L., He, J., Wang, H. Dataset Distillation via the Wasserstein Metric. 2025 International Conference on Computer Vision (ICCV).

[2] Shen, X., Yi, L., **Wang, P.**, et al. Split-and-Privatize Framework for Large Language Model Fine-Tuning, 34th International Joint Conference on Artificial Intelligence (IJCAI).

[3] Shao, Z., Li, B., **Wang, P.**, Zhang, Y., Choo, K. K. FedLoRE: Communication-Efficient and Personalized Edge Intelligence Framework via Federated Low-Rank Estimation. Transactions on Parallel and Distributed Systems.

[4] **Wang, P.**, Liu, X., Xiao, C. RePD: Defending Jailbreak Attacks Through a Retrieval-based Prompt Decomposition Process. NAACL 2025.

[5] **Wang, P.**, Liu, X., Wu, F., Cao, Y., Zhang, Y., Sun, L., Xiao, C. CVE-Bench: Benchmarking LLM-based Software Engineering Agent's Ability to Repair Real-World CVE Vulnerabilities. NAACL 2025.

[6] Xue, E., Li, Y., Liu, H., **Wang, P.**, Shen, Y., Wang, H. Towards Adversarially Robust Condensed Dataset by Curvature Regularization. The 39th Annual AAAI Conference on Artificial Intelligence.

[7] Shao, Z., Li, B., **Wang, P.**, Li, W., Zhang, Y. FedUFD: Uncertainty-Driven Feature Distillation for Heterogeneous Federated Learning. In IEEE International Conference on Computer Communications (INFOCOM), 2025.

[8] **Wang, P.***, Li, Q.*, Yu, L., Wang, Z., Li, A., Jin, H. Moderator: Moderating Text-to-Image Diffusion Models through Fine-grained Context-based Policies. In 31st ACM Conference on Computer and Communications Security (CCS), 2024. ***Distinguished Paper Award***.

[9] Chen, X., Meng, W., **Wang, P.**, Zhou, Q. Distributed Boosting: An Enhancing Method on Dataset Distillation. In 33rd ACM International Conference on Information and Knowledge Management (CIKM), 2024.

[10] Jiang, N., Wang, J. H., Wang, J., **Wang, P.** Top AS Router Geolocation in Databases: Performance and Techniques. In GLOBECOM 2023 - IEEE Global Communications Conference, pp. 2117-2122. IEEE, 2023.

[11] Jiang, N., Wang, J. H., Wang, J., **Wang, P.** TinyG: Accurate IP Geolocation Using a Tiny Number of Probers. In 2023 19th International Conference on Network and Service Management (CNSM).

[12] Li, B., **Wang, P.***, Shao, Z., Liu, A., Jiang, Y., Li, Y. Defending Byzantine Attacks in Ensemble Federated Learning: A Reputation-based Phishing Approach. Future Generation Computer Systems, 147 (2023): 136-148. **1st student author**.

[13] Li, B., Shao, Z., Liu, A., **Wang, P.** FedCliP: Federated Learning with Client Pruning. arXiv preprint.

[14] Li, B., **Wang, P.***, Huang, H., Ma, S., Jiang, Y. FlPhish: Reputation-based Phishing Byzantine Defense in Ensemble Federated Learning. In 2021 IEEE Symposium on Computers and Communications (ISCC), pp. 1-6. IEEE, 2021. **1st student author**. ***Best Paper Award***.

[15] Li, B., Jiang, Y., Sun, W., Niu, W., **Wang, P.** FedVANet: Efficient Federated Learning with Non-IID Data for Vehicular Ad Hoc Networks. In 2021 IEEE Global Communications Conference (GLOBECOM), pp. 1-6. IEEE, 2021.

[16] **Wang, P.**, Sun, Y., Huang, C., Du, Y., Liang, G., Long, G. Minedetector: JavaScript Browser-side Cryptomining Detection Using Static Methods. In 2021 IEEE 24th International Conference on Computational Science and Engineering (CSE).

[17] **Wang P.**, Liu Y., Lu Y., Cai Y., Chen H., Yang Q., Zhang J., Hong J., Wu Y. AgentArmor: Enforcing Program Analysis on Agent Runtime Trace to Defend Against Prompt Injection. In submission.

[18] **Wang P.**, Yu Y., Zhan X., Chen K., Wang H. Automatic Data Science Agent: A Survey. In submission.

[19] Jin H., Zhang P., **Wang P.**, Wang H. From Hallucinations to Jailbreaks: Rethinking the Vulnerability of Large Foundation Models. In submission.

[20] Jiang, Y.*, **Wang, P.***, Lin, C., Huang, Y., Cheng, Y. SecDT: Mitigating Label Leakage in Two-Party Split Learning. arXiv, in submission.

[21] **Wang, P.**, Wang, H. Who's Behind the Curtain: Discover and Understand the LLM-based Chat Agent on the Social Network. arXiv, in submission.

[22] Wang, J.*, Li, P.*, Ma, S.*, **Wang, P.***, Liu, X., Sun, J., Liang, Y., Xia, T., Wang, Y., Luo, W., Xiao, C. Prompt Injection Benchmark for Foundation Model Integrated Systems.

[23] **Wang, P.**, Wang, H. Understanding Political Stereotypes in Large Language Models through Ideological Testing. arXiv, in submission.

[24] **Wang, P.**, Liu, Y., Lu, Y., Tan, Z. What Are Models Thinking About? Understanding Large Language Model Hallucinations "Psychology" Through Model Inner State Analysis. Arxiv, in submission.

[25] **Wang, P.**, Li, H., Tian, R., Li, S., Wang, Y., Shen, D. Astra: Efficient and Money-saving Automatic Parallel Strategies Search on Heterogeneous GPUs. Arxiv, in submission.

[26] **Wang, P.**, Wang, H. DistDD: Distributed Data Distillation Aggregation Through Gradient Matching. Arxiv, in submission.

[27] Yue Huang, Chujie Gao, Siyuan Wu, ..., **Peiran Wang**, ... TrustGen: A Platform of Dynamic Benchmarking on the Trustworthiness of Generative Foundation Models. Arxiv, in submission.