Peiran Wang

mobile: (+86) 17883693551 · email: whilebug@gmail.com

Education Experience

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] 2019 Sichuan University First-Level Scholarship

[3] 2021 IEEE Symposium on Computers and Communications (ISCC) **Best Paper Award**

[4] 2022 Microsoft Research Asia (MSRA) Star of Tomorrow Award

[5] 2022 Sichuan University Bachelor Excellent Thesis

[6] 2024 ACM Conference on Computer and Communications Security (CCS) **Distinguished Paper Award**

Career Experience

ByteDance, Research Intern 2024.07-Present

Work in *Security Research Group* on LLM security.

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 Illinois at Urbana-Champaign (UIUC), Research Intern 2024.01-2024.11

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.

Tencent, Short-term collaboration 2023.04-2023.05

Work with the *PowerFL team* on split learning privacy.

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.

Accepted Papers

[1] 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.

[2] 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.

[3] **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***.

[4] 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.

[5] 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.

[6] 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).

[7] 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**.

[8] Li, B., Shao, Z., Liu, A., **Wang, P.** FedCliP: Federated Learning with Client Pruning. arXiv preprint arXiv:2301.06768 (2023).

[9] 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***.

[10] 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.

[11] **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).

Submitted Papers

[1] **Wang, P.**, Liu, X., Xiao, C. RePD: Defending Jailbreak Attacks Through a Retrieval-based Prompt Decomposition Process. In submission to NAACL 2025 (Meta-review: 4).

[2] **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. In submission to NAACL 2025 (Meta-review: 4).

[3] **Wang, P.**, Wang, H. DistDD: Distributed Data Distillation Aggregation Through Gradient Matching. arXiv, in submission to CPAL.

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

[5] **Wang, P.**, Wang, H. J., Wang, J. Reputation-assisted Network Collaboration Security System Based on Blockchain. In submission to IEEE Network.

[6] Shao, Z., Li, B., **Wang, P.**, Zhang, Y., Choo, K. K. FedLoRE: Communication-Efficient and Personalized Edge Intelligence Framework via Federated Low-Rank Estimation. In submission to TPDS.

[7] Liu, H., **Wang, P.**, Xing, T., Li, Y., Dalal, V., Li, L., He, J., Wang, H. Dataset Distillation via the Wasserstein Metric. In submission to CVPR.

[8] **Wang, P.**, Wang, H. Who's Behind the Curtain: Discover and Understand the LLM-based Chat Agent on the Social Network. In submission to ARR.

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

[10] 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. In submission to ICLR.

[11] **Wang, P.**, Wang, H. Understanding Political Stereotypes in Large Language Models through Ideological Testing. In submission to ARR.

[12] **Wang, P.**, Wang, H. J., Wang, J. Hierarchy Clustering-based Knowledge Transfer in Collaborative Intrusion Detection. In submission to Computer Network.

[13] **Wang, P.**, Lu, R., Li, C., Jiang, J., Wang, Z. DistPlanS: Automating Execution Planning for Distributed Deep Neural Networks. In submission to MLSys.

[14] Yi, L., **Wang, P.**, Yang Li. Split-and-Privatize Framework for Large Language Model Fine-Tuning, in submission to ARR 2025.

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