Ruikai Peng

peng@ruik.ai

WORK EXPERIENCE

Pwno **Mar. 2025 – Present**

Founder

- LLMs for memory security.
- We discovered over 60 previously undiscovered vulnerabilities across critical infrastructure, including Linux, V8, Chromium, FFmpeg, Firefox, PostgreSQL, MongoDB, and Redis, among the most heavily audited codebases in existence.

International Cybersecurity Olympiad (ICO) **May 2026 – Present**

US National Team *Tunisia*

- Selected as one of four U.S. National Team members

GGML **Mar. 2026 – Present**

Security *Remote*

- Help out security of Llama.cpp/GGML, LLM Inferencing.

OAIC **Oct. 2025**

Speaker *Oceanside, CA*

- Offensive AI Con (OAIC) is an invite-only technical conference dedicated to the use and development of AI for offensive cyber capabilities. Backed by Google DeepMind.
- Deductive Engine: Human-inspired Taint Reasoning. (my 3 months R&D project)
- Preparation based on Tree-of-AST from Black Hat. Deductive Engine found 17 CVEs and 3 0days in UniTree Robotics BLE stack

Black Hat **Aug. 2025**

Speaker *Las Vegas, NV*

- Black Hat, the world's leading cybersecurity conference series founded in 1997, draws 20,000+ attendees with annual flagship in Las Vegas
- Youngest Speaker in Black Hat USA History
- Black Hat USA: Thinking Outside the Sink: How Tree-of-AST Redefines the Boundaries of Dataflow Analysis
- Original R&D started on Feb 2024, Acceptance on May 25, 3 months

Independent **Sep. 2024 – Mar. 2025**

Independent Security Research *Avon, CT*

- Author of retr0.blog, 20,000+ monthly readers.
  - Featured/Republished by HackerNew, TheHackerNews, Checkmarx, Sonatype, Hackread, MalwareDotNews, InfosecWriteups, SecAlerts..
- ZeroCon25 (Seoul) Invited Speaker (with full honorarium) : Hardcore Inference Attack: Unraveling Llama.cpp's RPC Heap Puzzle
- ML Security Research / Application Security Research
  - Llama.cpp Distributed-Inferencing-Server RCE: Extensive research on Llama.cpp RPC and unique memory management; developed novel complex heap overflow exploitation techniques leading to Remote Code Execution (RCE) described in https://retr0.blog/blog/llama-rpc-rce.
  - Evernote RCE: Leveraged Electron's IPC mechanism and Evernote's internal BrokerBridge event listener to escalate the JavaScript injection into full RCE. This sophisticated exploit required reverse-engineering Evernote's obscured Electron application, detailed dynamic debugging, and constructing a multistep IPC payload, ultimately allowing attackers to silently execute malicious code on victim machines.
  - YoudaoNote RCE: Injecting malicious JavaScript payloads via LaTeX formula-rendering, bypassing Node.js integration restrictions, dynamically debugging Electron's internal IPC communications, and utilizing a

modified local cache to execute arbitrary executable files disguised as attachments. This chain enabled attackers to silently execute malicious code on victims' machines simply by viewing compromised notes.
    - Tenda AC8v4 Router RCE: Mips-based RCE through stack-based buffer overflow, employing ROP and register control techniques in order to bypass mitigation / limitations.
- ML Security Automation Development
    - AutoGDB.io: Founder and full-stack development of world's first dynamic debugging based binary-exploitation / reverse-engineering MCP SaaS AutoGDB. Reach two-hundreds users within two-days of beta stage.

Huntr **Dec. 2023 – Sep. 2024**

Security Researcher, Private Model-Format Threat Research | Aug. 2024 – Sep. 2024 *Remote*

- Located Model Format Security (Deserialization / Backdooring) Remote-Code Executions exploitation vectors in State-of-The-Art AI/ML Projects as TensorFlow, LlamaFile
- Identified Bypassing techniques on existing sophisticated Model-Format Security Systems, providing Hot-fixes / Mitigations on Identified Model-Format threats while developing threat-targeting scanner components. (Integrated as HuggingFace Third-party scanner: Protect AI)
- Contributed to Huntr's security blogs.

AI/ML Security Researcher | Dec. 2023 – Aug. 2024

- Located multiple critical vulnerabilities in state-of-the-art AI/ML projects; including Remote Code Execution (RCE) vulnerabilities in Transformers, Llama-cpp-python, PrivateGPT, PandasAI.
    - Llama-cpp-python Remote-Code Execution Supply-Chain Attacks: I discovered & exploited a Server-Side Template-Injection (SSTI) of llama-cpp-python in the GGUF format, affecting over 3,000 .GGUF Formats models, leading to Remote-Code Execution upon model deserialization, exposing a Supply-Chain Attack vectors for most of the exposed AI/ML inference endpoints. This exploitation is also known as "the-Llama-Drama" according to Checkmarx's review
- Located dozen critical vulnerabilities in LLM Inferencing endpoint security, working close with OSS Community on vulnerability patching / mitigating.

Tencent **Aug. 2023 – Aug. 2023**

Intern, AI Security and Binary Exploitation *Beijing, China*

- Intern as Tencent's T-Spark Talent Plan in Beijing (Youngest participant), exclusive Talent Plan of Tencent including NOI national team members and students from world-renowned institutions like MIT, Tsinghua, and Peking University.
- Researched in an advanced security research group with two research direction.
    - Traditional Security Research: Low-level Reverse-engineering of Telegram and exploited a sophisticated XSS to RCE zero-day in YouDao Note (~1 million daily active users in China) vulnerability entirely from scratch.
    - AI/ML Security Research: AI red team/blue team tackling high-level vulnerabilities such as prompt injection, context overflow, and linguistic-based attacks on large language models. Developed and implemented state-of-the-art defenses like IO detection and SoRA fine-tuning.
- Identified a unexpected zero-day cross-site scripting beyond the research requirements during the research process. Reported to the vendor, earned additional acknowledgment and credits for the discovery.

EDUCATION

The Webb Schools *Claremont, CA*

Avon Old Farms *Avon, CT*