# Ruikai Peng

retr0@retr0.blog | Canton, CT | Portfolio

## SUMMARY

Ruikai (Patrick) Peng is a 16-year-old founder, researcher, and writer at binary security.
He's the Youngest speaker at Black Hat USA, OAIC, ZeroCon.

From Llama.cpp heap overflows to Evernote RCE; Ruikai has found bugs in government systems, Google, Microsoft, and Intel. He's discovered RCEs in ML frameworks including Llama.cpp, Transformers, TensorFlow - amassing 27 CVEs over 25 RCEs by age 15. His work covers from ML to IOT security.

He runs a security blog, retr0.blog, over 20,000 monthly readers, where he story-tells his technical exploitation process. From a tens-of-thousands-word write-up on exploiting a heap overflow in Llama.cpp to step-by-step ROP chains for Tenda routers. His work was shared in multiple security publications (The Hacker News, Checkmarx, Sonatype, InfosecWriteups), and sparked discussions in YC Hacker News.

He's has turned his focus on finding cool vulnerability himself to finding cool vulnerabilities with AI. He presented his months of works and result at Black Hat, Offensive AI Con on thoughtful ways of bridging Transformers with low-level security, two fields he's most curious about.

He believes in the pureness of intellectual curiosity and creativity, and being kind, too.

## WORK EXPERIENCE

**Pwno** **Mar. 2025 – Present**

*Founder* *Avon, CT*

- AI in security of innermost working of computers.
- Full-stack development + R&D. From applied research (r&d, writing papers, talks) to full-stack (Langgraph, k8s), Founding team with two years in machine-learning security automation for binary-exploitation, working with research team at Tsinghua university, researcher from Google, MIT, CMU.
- Bloomberg coverage

**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* *Remote*

- 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

**Avon Old Farms** **Sep. 2024**

- Freshmen student; Working close with school Department of Techonology; Identified / Reported major physical / network vulnerabilities during Threat-Researching in a Private Facility. (H10301 HID based vulnerability allowing arbitrary entrance of protected buildings; and DMARC/SPF Based SMTP Spoofing); Reported severe vulnerability in boarding system REACH (reach.cloud) - school wide boarding control (student sign-in/out); that allows administration account-takeover. Received gratification from REACH CEO Brian Murray

## PROJECTS

**AutoGDB** **Dec. 2023 – Mar. 2024**

- Independently Developed AutoGDB: World's First Automatic / Dynamic Binary-Analysis Tool combining ML ReACT Reasoning and GDB Dynamic Debugging (This is before the emergent of the paper about the ML CTF Agent). AutoGDB were able to generate payloads that can solve real-life Binary-Exploitation challenges

**Educational Binary-Exploitation Series** **Jan. 2023 – Feb. 2024**

- 4.3K+ followers | 110K+ views | 10+ videos @ Bilibili
- Created niche binary exploitation tutorial series covering heap internals, glibc behavior, and memory corruption primitives.
- Bridged education and automation by making complex topics accessible to a new generation of researchers.

**AutoGDB.io (MCP SaaS)** **Dec 2023**

- Full-stack developed AutoGDB.op: World's First GDB enabled Automatic Dynamic Binary-Analysis/Exploitation MCP SaaS.
- Implemented MCP authorization based on original SSE session mechanism; Coordinated with GDB command backend (SSE+Backend) to accelerate the integration process

**PwnBERT** **Mar. 2023 – Apr. 2023**

- BERT-based vulnerability detection tool designed to identify and analyze Pwn-related vulnerabilities (e.g. UAF, heap overflow, etc.) in C language. By combining natural language processing techniques and security domain knowledge

**Chat-With-Binary** **Jan. 2023 – Apr. 2023**

- Independently Developed the World's First AI/ML Binary-Exploitation / Analysis Tool with individually based on RetDec / Retrieval-augmented generation (RAG). The project gain recognition and furthermore developed into chatwithbinary.com (binarychat.io, during development stage)

## CERTIFICATIONS, SKILLS & INTERESTS

- **Certifications:** 24 CVEs; Microsoft Security Response Center (MSRC) Q4 Leaderboard 2024; PicoCTF 24 (Ranked 10th/6957, 0.1438% Globally); PicoCTF 22 (Ranked 38/7794, 0.4875% Globally)
- **Skills:** Vulnerability Research; Binary-Exploitation; AI/ML Security; Application Security; Full-stack development; Electron Application Secuirty; Physical Security; Open Source Development; RF Security; HID Security; IoT Security; Firmware Security
- **Interests:** Songwriting; Playing John Mayer songs on my guitar; Writeup Writing; Snowboard; Watching 5 Broadway shows in three-days; Theater; Songwriting video making; Singing; FPV Drone Making; Late Night Thinking; Running; Creative Writing; Cycling; Walking in Foreign Cities; Jazz; Latin Jazz; Funk; Playing Life-is-Strange