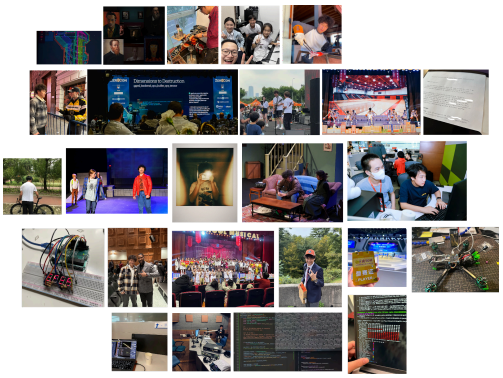
A few things we need:

- First, last name
  - Ruikai (Patrick) Peng
- Pronouns
  - he/him
- Researcher username
  - @retr0reg
- Twitter/Instagram handle (if you want to be tagged)
  - @retr0reg (X), @americanfish31 (Instagram)
- Photo of yourself



Little Abstract if you want to include *santerra.holler@bugcrowd.com*:

Ruikai (Patrick) Peng is a 15-year-old protégé security researcher in the intersection of ML security, binary exploitation and ML security automation, currently working with HiddenLayer ML Security Research. Ruikai uncovered RCEs in Machine-Learning frameworks including TensorFlow, Transformers, Llama.cpp, Llama-cpp-python, PrivateGPT. Beyond machine learning, Ruikai's research spans from sandbox-escapes RCE in Evernote, to ROP'ing Tenda Routers; From RCE in national examination system, to Privilege-Escalations in governmental education systems. While Ruikai had reported vulnerability and collaborated with big names Microsoft, Google, Evernote, and Managebac. His extensive research has been certificated in 25 CVEs. Ruikai is the youngest invited speaker for ZeroCon25, Seoul, and the youngest member of the Tencent Spark Talent Program.

Ruikai personal security blog, retr0.blog, attracts over 20k monthly readers, Ruikai's notable research includes Llama.cpp RPC heap-overflow to remote-code execution, exposing a Electron javascript-injection to remote-code execution escalation in Evernote, Llama-cpp-python supply chain attack via GGUF template injections, and intriguing ROP journey on Tenda Routers. His articles have been featured and republished by cybersecurity platforms, including The Hacker News, top-ten daily of YC HackerNews, Checkmarx, Sonatype, Hackread, MalwareDotNews, InfosecWriteups, Security Week, and SecAlerts.

"I love the low-levels, and things that curious me"

Tell us about yourself :)

- **What does your life look like outside of hacking (family/hobbies)?**
    - Oldest brother among two, dog *(Golden Retriever) /* cat *(Napoleon)* owner
    - I love everything, what that means is both I am interested in everything while treating them with love and joy.
        - **actor** (I really wanted to be an actor when I was around five, it just really seems like my thing. I have become less active and more shy recently but I still enjoy it), **booth operator** *(I was in the tech two years before I actually start acting)***, director** (I love writing my own movie when I was younger, and put it together by snapping and cropping in iMovie, and show my family with it)
        - **guitarist, songwriter** (*I can play Neon by John Mayer* and I love to do some write little songs when I am free (you can see on my instagram), **music-theory nerd**, and **pedal nerd** *(My favorite pedal would be Nordvang Gravity, with both Klon and TS10 channels, my Gravity is like 5th hand I found it on ebay, I love it)), mostly on my Instagram.* **producer** *(I learn about via youtube on music production, mixing, eq-ing, reverbing and space design… I have logic and bunch of plugins, my favorite one would be Cradle Hub's "The spirit", the doubler works great on vocals)*
        - ***midfielder*** *(soccer)* **cyclist**, **archer**, **runner, FPV pilot** *(I built my own drone from soldering, buying FCs)*
        - **photographer** (Since my mom gave me the little Canon camera since I was seven, I was in love with photographer, remember back in the summer breaks when I was younger, I would spend all day in my room figuring out how ISO works (internally) looking into different explanation, hour long videos, and I would write a whole page long camera internals down and post it to my wechat moments)
        - **content creator** *(I taught binary-exploitation and breaking down real-life exploits on Bilibili, 4.3K followers, 110K views),* **gamer** *(I built a register system in Minecraft, used to look into cheating for practising reverse-engineering),* **CG animator** (I took a while to learn how to use AE/PR/etc), **snowboarder** (I took about one and a half months spending each weekend in the ski park to figured out how to snowboard, I fell probably around fifty times, and trying to jump on the slope took me the most courage and time (after fluent edge curving), my neck still hurts now) …
        - I try to find my passion in everything, and love doing what I enjoyed!
- **Who influences you? (hacking and/or life)**
    - Sam Altman *(Y Combinator, OpenAI)*, Richard Hendrix (HBO Silicon Valley), Luis von Ahn *(Duolingo, CAPTCHA)*.

        - Sam's writings felt strangely personal, before I knew his background, I was into yc's ethos, later to realize he was part of the original founders who helped shape it, while doing research for my location-based campus network startup Lono (pivoted from Questy), I learned about Loopt, his early project. It felt like deja vu. His blog is really authentic and inspiring
        - Richard Hendrix, I love how *Silicon Valley* teaches you little lessons while having a laugh. You can also see Richard's growth from a nerd
        - Luis von Ahn, founder of Duolingo and *CAPTCHA, comp-sci professor in CMU and sold the company to Google twice. He's really an authentic and low-key figure, he's the proof that if you have an astonishing, momentum and persistence with execution, that will be enough to start great things (interview)
- **Do you enjoy sports or any sort of physical activity? *(lighter blue)***
    - I do theater! Always a secure and open space for me. I debuted as Troy Bolton in *High School Musical* During Middle School Production and that was a fantastic experience, after that, I played *Tierney* in *The Birds* (Yes the original of that Hitchcock Movie),
    - I did varsity soccer for awhile back in middle school, I can never forget getting on the field with dozens of my friends. We'd still play outside in the mid summer with shirts and our Khakis, we would be literally soaking wet after a fair match, and I still remember we'd split a bottle of ice coke.
- **Where did you grow up?**
    - I was born and raised in Shanghai, Pudong area. It's a nice place and rather the slower part of the city. Sometimes I miss going back there as it's still my most comfortable place to be. My favorite thing to do will probably be just riding my bike around the river with my headphones on, or just playing soccer on my old middle school field.
    - I am in Avon, Connecticut for my high school. Avon is a small typical new england town, fantastic environment and you're able to see the mountain (technically the hill) side, while the woods with a small running stream. I love taking a rest sitting on the bench with my earbuds on, just looking into the woods. It's relaxing even though it gets a little bit chilly sometimes.

- **Tell us a fun fact about yourself!**
    - Interests: Staring at crowds in Rockefeller Plaza, wondering what story each person carries; Late-night thinking, occasionally solving security flaws, occasionally overthinking texts; Obsessed with rainy window reflections and also reverse engineering HID protocols; Buying a Holden Caulfield hat that looks horrend tous but still feels symbolic; Writing notes that start as technical logs and end as personal essays; Believing every exploit has a story, and every person is a little unsolved puzzle
    - Fun fact: I spent a week building an 8-bit register system in Minecraft, and I can perform Troy Bolton's 'Getcha Head in the Game' with full choreography simultaneously!

Let’s talk about hacking!

- **What do you specialize in?**
  - I probably refer to it as the “low-levels”, I really like that word since it can be “low-level” as in binary-exploitation and working with registers, or can be “low-level” as in working in basic implementations and internals of frameworks and applications.
    - ML Security, binary-exploitation,
- **How long have you been hunting?**
  - I have always been looking but never really hunting.
    - I have been looking into the AI/ML field since the early 2024s, back in the langchain era, people were still confused about GPTs and Tools and agents were at the very beginning of being a thing.
  - The habit of looking a little deeper into the daily tangibles really leads me to some interesting finds: privilege escalation in the governmental education system, RCEs on examine taking systems, account takeovers in IB school systems and boarding school student management system *(both I heard back from the development team, actually the second one, REACH, the ceo wrote back to me)* HID authorization bypass in private facilities… (These’re all responsibly disclosed)
- **How has hacking changed your life?**
  - In a way that is already very hard to describe how much it already did.
  - Most impactful on how to deal with things with a more critical perspective, thinks things in a reversal mindset.
- **Can you talk about a vulnerability/bug that you think is particularly dangerous? Or, not talked about enough? Or, is something to look out for in the future? (let me know if this doesn’t make sense)**
  - Tokens. Tokens for are words *(or “meanings”)* for LLMs but haven't been translated in a way that they can understand *(semantically rich, high dimensional tokens)*, acts like an immediate-language between natural-language and hidden latents. (we project tokens on embeddings to map to latent, while during the last unembedding layer we calculate the possibility for next token by latent)
  - Tokens are fun,you can do really cool things with a token without really messing with the model's weight / hidden internals.
    - Manipulating tokenizer, the simplest tweak on the tokenizer.json can results in the most interesting result, fantastic writeup from Joe Lucas: https://developer.nvidia.com/blog/secure-llm-tokenizers-to-maintain-application-integrity/
- **Depending on whether you’re hardware, web or general, do you have any favorite tools or resources? What are they?**
  - GDB, the GNU debugger, allows you to see what is going on 'inside' another program while it executes -- or what another program was doing at the moment it crashed.

        - I love it because it is precision control over the most low-level components *(having everything under control,* seeing each register, heap allocations during an exact frame*)* while with fantastic community support.

Extra bits:

- **How do you use AI/MLs in your hacking?**
    - Great question, I imagine MLs as another version of me with the ability to multi-thread,
    - but I always want to reinforce how tech-debts can be fatal in the security research field, the
    - Cut corners, sloppy setup, half-checked proofs, flawed foundations. Leads to propagation effects
        - This is a little bit off topic, but tech-debts is structured in trees and graphs, imagine tech-debts branching out from another tech-debt. Tech Debts only leads to techdebts
        - Research depends on accumulation of experience and knowledge. False accumulation of knowledge can hurt back when we're not expecting.
        - After all in-depth is better in-breadth
- **How do you use ML automations in your attacks?**
    - AutoGDB w/ Binary Ninja MCP, go and try it out!
- **What are your thoughts on AI and how it'll change the landscape of cybersecurity?**
    - On the security researching (vulnerable) perspective (as AI as a target)
    - The fact that critical vulnerabilities in AI/ML Framework can exist in an "old-fashion" way, security is about "adaptation", novelty is definitely the key, but like how machine-learning is still optimizations problems in Calculus but in a more complex, clever and interesting way, same goes with ML Security, undoubtedly a part of it is something new that we haven't seen before at all (posing existential crisis to LLMs to)
    - With the lack of security in AI enabled programmings, you will see a bloom on the amount of reported vulnerabilities, and CVEs
    - I have seen lots of people pulling the coolest hack using the oldest tricks.
- **What vulns are common in AI/ML frameworks?**
    - Deserializations (Model Parsing), this can be pickle state machines, tensorflow operators, keras lambda layers and SavedModel *(this is more as a backdooring)*.
- **What is Llama.cpp? How do you hack it? (I've never heard of this before. I'm very curious about this)**
    - Llama.cpp is a low-level inference library written in C/C++, and today pretty much every locally deployed LLMs builds on it. Is written in a low-level way that you can describe it as the keystone/fundamentals of nowaday LLMs inferencing.
    - Turned a heap-overflow on tensor operations to remote-code execution (first exploitation after) by using Llama.cpp internals (memory layout) and interesting techniques (partial-writing, structure-oriented exploitation)

        - This allows you to fully take over computation RPC nodes and clusters (usually with massive computational resources) via an endpoint.
        - To draw an analogy: imagine OpenAI trying to serve millions of inferences. Instead of frying a single backend, they offload computation across many machines in a cluster. Each node handles a subset of the load. And exploitation in it means you can take over control of the very machines doing the heavy lifting.
    - At first with a prior month of research on the internal implementation of the RPC server, I was able to discover a limited heap-overflow caused by tensor-miscalculation. However, unique memory structures and implementations of llama.cpp RPC made it extremely difficult to escalate it to something meaningful. A novel exploit methodology only used in CTFs was able to turn the situation a bit; but then it comes with more setbacks and a more complex situation.
    - With 50 straight hours of GDB'ing, the exploitation was finally constructed with countless setbacks, obstacles, snags, and tangled paradox of memory state and object reuse. The exploitation finds its own way out of the heap-maze through nothing but weird behaviors, unpredictable object layouts, whimsical exploitation approach and working-with-what-you-got. There's actually an awe moment where one single line of code was able to solve the two major "paradoxes" in the exploitation journey.

- **How did you get into the Cybersecurity space? What's your origin story? / Tell us about your blog: retr0.blog**

    - retr0.blog is somewhere I wanted to blog about these intellectual-curiosity arousing *(I learned that word from yc hacker-news and I just can't stop using it)* cool-reading exploitations usually on a novel vector or target, which is the type of writeup I appreciate the most where I can learn cool stuff, and the type of research I am always working towards to.

    - I started technical writing when I was about eleven, where I found a simple way to exploit an information leak into a get-shell using a n-day exploit on my middle school's website. It was that kind of the most simple exploitation but it meant so much to me back then. (Always my dream is to get a little closer to Marcus Holloway, the O.G. Retr0 in WatchDogs2) I wrote my first writeup about it and posted it Zhihu sort of like Reddit + Quora in China, and it got a little attention. That little attention motivated me alot.
    - After that I was writing things on physical hackings and cool gadgets, GSM spoofing and Rubber-duckys, back then I would spend all my allowance just on a HackRF device (these like Chinese homemade ones, still expensive to me).

    - I started to get into little bit more advance techniques on applicational exploitation, mostly I started to look into things I used everyday, the daily tangibles; That's when I

discovered the mostly cool sounding ones I mentioned before: *(RCEs on examine taking systems, account takeovers in IB school), as we mentioned before)*

- After that I spend most of time on open-source software development, At the meantime diving into binary exploitation, house heap attacks that was extremely hard, remember I spent around a month just figuring out unlink attacks, the purpose was to teach things I spent days or even weeks to understand in within seconds of logical straightforward and accessible way (you can reference Grant Sanderson, author of 3Blue1Brown is something I admire), this is when I am focusing more on Tenda Router RCEs, kernel level Glibc allocation analysis.
    - On Aug 2023, my previous OSS projects on ML security automation, previous research and PicoCTF rankings took me into Tencent's T-Spark Talent program, where I worked with top college students in Tsinghua, Beijing University, MIT on AI/ML Security (the T-Spark program have different field focus: quantum-computations (they had to sign NDAs), machine visions). I was the youngest participant of all time (I was 14, others was around 19-20) additionally found an extra zero-day in the research process.

- Then I woke up one night and had an epiphany. I pivoted into trying to look for *"low-level, sophisticated and elaborate exploitation vectors that led to understandable severe consequences",* something like Google Project Zero, that's when I started putting essentials on retr0.blog, and start to look for something deeper and more interesting
    - ML Security researches: I started to research deep in ML projects, found bugs in Transformers, Tensorflow, these huge machine learning libraries, and inference framework PrivateGPTs… Llama-cpp-python which is one of the most effective supply chain vulnerabilities. (GGUF header template injections); Evernote IPC RCE, Youdao Note RCE, started to spend time on research that required larger time commitments and fundamentals.
- Llama.cpp heap-overflow exploitation was an exploitation that really required time and effort from the pre-research, discovery and exploitation, the whole exploitation was looking into a framework that no-one exploited before, not much of exploitation methodology you could reference and you have to figure each step by step with no guarantee next step would be exploitable at all. The overall process took around 3 months to complete (counting disclosure process). But it leads to speech invitation in security conference Seoul, opportunities from all over the field, from big names to hedge funds teams
    - I wouldn't say it directly leads to these opportunities, but the fact is the exposure told more people about what I did, the writing got on y-combinator hacker-news frontpage (Heap-overflowing Llama.cpp to RCE | Hacker News), brought 20,000 readers for my blog, and bought on to a famous technical daily sharing site (fefe.de) (the site is entirely germany, and itself bought 8k readers). I am grateful for that finding!

    - Now that I am doing researches that I love, working on few projects that bring out the curious out of me, working with wonderful people and figuring out what's next.

- **Do you have any advice for new hackers?** *(this and the next one is overlapping a bit, I'll just answer it in the one below:))*
- **What's an important lesson that you wish you learned early on in your hacking career?**
    - Simplicity over complexity.
    - Eliminate vagueness (Deep understanding over the target framework, architecture, tech stack is always the number one to go.)
    - Genuine intellectual curiosity is the best workmate for momentum.
    - Aim small to grow, aim big to jump.
    - Focus is extremely important.
- **Why do you hunt with Bugcrowd?**
    - fantastic community, matured triage/reward system, technical-heavy programs and VRT of Bugcrowd provides a deliberately considered standard between hackers vendors relation environment. Also the overall research community can be a cradle for quality research.

- **Your future is bright! What's next for you? / *You're so young, where do you see your journey going from here?***
    - That's a great question. Usually these is often time how I decide what's next

    1. I usually follow my passions and let curiosity guide me, this is the most dominant momentum derivative, I believe they'll take me where I'm meant to go. After wrapping up a project or research, I like to pause and see what sparks my curiosity next, then dive deep into it. Remember that genuine curiosity is often what drives real depth and discovery, counterintuitively.
    2. I follow the steps of others who pushed changes and impacts to the world, and walk along the path of becoming the person I want to become. *(This is similar to the who-inspire-me-the-most question)*. Learning about people who accomplished great things, find similarities between them, however instead of becoming them, more on how they can be a part of me?

    - This is great because it gave me both Idea-driven and Goal-driven[1] option, usually I feel happy when I realize I am on the route of both/either of them, but this doesn't mean I throw away everything that's on the way / not related.

    - The growing process of me is always about learning to interweave and braid these two parts deliberately, I gradually start to form a more vague-less path on the roads ahead, treating each decision deliberately. But I wouldn't try to predict where I am in a few years or more, rather go with the flow, keep learning while

[1] http://joschu.net/blog/opinionated-guide-ml-research.html, *Roughly speaking, there are two different ways that you might go about deciding what to work on next…*

working with what I have and trying to earn and make the best opportunity and decision.

- **Hacking is hard work and you're balancing school and home with it. How do you take care of your mental health and avoid burnout?**
    - It's extremely hard, especially between trying to start new things, security research, and paying enough attention to the emotional world while getting on track of the real world and the academics. Focus, prioritize and sometimes you have to trade off one for another and think deliberately
    - Taking a break is most essential for burnout, on this line of field it usually comes with imposter syndrome and anxiety. I always felt so much behind and I am not keeping up with the work.
    - But the fact is if this helps, the most badass researches I read about came from a random glance or a side thought, although behind it might be considerably large amount of knowledge base, but what I mean is don't overwhelm yourself, build your own introspective value matrix instead letting someone else to tell you what you're worth, and always cut yourself some leeways.

-
- **What are some goals you have for this year?**
- **What is your ideal career?**
    - On a path of starting something big *(vision, changing-the-world-ish)* that I am passionate about.
- **Anything else you want to include!**