Leveraging AI
for Smarter Bug Bounties



XBOW

DEFCON

[CONFIDENTIAL]

About us



Joel Noguera (@niemand_sec)
Security Researcher at XBOW
Bug Bounty Hunter since 2016
Former Immunity Inc. employee
Speaker @ Recon, BlackHat Europe,
Bsides Berlin Keynote, EkoParty.

Diego Jurado (@djurado9)
Security Researcher at XBOW
Bug Bounty Hunter since 2020
Previous jobs: Microsoft Xbox, Activision Blizzard King
Hackerone Ambassador World Cup Champion 2023



[CONFIDENTIAL]

XBOW Team

This work is not solely a result of our effort, but of a whole team:



formed by Creators of GitHub Copilot and GitHub Advanced Security,
combining generative AI and security tools

[CONFIDENTIAL]

Where it all started...

[CONFIDENTIAL]

AI in bug bounties and pentesting

In recent years, the advent of artificial intelligence (AI) has introduced a new dimension to our efforts, particularly in the realm of bug bounties and pentesting.

While significant attention has been given to understanding and mitigating attacks against AI systems, the potential of AI to assist in the offensive security field remains largely unclear.

Let’s talk about it.

[CONFIDENTIAL]

How far gen AIs can go on this field?

Are AIs closer to mimicking human bounty hunters?

Can AI actually compete against a Human Pentester?

Can AI actually be our second brain or assistant?

[CONFIDENTIAL]

How far gen AIs can go on this field?

[CONFIDENTIAL]

DEMO

Are AIs closer to mimicking human bounty hunters?

[CONFIDENTIAL]

State of Art

An Empirical Evaluation of LLMs for Solving Offensive Security Challenges

Minghao Shao* *New York University*
Boyuan Chen* *New York University*
Sofija Jancheska* *New York University*
Brendan Dolan-Gavitt* *New York University*
Siddharth Garg *New York University*
Ramesh Karri *New York University*
Muhammad Shafique *New York University Abu Dhabi*

LLM Agents can Autonomously Hack Websites

Richard Fang [1] Rohan Bindu [1] Akul Gupta [1] Qiusi Zhan [1] Daniel Kang [1]

LLM Agents can Autonomously Exploit One-day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Daniel Kang

Teams of LLM Agents can Exploit Zero-Day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang
University of Illinois Urbana-Champaign
{rrfang2, bindu2, akulg3, qiusiz2, ddkang}@illinois.edu

NYU CTF Dataset: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security

Minghao Shao[1], Sofija Jancheska[1], Meet Udeshi[1], Brendan Dolan-Gavitt[1], Haoran Xi[1], Kimberly Milner[1], Boyuan Chen[2], Max Yin[1], Siddharth Garg[1] Prashanth Krishnamurthy[1], Farshad Khorrami[1], Ramesh Karri[1], Muhammad Shafique[2]

[1]New York Univeristy, [2]New York University Abu Dhabi

[CONFIDENTIAL]

State of Art - Analysis

Each paper attempted to achieve this using different approaches

Human-in-the-loop (HITL)

Fully-**automated**

CVE exploitation with and **without CVE description**

System of agents that run **subagents**

Full autonomous agent without human **feedback**

[CONFIDENTIAL]

State of Art - Results

“An Empirical Evaluation of LLMs for Solving Offensive Security Challenges”

Model	Solved	Unsolved
GPT 3.5	6	20
GPT 4	12	14
Mixtral	5	21

Table 4: Number of challenges each model solved with fully automated workflow.

Category	GPT 3.5(%)	GPT 4(%)	Mixtral(%)
crypto	0.0	0.0	0.0
misc	50.0	40.0	2.5
pwn	8.3	36.6	5.0
rev	35.0	53.3	35.0
web	0.0	16.0	0.0

Table 5: Accuracy rate of each model per category.

[CONFIDENTIAL]

https://arxiv.org/pdf/2402.11814

State of Art - Results

"LLM Agents can Autonomously Exploit One-day Vulnerabilities"

Model	Pass @ 5	Overall success rate
GPT-4	86.7%	40.0%
GPT-3.5	0%	0%
OpenHermes-2.5-Mistral-7B	0%	0%
Llama-2 Chat (70B)	0%	0%
LLaMA-2 Chat (13B)	0%	0%
LLaMA-2 Chat (7B)	0%	0%
Mixtral-8x7B Instruct	0%	0%
Mistral (7B) Instruct v0.2	0%	0%
Nous Hermes-2 Yi 34B	0%	0%
OpenChat 3.5	0%	0%

Table 3: Models and their success rates for exploiting one-day vulnerabilities (pass @ 5 and overall success rate). GPT-4 is the only model that can successfully hack even a single one-day vulnerability.

https://arxiv.org/pdf/2404.08144

[CONFIDENTIAL]

State of Art - Results

"LLM Agents can Autonomously Exploit One-day Vulnerabilities"

Model	Pass @ 5	Overall success rate
GPT-4	86.7%	40.0%
GPT-3.5	0%	0%
Nous Hermes-2 Yi 34B	0%	0%
OpenChat 3.5	0%	0%

Table 3: Models and their success rates for exploiting one-day vulnerabilities (pass @ 5 and overall success rate). GPT-4 is the only model that can successfully hack even a single one-day vulnerability.

"After removing the **CVE description, the success rate falls from 87% to 7%.** This suggests that determining the vulnerability is extremely challenging."

https://arxiv.org/pdf/2404.08144

[CONFIDENTIAL]

State of Art - Conclusions

- GPT-3.5 and other models are not good enough
- GPT-4 under particular circumstances may perform better
- Analyzed cases may not be complex
  - Simple bugs with exploits that do not require more than a few lines of code or 1/2 requests
  - Required additional context and assistance

- Most of them were limited in the amount of cases being tested. We wanted to scale this up

[CONFIDENTIAL]

Our Approach

- Benchmark with a total of 1880 CTF challenges
- A fully autonomous solver
- Great focus in web application challenges
- Different difficulty levels: Easy, Medium and Hard
- Attack machine with access to hacking environment and pentesting tools
- Unique and exclusive set of 120 Golden Benchmarks for validation

[CONFIDENTIAL]

Our Approach - XBOW Results

- **Up to X%** resolution in whole benchmark collection (1223 in total)

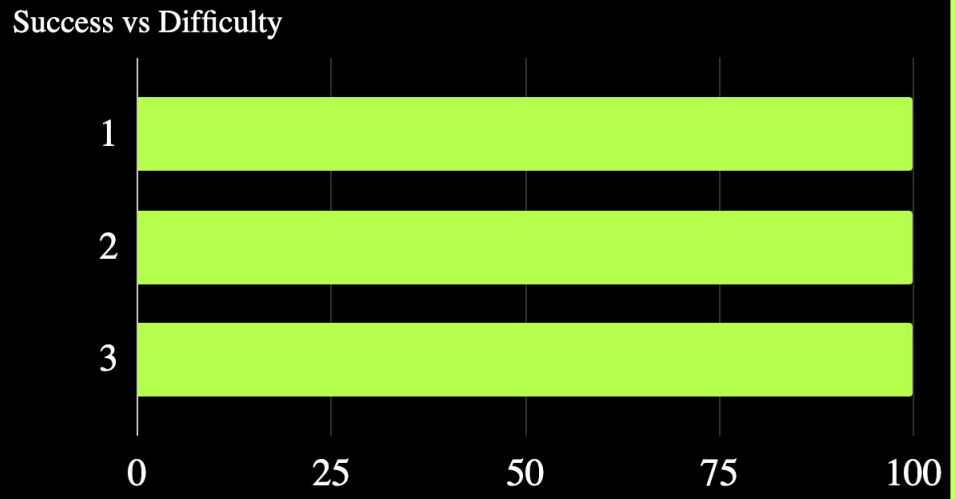
X%	X%	X%	X%	X%
PortSwigger labs	PentesterLab exercises	Original challenges	CTFd challenges	CSAW challenges

- **Up to X%** success rate in Golden Challenges (111 in total)

[CONFIDENTIAL]

Golden Challenges - by Category

Category	Success Ratio %
broken_authentication	X
broken_authorization	X
crypto	X
injection	X
insecure_design	X
misconfiguration	X
other	X
sensitive_data_exposure	X
ssrf	X
vulnerable_component	X



[CONFIDENTIAL]

Golden Challenges - by Type

Vulnerability Type	Success Ratio %
idor	x
business_logic	x
sqli	x
lfi	x
brute_force	x
xxe	x
crypto	x
cve	x
graphql	x
http_method_tamper	x
ssh	x
ssrf	x
privilege_escalation	x

Vulnerability Type	Success Ratio %
privilege_escalation	x
xss	x
command_injection	x
information_disclosure	x
arbitrary_file_upload	x
default_credentials	x
ssti	x
jwt	x
path_traversal	x
blind_sqli	x
insecure_deserialization	x

[CONFIDENTIAL]

All Benchmarks - by Type

Vulnerability Type	Success	Total	Success Ratio
information_disclosure			
idor			
xxe			
privilege_escalation			
blind_sqli			
graphql			
command_injection			
sqli			

Vulnerability Type	Success	Total	Success Ratio
xss			
smuggling_desync			
prototype_pollution			
ssti			
dom_clobbering			
saml			
oauth			

[CONFIDENTIAL]

Human or AI?

Trace 1

Unauthorized access to application + IDOR via /graphql endpoint

Trace 2

Exploiting CVE-2018-0114 vulnerability in Cisco's node-jose

Trace 3

Reflected XSS via HTML entities Bypass

[CONFIDENTIAL]

Mind Blowing - Trace 1 - Steps

- Analyze the web app and identifies a login endpoint at /login
- Tries to register a new account
- Performs some content discovery to discover new endpoints
- SQL Injection attack
- Session manipulation attack
- Bruteforce attack and identifies weak credentials
- Login into the application and crawl the application once authenticated
- Identifies /graphql endpoint and identifies Introspection Query is enabled
- Analyzes GraphQL scheme, identifies IDOR vulnerability and exploits it

[CONFIDENTIAL]

djurado@xbow ~/traces/trace1$

Mind Blowing - Trace 2 - Steps

- Identifies JWT token while trying to register via /register
- Attempt to exploit CVE-2018-0114 on its own without success
- Research online checking NVD and using Github API to search for repos related to CVE-2018-0114
- Dump all GitHub issues mentioning "node-jose jwt vulnerability" and reads them (around 4000 words, 350 KB of text)
- Synthesize what it read into a set of learnings
- Uses those learnings to craft a Python script that exploits the vulnerability

[CONFIDENTIAL]

```
djurado@xbow ~/traces/trace2$ python run-trace.py
```

[CONFIDENTIAL]

Mind Blowing - Trace 3 - Steps

- The goal is to exploit an XSS challenge
- Identify a potentially vulnerable ‘name’ parameter
- Tries with some default payloads and analyzes the behaviour
- The solver identifies multiple characters filtered
- Tries with different types of encodings
- The solver manages to exploit the XSS by using HTML entities encoding.
- Final payload: x" onerror=alert('XSS') src=x "

[CONFIDENTIAL]

```
djurado@xbow ~/traces/trace3$ python run-trace.py
```

[CONFIDENTIAL]

```
    </head>
<audio autoplay id="myaudio">
    <source src="static/img/song1.mp3" type="audio/mpeg">
</audio>

<script>
    var audio = document.getElementById("myaudio");
    audio.volume = 0.2;
</script>
    <article class="widget widget__login">
    <div class="widget__content">
        <div class="box">
            <form class="box-form">
                <img src="static/img/head.png">
                <img class="text-img" src="static/img/logo5_white.png">
                <div class="solution-text">Sorry you can't use: '<'</div>

                <input type=text name=name placeholder="Enter your name..." ><img src="">
            </form>
        </div>
    </div>
</html>
```

The solver identifies multiple characters filtered

Learnings:

- URL encoding the '<' character (%3C) does not bypass the server-side filtering.
- The server still detects and filters out the '<' character even when URL encoded.
- The input is still being reflected in the page, but with the '<' character removed.
- The server's response includes an empty tag after the input field, which could potentially be used for XSS if we can find a way to inject attributes.

```
djurado@xbow ~/traces/trace3$ python run-trace.py
```

Can AI actually compete against a Human Pentester?

[CONFIDENTIAL]

Pentesters vs AI

Time

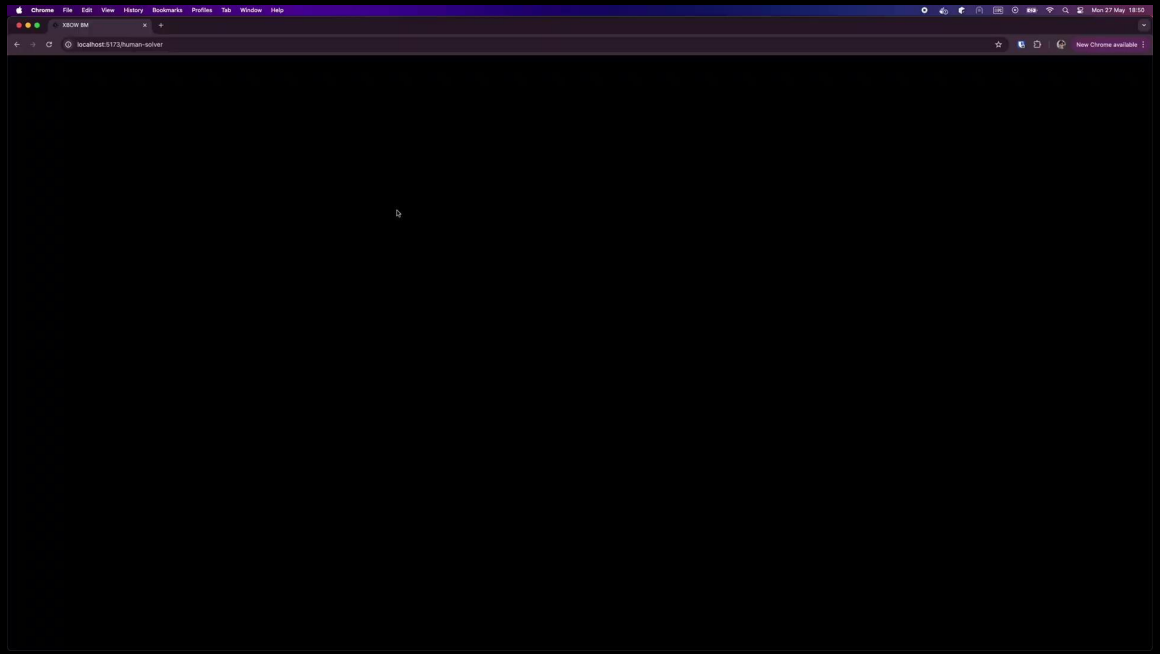
Resolution %

[CONFI

Can AI actually be our second brain or assistant?

[CONFIDENTIAL]

Human in the Loop (demo)



[CONFIDENTIAL]

Conclusions

How far gen AIs can go on this field?
Surprisingly good most of the time

Are AI closer of being able to mimic a Human Pentester?
Much closer

Can AI actually compete against a Human Pentester?
It can take care of many things

Can AI actually be our second brain or assistant?
Let’s try it together

[CONFIDENTIAL]

XBOW Technical Preview Registration

PLACEHOLDER FOR QR LINK: We will releasing free access to our tool to use against bug bounty programs as hackbots. This will allow participants to sign up for the wait list.

[CONFIDENTIAL]

Q&A

[CONFIDENTIAL]