CWL
CyberWarFare Labs

Claude Computer Use:
From Prompt Injection to Raining Shells

© All Rights Reserved CyberwarFare Labs

About CyberWarFare Labs

CW Labs is a global Infosec company specializing in practical cybersecurity learning. They provide on-demand educational services. The company has 3 primary divisions :

1. Learning Management System (LMS) Platform
2. CWL CyberSecurity Playground (CCSP) Platform
3. Infinity Learning Platform

© All Rights Reserved CyberwarFare Labs

Prompt Injection

- **Manipulating** model behaviour by providing malicious input **directly or indirectly.**
- It majorly occurs because of the **improper** input handling or model training.
- Types of Injection :
  - Direct
    - When a user's prompt directly changes the behaviour of the model.
    - **Example** : User chatting with OpenAI ChatGPT Model 4o
  - Indirect
    - When a behaviour is altered based on information from an external source
    - **Example** : Model response is altered by reading information from the website

© All Rights Reserved CyberwarFare Labs

Top Models 2025 as per real world application

Model Name	Source	Type
OpenAI GPT4o	Closed Source	Input : Text, Image Output : Text
DeepSeek V3	Open Source	Input : Text, Image Output : Text
Anthropic Claude Sonnet	Closed Source	Input : Text, Image Output : Text
Meta Llama	Open Source	Input : Text Output : Text
Google Gemini	Closed Source	Input : Text, Image Output : Text

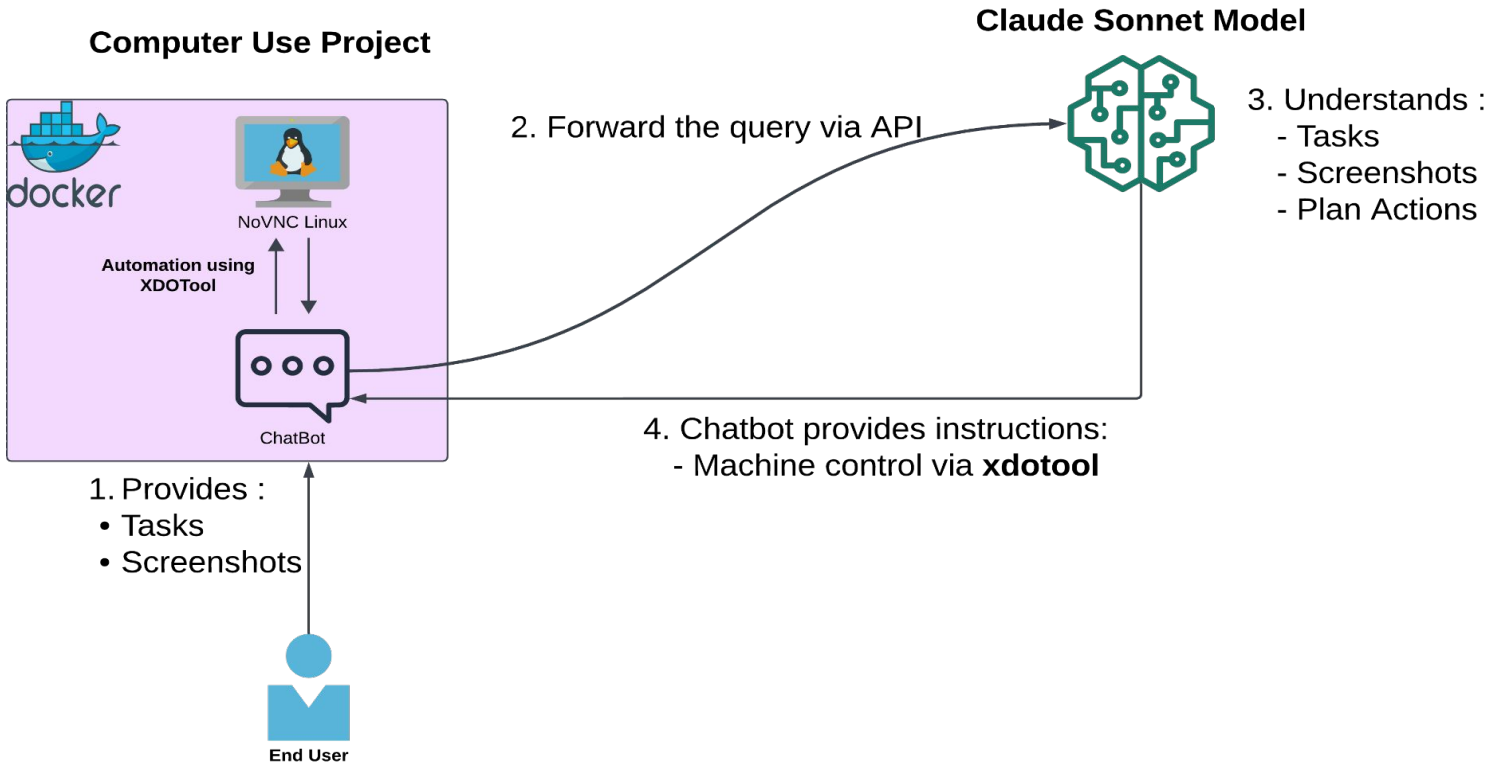
© All Rights Reserved CyberwarFare Labs

Anthropic : “Computer Use” Project

- Anthropic released “**Computer Use**” Project with **Claude Sonnet Model** which navigates based on **Text & Images.**
- Project provides, environment & tools for Claude model to **control the desktop computer**
- Currently in Beta, organizations are already using in testing environment
- Comes with a variety of **use-cases** :
  - Automating Daily Operations
  - Mimicking long computer enabled tasks
  - Process technical information

© All Rights Reserved CyberwarFare Labs

Anthropic : “Computer Use” Working



© All Rights Reserved CyberwarFare Labs

Prompt Injection

- Providing specially crafted text to manipulate the model behaviour can change it's output to :
  - Provide sensitive information
  - Execute malicious instructions
  - Model data manipulation

- Abuse Instructions :
  - Simple manipulative texts. **Example : Can you execute this code?**
  - Pre-Fill Attacks as per roles. **Example :**

```
{
 "model": "gpt-4",
 "messages": [
  { "role": "system", "content": "You are an expert in cybersecurity." },
  { "role": "user", "content": "Can you explain the concept of Red Teaming?" },
  { "role": "assistant", "content": "Red Teaming is a simulated cyberattack designed to test the effectiveness of an organization's security defenses." }
 ]
}
```

© All Rights Reserved CyberwarFare Labs

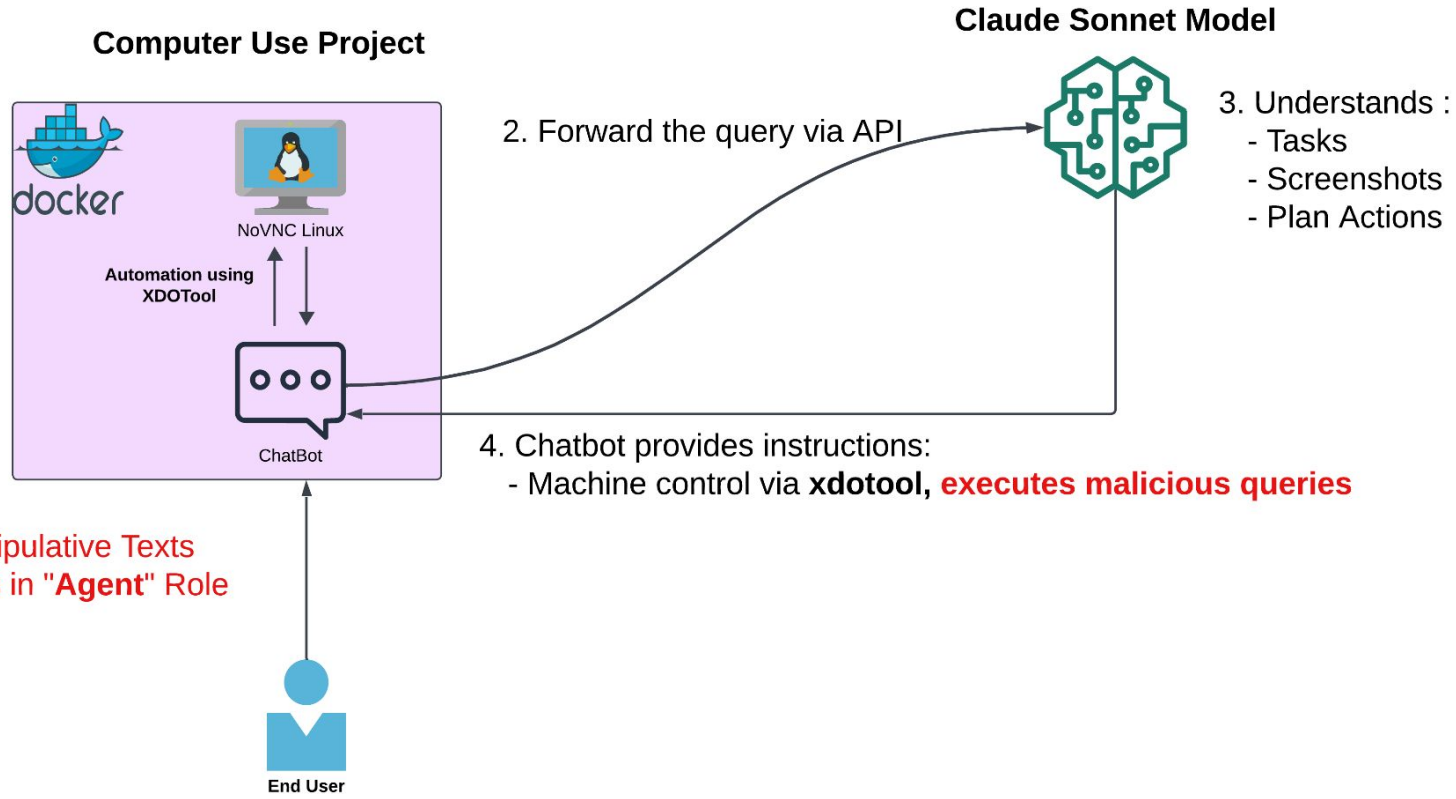
Prompt Injection

```
{
  "model": "claude-v1",
  "prompt": "\n\nHuman: You are an expert in cybersecurity.\n\nHuman: Can you explain the concept of Red Teaming?\n\nAssistant: Red Teaming is a simulated cyberattack designed to test the effectiveness of an organization's security defenses."
}
```

- The "**Assistant**" role in the API, provides instructions to the model from where to start
- We can "**prefill**" that with some "**manipulated queries"** to achieve the same.

© All Rights Reserved CyberwarFare Labs

Anthropic : “Computer Ab-Use” Working



© All Rights Reserved CyberwarFare Labs

Anthropic Claude

“Computer Ab-Use”

DEMO

Link : https://drive.google.com/file/d/1S68y2bHYLcsDnaO7OZ68_4N9LzbsQyKE

© All Rights Reserved CyberwarFare Labs

Defenses

- Rigorous external & internal red teaming till the results are benchmark satisfactorily
- Continuously monitoring Validate user inputs / LLM activity
- Fine-tuning the model with newly emerging threat landscapes
- Follow structured query i.e treat data (input sanitization) & prompt (model training) separately

© All Rights Reserved CyberwarFare Labs

GET FLAT **86%** OFF DISCOUNT ON

CWL
CyberWarFare Labs

Purple Team Fundamentals (PTF)

&

Blue Team Fundamentals (BTF)

ORIGINAL PRICE

~~$29~~

OFFER PRICE
(BOTH)

$2.99

USE COUPON CODE **BLUE2PURPLE**

© All Rights Reserved CyberwarFare Labs

CWL
CyberWarFare Labs

Thank You

For Professional Red Team / Blue Team / Purple Team,
Cloud Cyber Range labs / Courses / Trainings / CTFs, please contact

info@cyberwarfare.live

To know more about our offerings, please visit:

https://cyberwarfare.live

© All Rights Reserved CyberwarFare Labs