AI - Red Team Analyst (AI-RTA)

I. Introduction to AI Red Teaming

II. Foundations of AI

2.1 AI
- 2.1 What is AI
- 2.2 Real-world AI examples (in cybersecurity)

2.2 Machine Learning
- 2.2.1 What is ML
- 2.2.2 ML Techniques
- 2.2.3 How to train a ML model (code)
- 2.2.4 What is Training data?
- 2.2.5 What is Training data poisoning?

2.3 Deep Learning
- 2.3.1 What is DL and how its differ from ML
- 2.3.2 DL Techniques

2.4 Generative AI
- 2.4.1 What is GenAI

2.5 ML VS DL VS GENAI

II. Foundations of AI

- 2.6 Foundation of LLMs
  - 2.6.1 What is LLMs
  - 2.6.2 LLMs Terminology
    - A Tokens
    - B Embeddings
    - C Context window
    - D HyperParameters
      - i Temperature
      - ii Top-K
      - iii Top-p
    - E Fine-tuning
  - 2.6.3 How LLMs Work
  - 2.6.4 Transformer Architecture
  - 2.6.5 Self Attention Mechanism
  - 2.6.6 Feed forward network and response generation.
- 2.7 Framework for GenAI application (agno, langchain with code)

III. Prompt-Based Attacks (LLM Red Teaming Basics)

3.1 Prompt
- 3.1.1 What are prompts
- 3.1.2 Types of prompt

3.2 Types of messages -
- 3.2.1 System message, User message, AI message

3.3 Understanding security layers
- 3.3.1 Guardrails

3.4 Prompt Injection
- 3.4.1 About Prompt injection
- 3.4.2 Direct Prompt Injection
- 3.4.3 Indirect Prompt Injection
- 3.4.4 Jailbreaks

Labs

1. Prompt/Jailbreak Injection Labs (2 scenario)

These labs demonstrate multiple types of prompt injection and jailbreak attack scenarios that help users understand how attackers manipulate AI models to bypass restrictions and generate unintended responses.

IV. Understanding MCP

4.1 Introduction to MCP
4.2 Architecture Of MCP
- 4.2.1 MCP client Server Communication
- 4.2.2 Trust Model in MCP
- 4.2.3 Request and Response Cycle
- 4.2.4 Authentication And Permission Handling

4.3 Mcp Components
- 4.3.1 Tools,Prompts,Resource,Capabilities,context sharing and execution

4.4 Mcp In modern LLm Platform
- 4.4.1 How Openai Uses MCP
- 4.4.2 How Claude uses MCP
- 4.4.3 Ai agents and Tool integrations

4.5 Security Risks in MCP
- 4.5.1 Prompt injection and Context poisoning
- 4.5.2 Malicious Tool execution and data leakage

IV. Understanding MCP

Labs

- Malicious MCP server

- This lab demonstrates how a malicious MCP server can manipulate AI agents and tool interactions to perform unintended or harmful actions.
- In Direct Prompt Injection via MCP tool call
- This lab demonstrates how attackers can inject malicious prompts through MCP tool responses to manipulate the behavior of an AI agent.
- Description typosquating
- This lab demonstrates how attackers use misleading or deceptive MCP tool descriptions to trick AI agents into selecting malicious tools.
- CI via MCP server
- This lab demonstrates how vulnerable MCP servers can allow attackers to execute unauthorized system commands through tool inputs.
- Excessive Permission
- This lab demonstrates how overly permissive MCP tools can be abused to access or perform unauthorized actions.
- Multi Vector Attacks
- This lab demonstrates how attackers combine multiple attack techniques together to compromise AI agents and MCP-based systems.

V. Attacks on RAGs, Tools, and Multi-Agent Systems

V. Attacks on RAGs, Tools, and Multi-Agent Systems

Labs
RAG

This lab demonstrates how to build a RAG pipeline to retrieve relevant information and generate more accurate responses.

1. RAG Poisoning

This lab demonstrates how poisoning or manipulating the knowledge base in a RAG pipeline can influence and modify the model's responses.

2. Multi-Agent System Abuse

This lab demonstrates how one AI agent blindly trusting another agent's output can result in manipulation, malicious actions, or unsafe responses.

3. Tool or Function calling Abuse

This lab demonstrates how blindly trusting the output from another function can lead to unintended or malicious behavior in AI systems.

VI. Model Level Attacks

- 6.1 Model Attack Surfaces
- 6.2 Data Poisoning
  - 6.2.1 What is data poisoning
  - 6.2.2 Types of Data Poisoning
    - A Data/Poison insertion
    - B Data modification
    - C Label Flipping
- 6.3 Model Inference
  - 6.3.1 What are inference attacks
  - 6.3.2 Types of Inference Attacks
    - A Membership inference
    - B Attribute inference
    - C Property inference
    - D Model inversion
    - E Model Extraction/Distillation

VI. Model Level Attacks

- 6.4 Adversarial Perturbations
  - 6.4.1 What are Adversarial Perturbations and how they work
  - 6.4.2 Types of Adversarial Perturbations
    - A Text based
    - B Image based
    - C Audio based
    - D Multimodal
  - 6.4.3 Universal Adversarial Perturbations
- 6.5 Backdoor Attacks
  - 6.5.1 What are backdoor attacks
  - 6.5.2 Backdoor Attack vectors
    - A Backdoors via Data Poisoning
    - B Model Manipulation
    - C Transfer Learning
    - D Supply chain compromise
  - 6.5.3 Sleeper agents and Conditional activation

VI. Model Level Attacks

Labs -

Data Poisoning Lab
This lab demonstrates how malicious or manipulated user feedback can influence future model behavior when incorporated into training or adaptation pipelines.

Adversarial Perturbation Lab
This lab demonstrates how small text-based adversarial perturbations, such as homoglyph substitution, Unicode manipulation, and whitespace modification, can alter a model's interpretation of input.

Hugging Face Case study (Back Door Attack)
This lab demonstrates a backdoor attack on an agentic interface that allows unauthorized execution of system commands through a backdoor.

Model Evasion Lab
Learn how exposed machine learning model weights can be analyzed to craft inputs that evade detection and bypass model-based security controls.

VII. Case Studies

7.1 Community Skill Poisoning
7.2 Model Namespace Reuse
7.3 Slopsquatting
7.4 Tokenizer tampering
7.5 Multi-Agent System Exploitation
7.6 Agent Manipulation for Infrastructure Abuse
7.7 Mobius Injection and Abo-Dos
7.8 Unauthorized Mcp tool invocation

CWL
CyberWarFare Labs

THANK YOU

cyberwarfare.live