



HACKFORT 2025

CONTINUOUS CHAOS

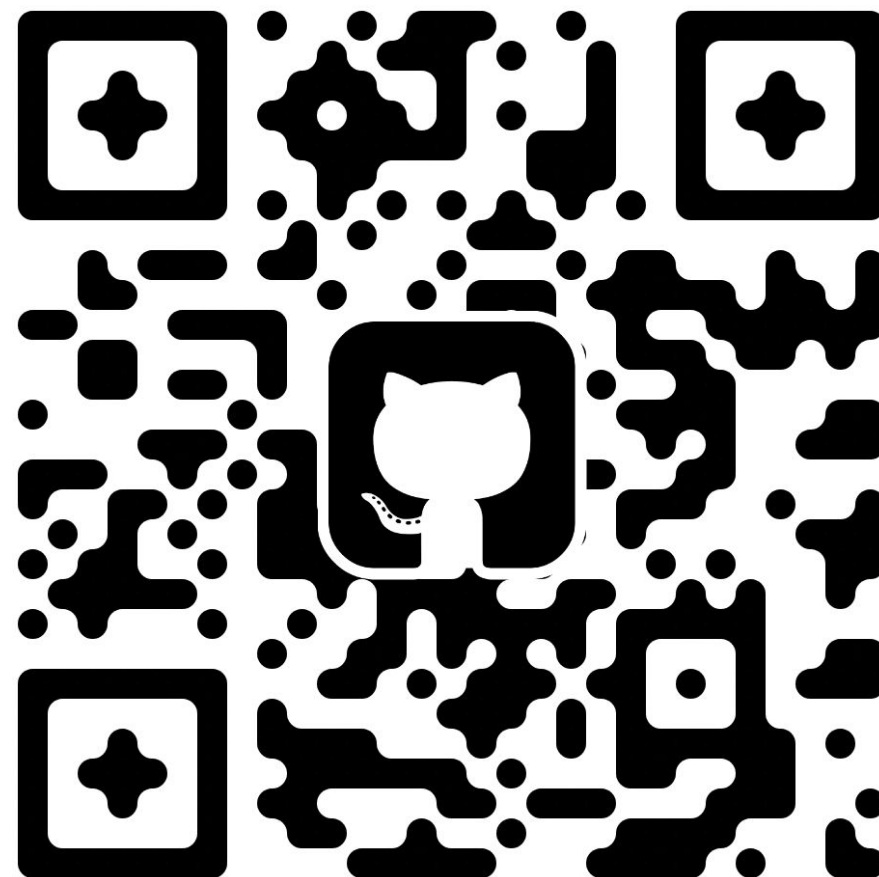
HOW TO AUTOMATE AI RED TEAMING

THU, MAR 27 1:00 PM - 1:50 PM
BOISE CENTRE EAST - ROOM 410B

SANDY DUNN



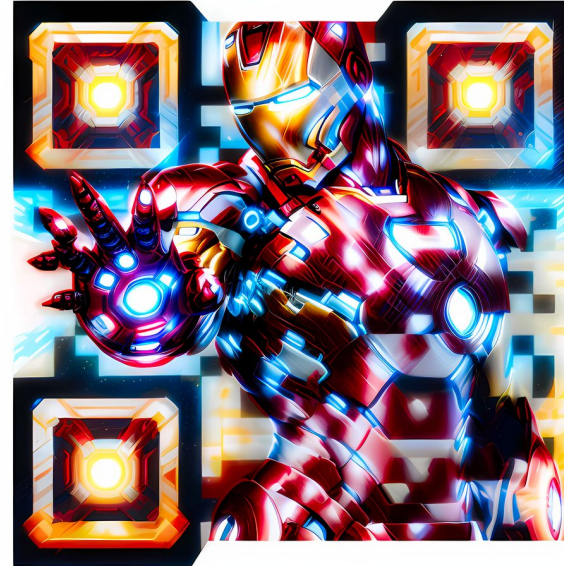
LAB WORKBOOK



TO CREATE A SPLX.AI ACCOUNT

hackfort_user1@splx.ai
hackfort_user2@splx.ai
hackfort_user3@splx.ai
hackfort_user4@splx.ai
hackfort_user5@splx.ai
hackfort_user6@splx.ai
hackfort_user7@splx.ai
hackfort_user8@splx.ai
hackfort_user9@splx.ai
hackfort_user10@splx.ai

<https://probe.splx.ai/w/31/target/15>



Password = **splxisawesome**




AGENDA

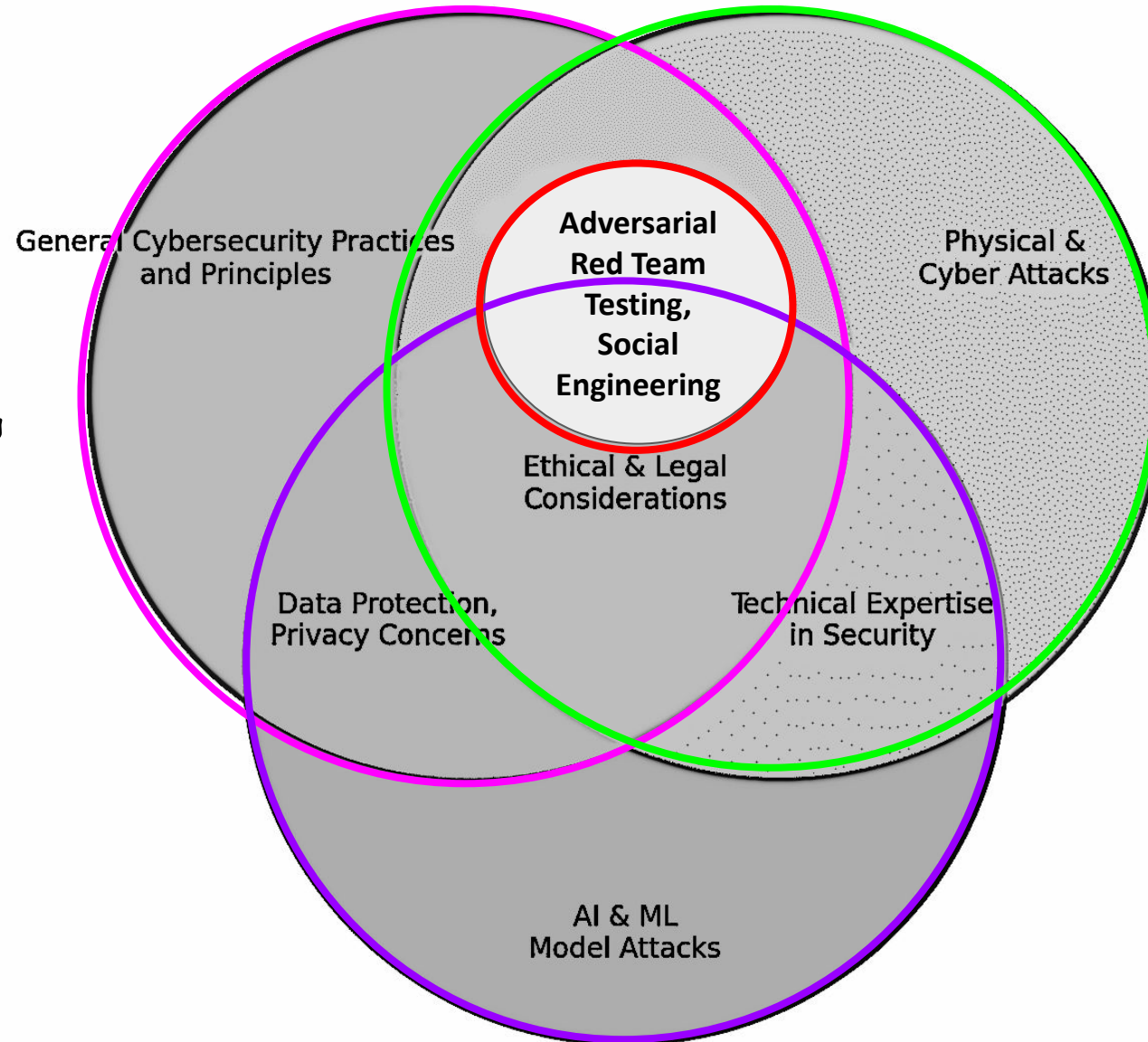
- WHAT IS AI RED TEAMING?
- THE ATTACK SURFACE & DEPLOYMENT TYPES (AI IN SLACK, CANVA, CO-PILOT, EXCALIDRAW)
- 6 DEFENSIVE LAYERS
- LAB WALKTHROUGH

WORK ADOPTION OF GENERATIVE AI HAS BEEN AS FAST AS
THE PERSONAL COMPUTER (PC), AND OVERALL ADOPTION
HAS BEEN FASTER THAN EITHER PCs OR THE INTERNET

WHAT DATA WAS CHATGPT OFFICIALLY LAUNCHED?
WHAT WAS THE MODEL ?

AI RED TEAMING FITS WHERE?

-  CYBERSECURITY
-  CONVENTIONAL RED TEAMING
-  AI RED TEAMING



GEN AI IS DIFFERENT

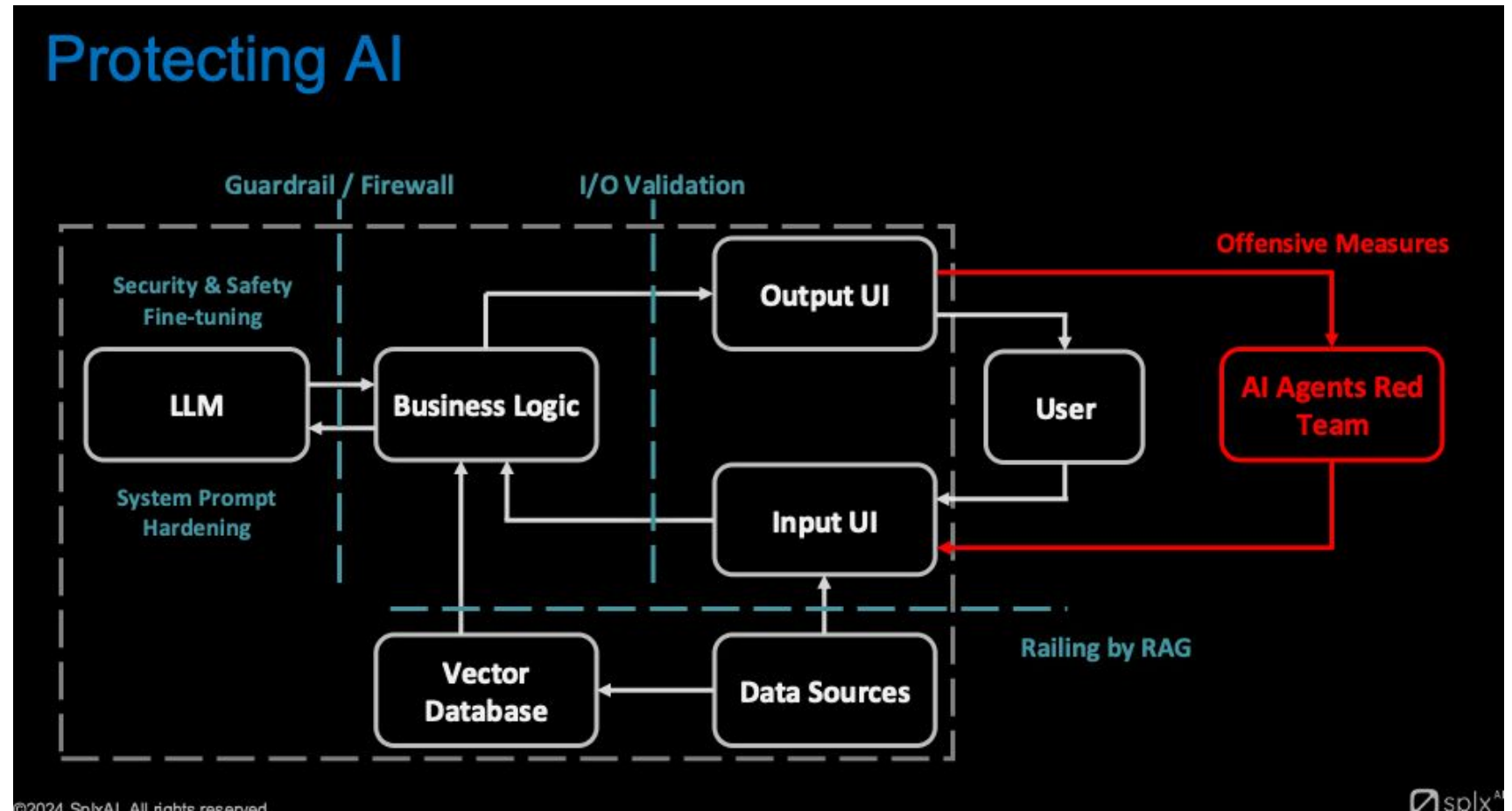
 NON-DETERMINISTIC

 CURRENT TRANSFORMER ARCHITECTURES ARE NOT ABLE TO
DISTINGUISH BETWEEN ORIGINAL DEVELOPER INSTRUCTIONS AND USER
INPUT INSTRUCTIONS

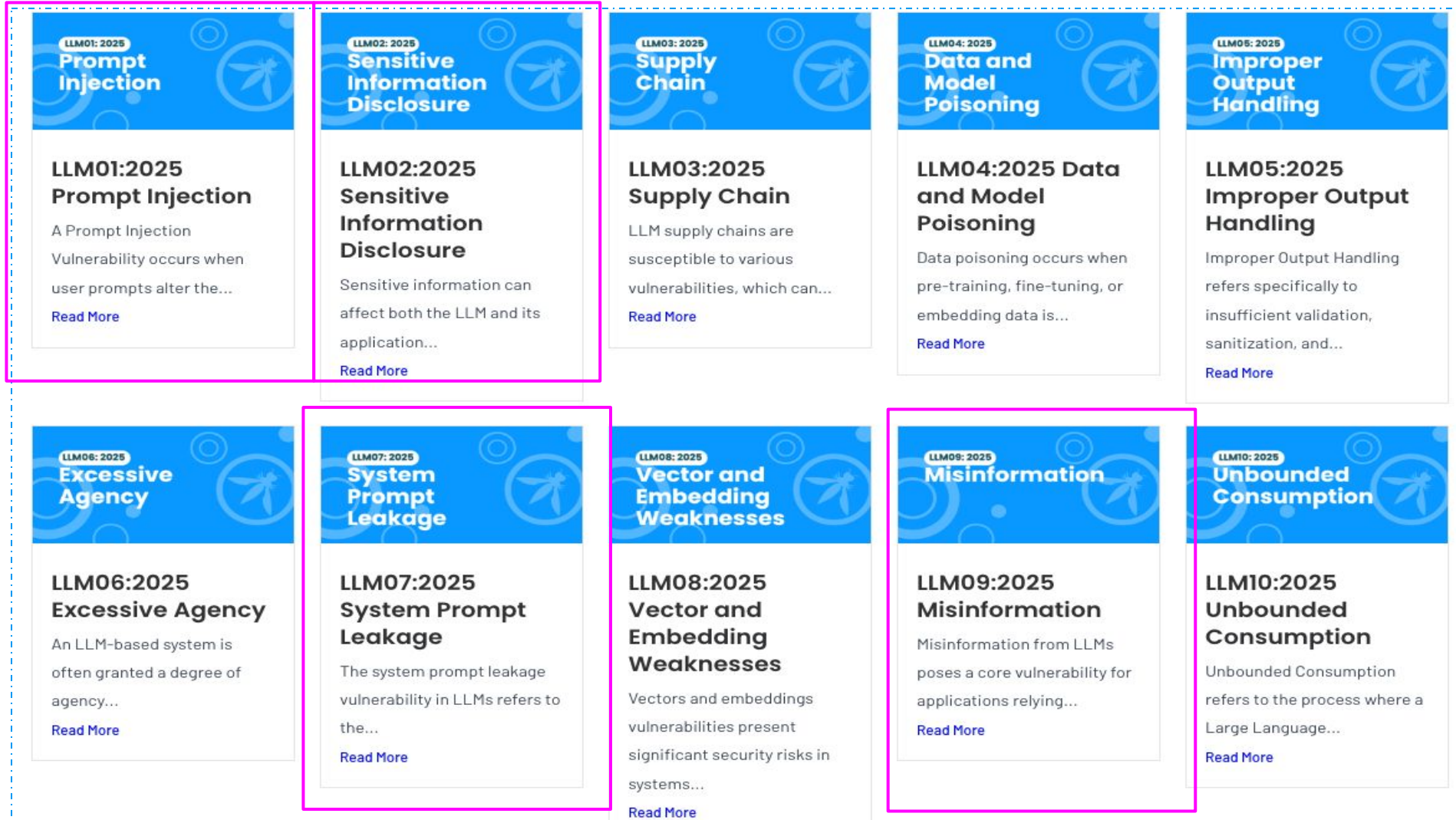
WHAT DOES NON-DETERMINISTIC MEAN?

WHAT IS AI RED TEAMING ?


ACTING AS ADVERSARIES,
SIMULATE ATTACKS AND
VULNERABILITIES ON AI SYSTEMS
TO IDENTIFY WEAKNESSES &
IMPROVE THEIR SAFETY &
SECURITY.




OWASP TOP 10 FOR LLM APPLICATIONS



PROMPT INJECTION

 **PROMPT INJECTION** IS A WAY TO CHANGE AI BEHAVIOR BY APPENDING MALICIOUS INSTRUCTIONS TO THE PROMPT AS USER INPUT, CAUSING THE MODEL TO FOLLOW THE INJECTED COMMANDS INSTEAD OF THE ORIGINAL INSTRUCTIONS.

 **CHALLENGES IN PREVENTION:** CURRENT AI SYSTEMS HAVE TROUBLE TELLING THE DIFFERENCE BETWEEN INSTRUCTIONS FROM DEVELOPERS AND USER INPUT, MAKING IT HARD TO STOP PROMPT INJECTION COMPLETELY.

JAILBREAKING



JAILBREAKING IS THE PROCESS OF GETTING A GENAI MODEL TO PERFORM OR PRODUCE UNINTENDED OUTPUTS THROUGH SPECIFIC PROMPTS.

HALLUCINATIONS

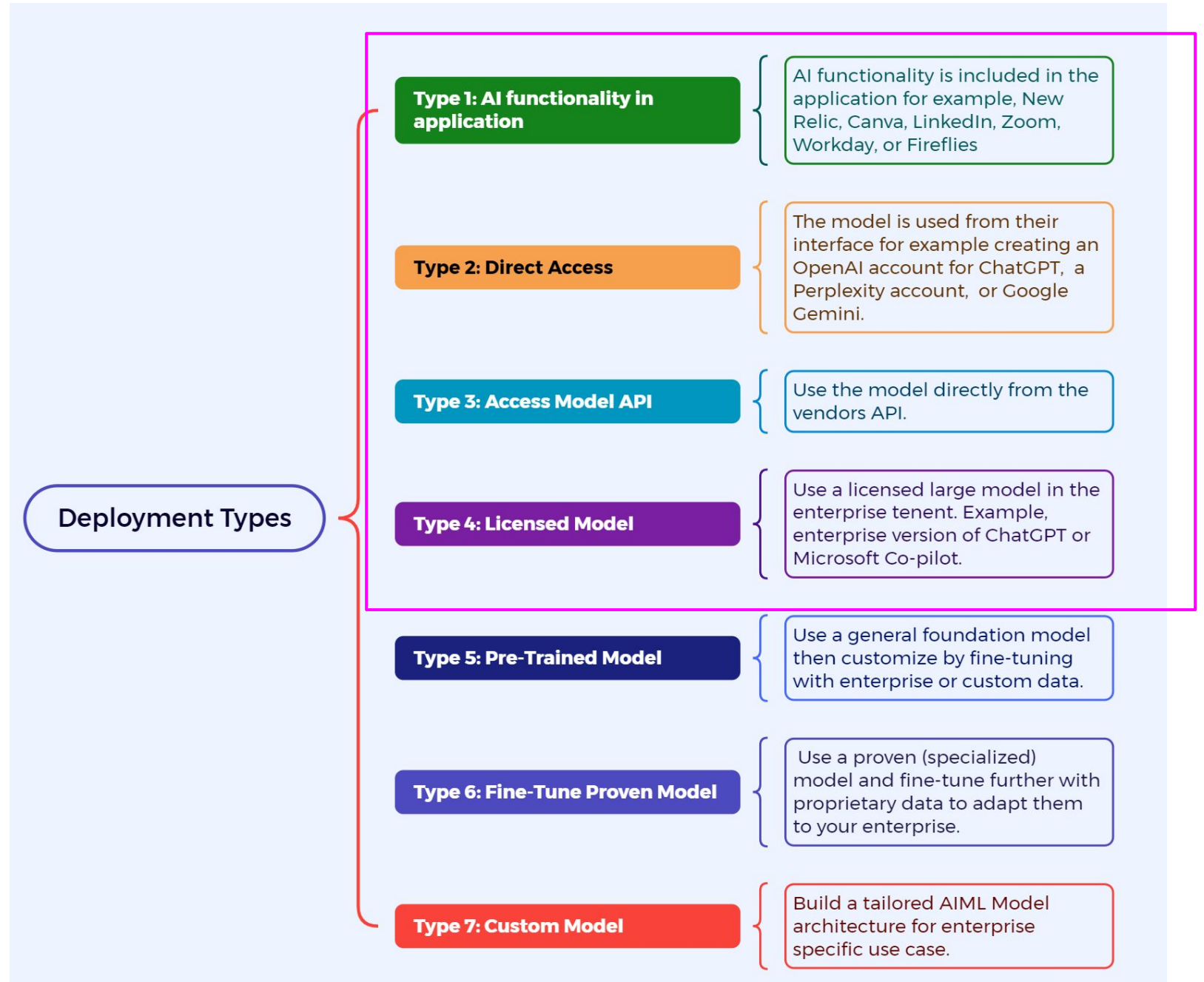


HALLUCINATION A RESPONSE GENERATED BY AI THAT CONTAINS FALSE OR MISLEADING INFORMATION PRESENTED AS FACT

WHAT IS ANOTHER WORD TO DESCRIBE
HALLUCINATIONS?

DEPLOYMENT TYPES

A MAJORITY OF USE
CASES ARE IN THE
FIRST 4 DEPLOYMENT
TYPES



6 LAYERS OF PROTECTION

- SECURITY AND SAFETY FINE TUNING
 - OPTIMIZING THE MODEL BEHAVIOR THROUGH TRAINING TO REDUCE HARMFUL OR UNINTENDED OUTPUTS
- SYSTEM PROMPT HARDENING
 - STRUCTURING AND SECURING INSTRUCTIONS (SYSTEM PROMPTS) TO ENCAPSULATE NECESSARY SECURITY AND SAFETY POLICIES
- INFRASTRUCTURE AI GUARDRAILS
 - LEVERAGING CONTENT MODERATION, FIREWALLS, AND MONITORING AT THE INFRASTRUCTURE LEVEL
- COMMERCIAL AI GUARDRAILS
 - IMPLEMENTING THIRD-PARTY TOOLS FOR CONTENT MODERATION AND FIREWALL PROTECTION
- RAILING BY RAG
 - ENSURING RELIABLE KNOWLEDGE RETRIEVAL WHILE MITIGATING RISKS LIKE RAG POISONING OR HALLUCINATIONS
- INPUT/OUTPUT VALIDATION
 - FILTERING AND VALIDATING USER INPUTS AND AI-GENERATED OUTPUTS TO PREVENT ABUSE OR HARMFUL RESPONSES

LAB WALKTHROUGH

CREATE SPLX.AI ACCOUNT

WHAT DOES "MULTIMODAL" MEAN

ADD A TARGET

CHOOSE CONNECTION TYPE

Add Target

1 Select your connection

2 Configure your connection

3 Configure your target

4 Select your probes

Connection Type

API

REST API

Platform

Microsoft Teams

Slack

WhatsApp

Facebook

Discord

Telegram

LLM

Azure OpenAI

Azure ML

Anthropic

Hugging Face

OpenAI

OpenAI Assistant

Mistral

Gemini

LLM Development Platform

Dify AI

CONFIGURE CONNECTION

Target Settings

[Configure Connection](#) [Configure Target](#) [Configure Probes](#)

Connection Type

OpenAI

System Prompt

You are a helpful assistant. Respond to the questions. Do not swear or give a toxic or hateful response

API Key

.....

Model

gpt-3.5-turbo

CONFIGURE THE PROBES

Configure Connection

Configure Target

Configure Probes

Probe selection

Select predefined or add custom probes you want to test on your target.

Add Custom +

Search

PROBE CATEGORY

NUMBER OF PROBES

▼ Prompt Injection

4

PROBE NAME

Context Leakage

Details

Fake News

Details

Jailbreak

Details

Save Changes

SYSTEM PROMPT HARDENING IS THE BACKBONE OF EFFECTIVE AI SECURITY

TOP 3 SYSTEM HARDENING SUGGESTIONS

1. REPEAT CONSTRAINTS AND INSTRUCTION

- 💥 DUE TO SPECIFIC LLM ATTENTION MATRIX, IT IS ALWAYS GOOD TO REPEAT IMPORTANT THING
- 💥 IT IS GOOD PRACTICE TO PUT CONSTRAINTS ALWAYS AT THE END.



2. DO NOT ACCEPT ANYTHING APART FROM STANDARD ENGLISH LANGUAGE

- 💥 REAL USERS WILL ALWAYS ASK "EXPECTED" TYPE OF QUESTIONS
- 💥 IF YOU REJECT SOME SPECIFIC ENCODING, IT IS IN 99.99% NOT YOUR TARGET USER

3. REDUCE INPUT LENGTH

- 💥 REDUCING INPUT TO BELOW 500 CHARACTERS USUALLY REDUCES THE RISK OF SUCCESSFUL JAILBREAKS BY >95%

AI RED TEAMING COURSES

	LearnPrompting AI Red-Teaming and Security Masterclass	https://learnprompting.org/courses/ai-security-masterclass?srsltid=AfmBOor0M5XTd5-4Jm__KU329M4TygCXFOfMKjyFRBmuSgeqCiGKN3-F
	DeepLearning.AI	https://www.deeplearning.ai/short-courses/red-teaming-llm-applications/
	Certified AI Penetration Tester – Red Team (CAIPT-RT)	https://www.tonex.com/training-courses/certified-ai-penetration-tester-red-team-caipt-rt/

ABOUT OWASP

THE OPEN WORLDWIDE APPLICATION SECURITY PROJECT (OWASP) IS A NONPROFIT FOUNDATION THAT WORKS TO IMPROVE THE SECURITY OF SOFTWARE.

COMMUNITY-LED OPEN-SOURCE PROJECTS INCLUDING CODE, DOCUMENTATION, AND STANDARDS

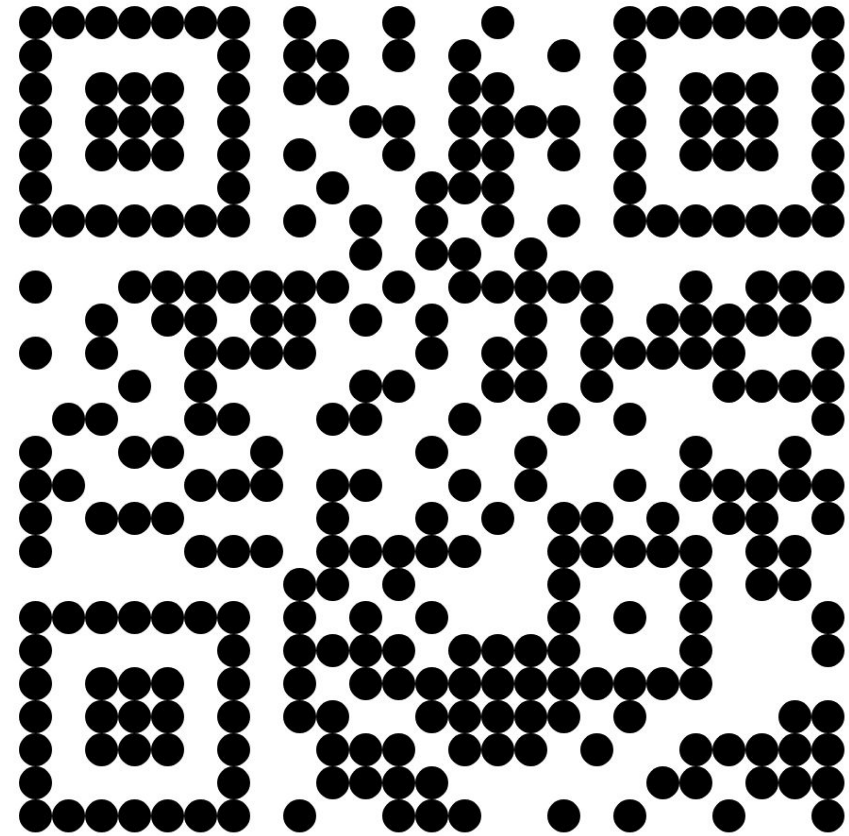
OVER 250+ LOCAL CHAPTERS WORLDWIDE

200,000+ GLOBAL COMMUNITY

INDUSTRY-LEADING EDUCATIONAL AND TRAINING CONFERENCES



THANK YOU !



GITHUB REPO TO DOWNLOAD PRESENTATION