

GenAi en AWS

Mitre ATLAS:

Seguridad en la

Era de IA





Nosotros Somos



Kayck Silva

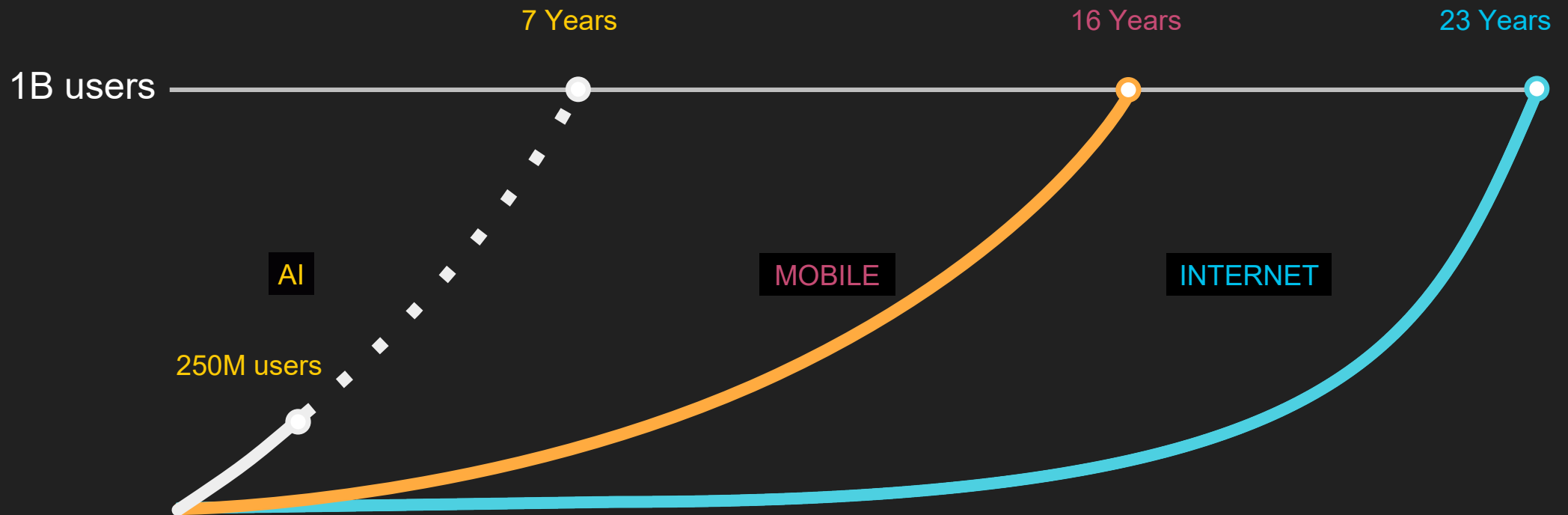
Security Delivery
Associate Manager
Cloud&IA Security Lead
at Accenture



Max Calderon

Security Delivery
Associate Manager
Cloud&IA Security Lead
at Accenture

IA ya es la tecnologia de mas rapida adopcion de la historia



Source: Internet World Stats; assumed internet inception in 1982 after US DoD standardized TCP/IP protocol

AI hiring tools may be filtering out the best job applicants

16 February 2024

5 per month

Share

Opinion

Sport

Culture

Lifestyle



The Guardian

[Americas](#) [Asia](#) [Australia](#) [Middle East](#) [Africa](#) [Inequality](#) [Global development](#)

• This article is more than 1 year old

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare



Energy & Environment

AI could strain Texas power grid this summer

KUT 90.5 | By [Mose Buchele](#)

Published June 21, 2024 at 2:20 PM CDT



La Intersección: IA Generativa & Seguridad



Herramienta

- Alertas
- Resumen de datos e Informes
- Automatización de Consultas - Código
- Automatización de Pruebas - Pentesting



Objetivo

- Vector de entrada;
- Fuga de datos;
- Interrupción del servicio;
- Envenenamiento de modelos de IA;
- Vulnerabilidades en la cadena de suministro;



Amenaza

- Deep Fake;
- Automatización de ataques;
- Códigos maliciosos;
- Personalización de Campañas de Phishing;
- Ej: Malware PoC BlackMamba.

Adversarial Threat Landscape for Artificial-Intelligence Systems

ATLAS Matrix

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 adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the [ATLAS Navigator](#).

Reconnaissance ^{&}	Resource Development ^{&}	Initial Access ^{&}	AI Model Access	Execution ^{&}	Persistence ^{&}	Privilege Escalation ^{&}	Defense Evasion ^{&}	Credential Access ^{&}	Discovery ^{&}	Collection ^{&}	AI Attack Staging	Command and Control ^{&}	Exfiltration ^{&}	Impact ^{&}
6 techniques	12 techniques	6 techniques	4 techniques	4 techniques	4 techniques	2 techniques	8 techniques	1 technique	7 techniques	3 techniques	4 techniques	1 technique	5 techniques	7 techniques
Search Open Technical Databases^{&}	Acquire Public AI Artifacts	AI Supply Chain Compromise	AI Model Inference API Access	User Execution^{&}	Poison Training Data	LLM Plugin Compromise	Evade AI Model	Unsecured Credentials^{&}	Discover AI Model Ontology	AI Artifact Collection	Create Proxy AI Model	Reverse Shell	Exfiltration via AI Inference API	Evade AI Model
Search Open AI Vulnerability Analysis	Obtain Capabilities^{&}	Valid Accounts^{&}	AI-Enabled Product or Service	Command and Scripting Interpreter^{&}	Manipulate AI Model	LLM Jailbreak	LLM Jailbreak		Discover AI Model Family	Data from Information Repositories^{&}	Manipulate AI Model		Exfiltration via Cyber Means	Denial of AI Service
Search Victim-Owned Websites^{&}	Develop Capabilities^{&}	Evade AI Model	Physical Environment Access	LLM Prompt Injection	LLM Prompt Self-Replication		LLM Trusted Output Components Manipulation						Extract LLM System Prompt	Spamming AI System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application^{&}	Full AI Model Access	LLM Plugin Compromise	RAG Poisoning		LLM Prompt Obfuscation						AI Data Package	Erode AI Model Integrity
Active Scanning^{&}	Publish Poisoned Datasets	Phishing^{&}					False RAG Entry Injection						AI Response Rendering	Cost Harvesting
Gather RAG-Indexed Targets	Poison Training Data	Drive-by Compromise^{&}					Impersonation^{&}							External Harms
	Establish Accounts^{&}						Masquerading^{&}							Erode Dataset Integrity
	Publish Poisoned Models						Corrupt AI Model							
	Publish Hallucinated Entities													
	LLM Prompt Crafting													
	Retrieval Content Crafting													
	Stage Capabilities^{&}													



15
TACTICS

74
TECHNIQUES

20
MITIGATIONS

31
CASE STUDIES

Home > Studies > ChatGPT Plugin Privacy Leak

ChatGPT Plugin Privacy Leak

Exercise

Incident Date: N
Actor: Embrace

Summary

Researchers un
plugins to take c
leakage from the

Heads up: Hallucinated packages in the wild?

During our research encountered an interesting python hallucinated package called "huggingface-cli".

DOWNLOAD DATA ▾

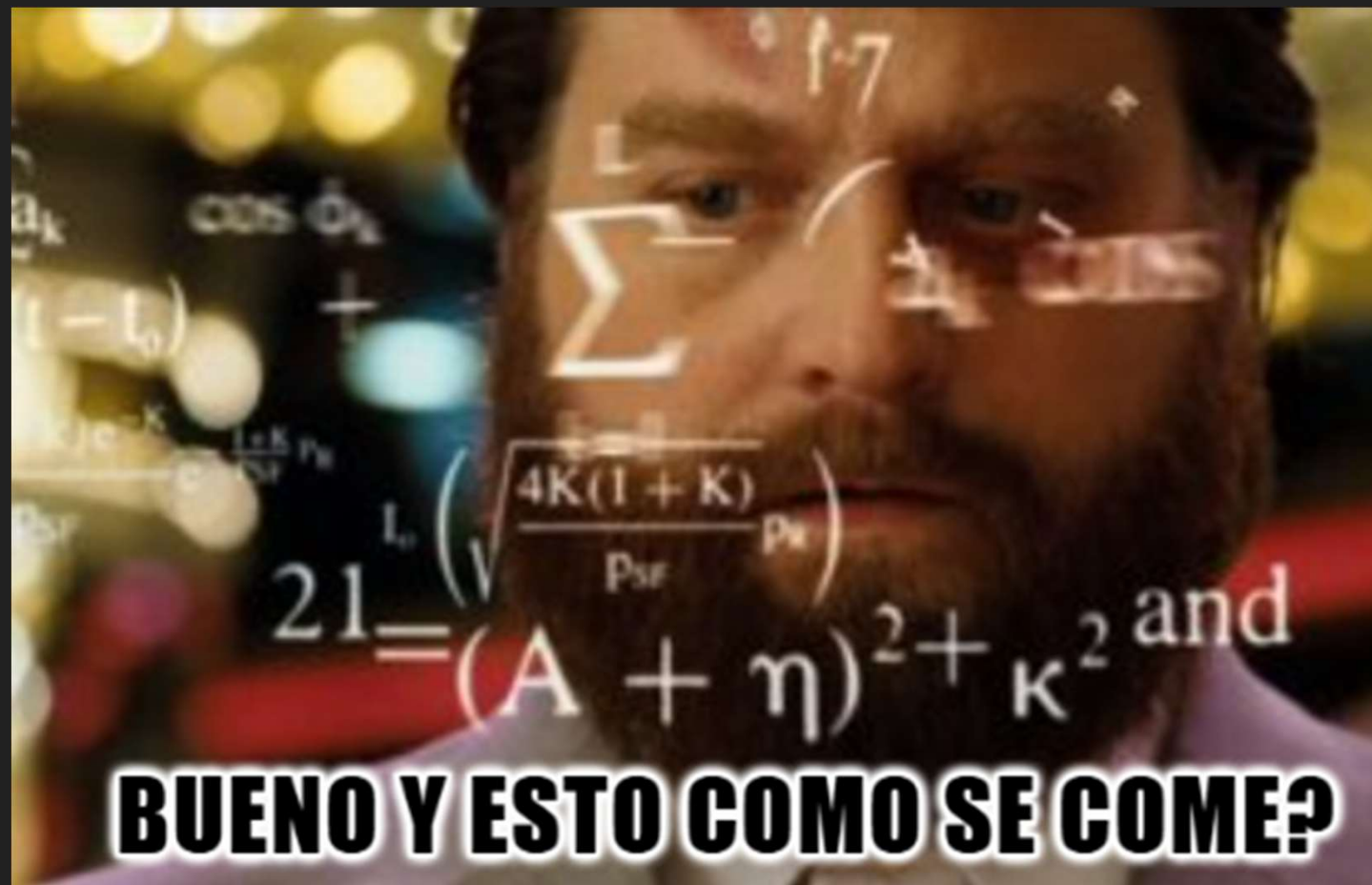
- You**
how to upload a model to huggingface?
- ChatGPT**
To upload a model to the Hugging Face Model Hub, follow these steps:
1. Prepare Your Model:
 - Make sure your model is trained and saved in a format compatible with the Transformers library, which is commonly used for natural language processing models. Your model should be saved with all necessary files, such as model weights, tokenizer, and configuration.
 2. Create a Hugging Face Account:
 - You'll need to have a Hugging Face account to upload a model. If you don't have one, sign up at <https://huggingface.co/signup>.
 3. Install 'transformers' Library:
 - Make sure you have the 'transformers' library installed. You can install it using pip:

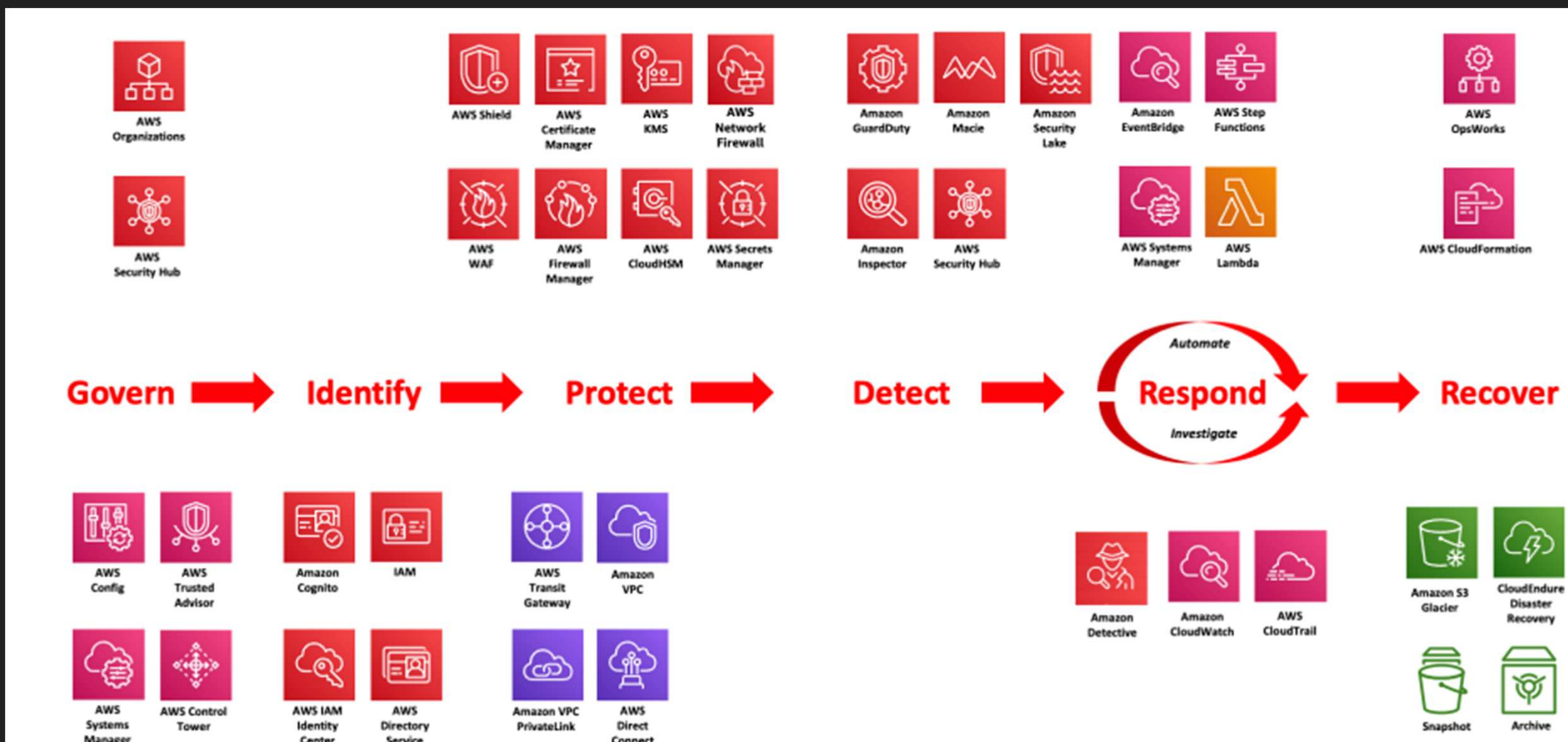
```
bash
```

Copy code

ed malicious websites through ChatGPT
back, users may be vulnerable to PII



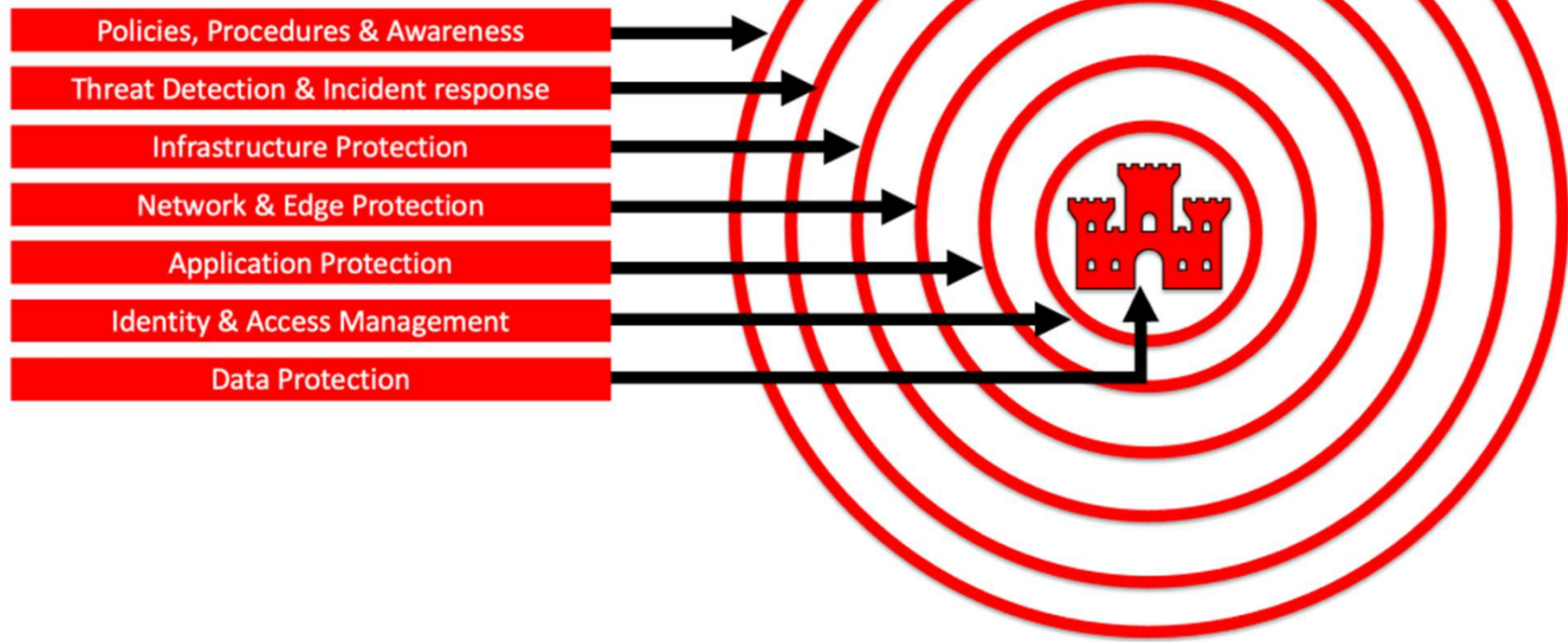




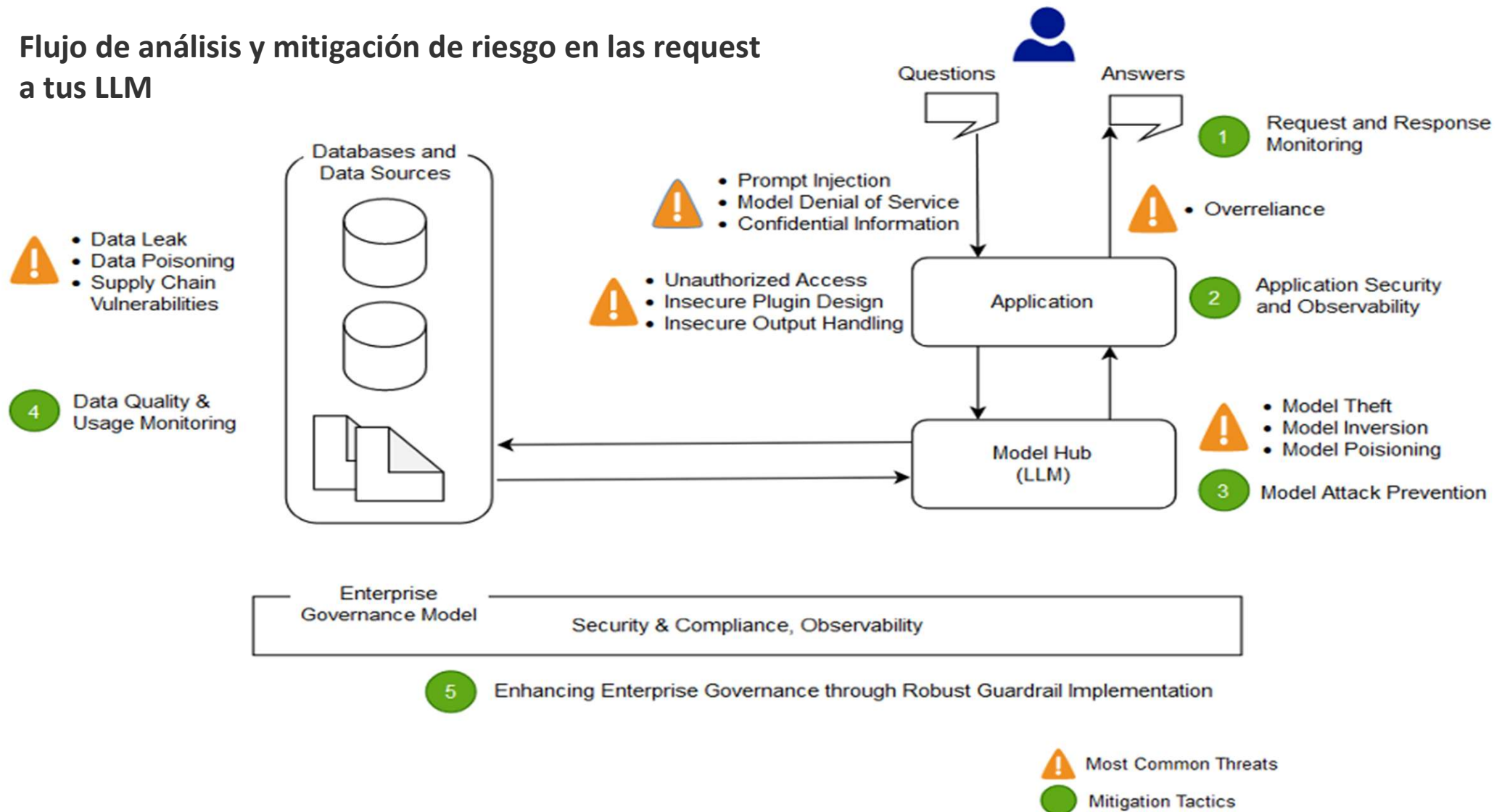
<https://aws.amazon.com/es/blogs/machine-learning/architect-defense-in-depth-security-for-generative-ai-applications-using-the-owasp-top-10-for-llms/>

Defense-in-depth

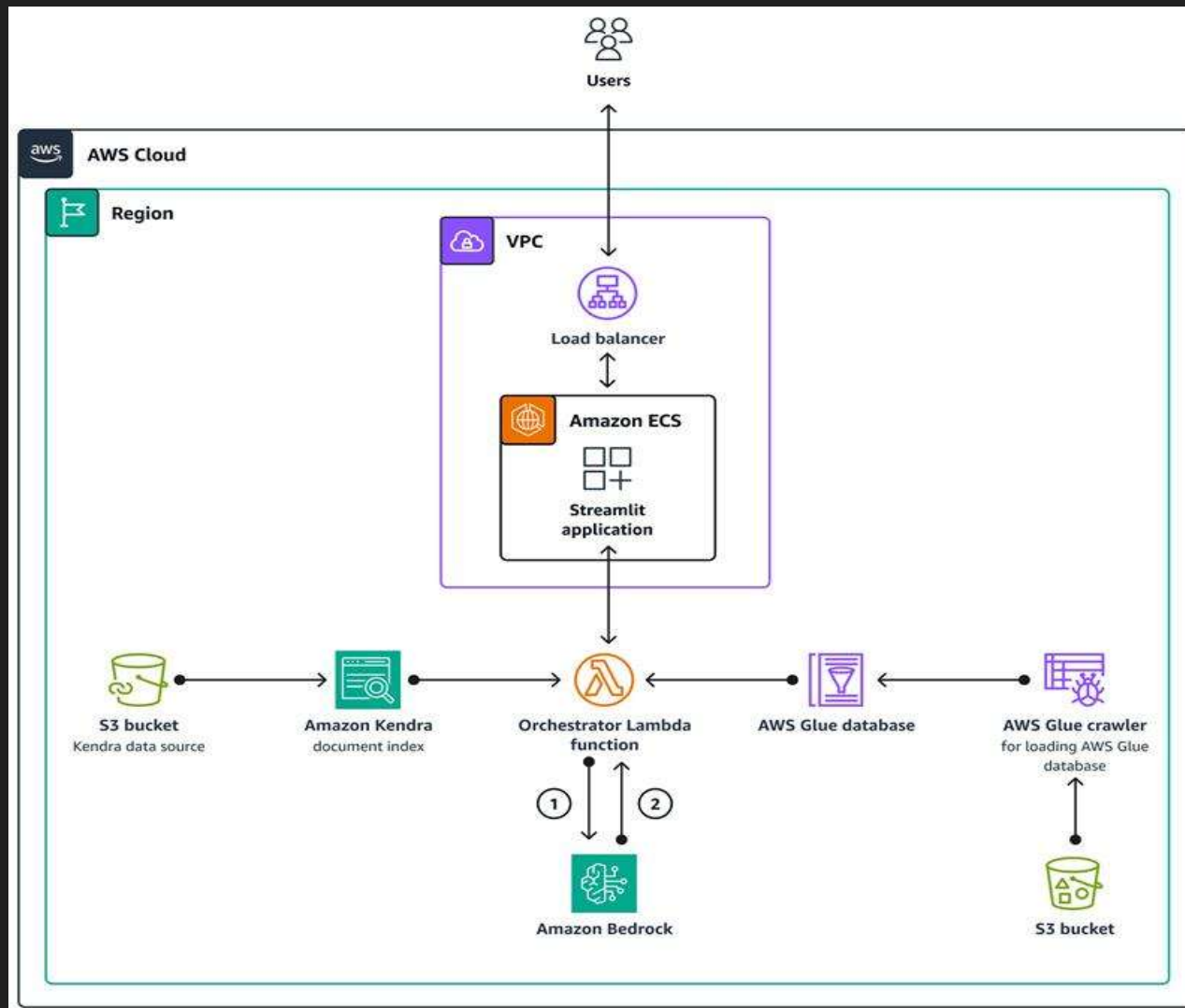
PROTECTION WITH MULTIPLE LAYERS OF SECURITY CONTROLS



Flujo de análisis y mitigación de riesgo en las request a tus LLM



Arquitectura Alto nivel para-MVP en Bedrock





Un gran **poder** conlleva una gran **Responsabilidad**



Kayck Silva

Security Delivery
Associate Manager

Cloud Security



Max Calderon

Security Delivery
Associate Manager

Cloud Security



¡GRACIAS!