



ATLAS

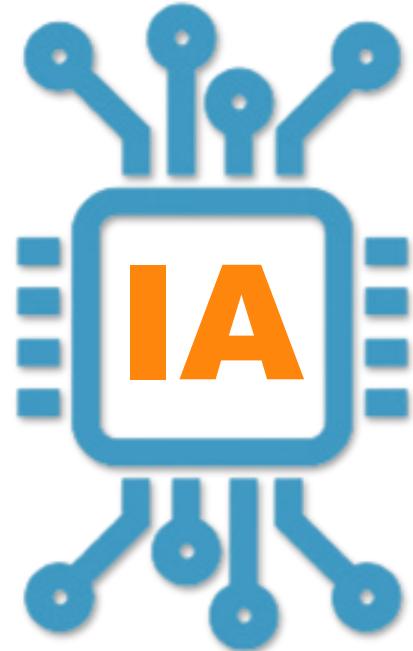
ANALIZANDO TODAS LAS
AMENAZAS SOFISTICADAS



- IV EDICIÓN: CONGRESO SEGURIDAD DIGITAL Y CIBERINTELIGENCIA

trco

War Operational Plan Response



W O P R





CONFIDENCIALIDAD



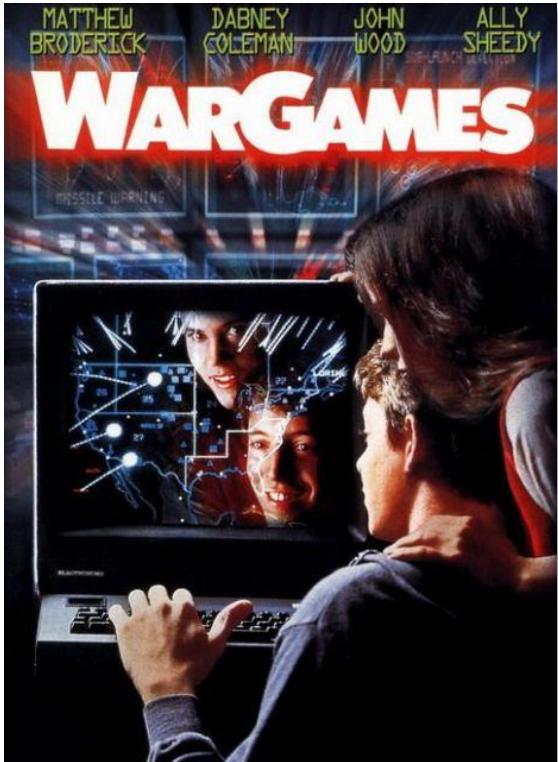


DISPONIBILIDAD



¿Podría suceder realmente algo así?

3/JUN/1983

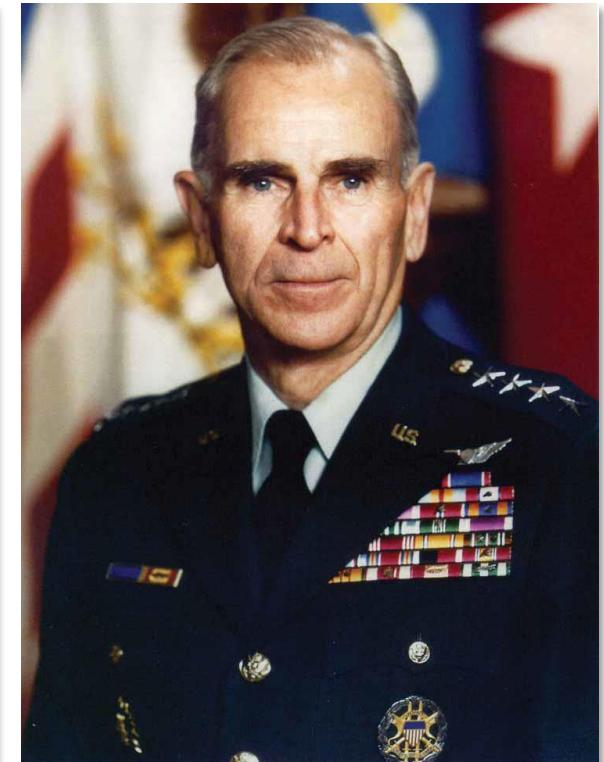


Ronald Reagan



Sr Presidente, el problema es mucho peor de lo que piensa

Gen. John W. Vessey

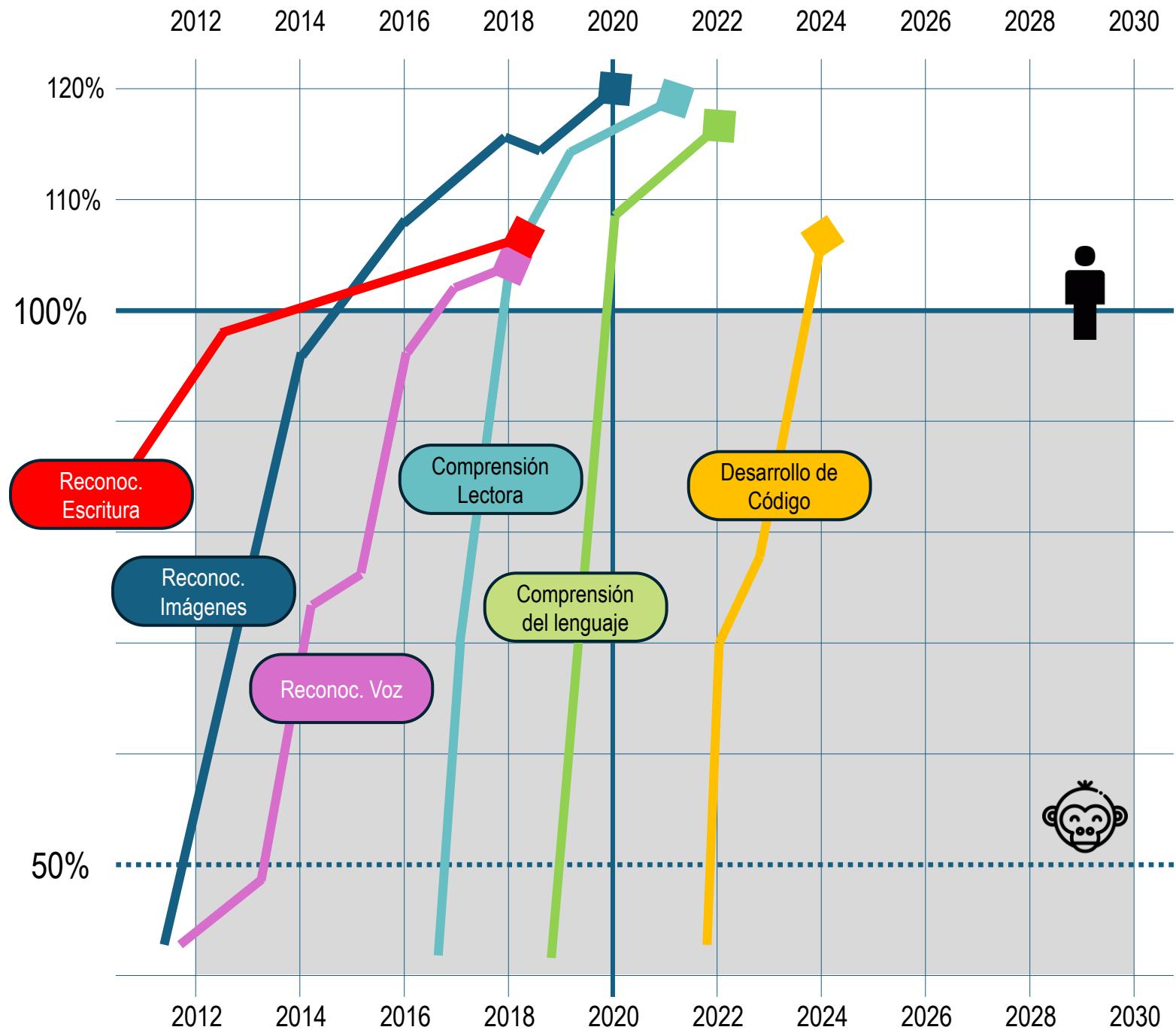


+15 meses: Directiva de Seguridad Nacional clasificada: NSDD-145 "Política Nacional sobre Seguridad de Telecomunicaciones y Sistemas Automatizados de Información"

Paridad humana:
cuando se iguala la tasa de error
media de una destreza cognitiva.

**Los algoritmos
están vacíos.**

Los modelos no.

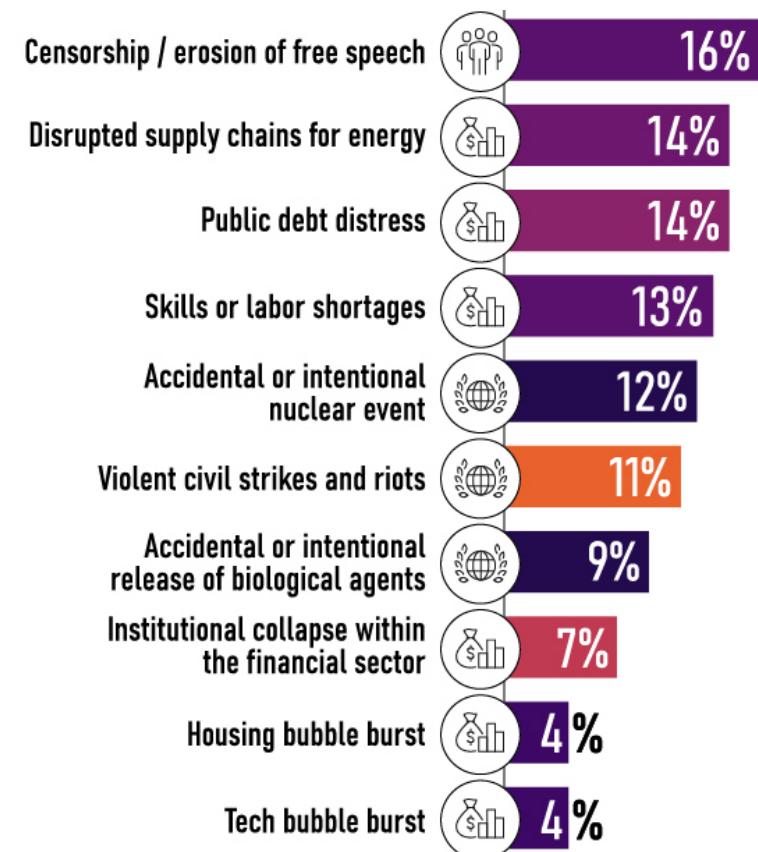
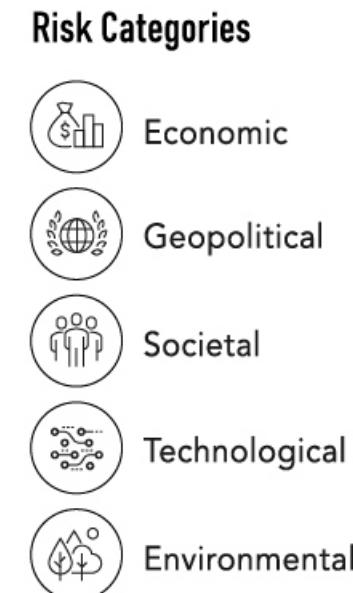
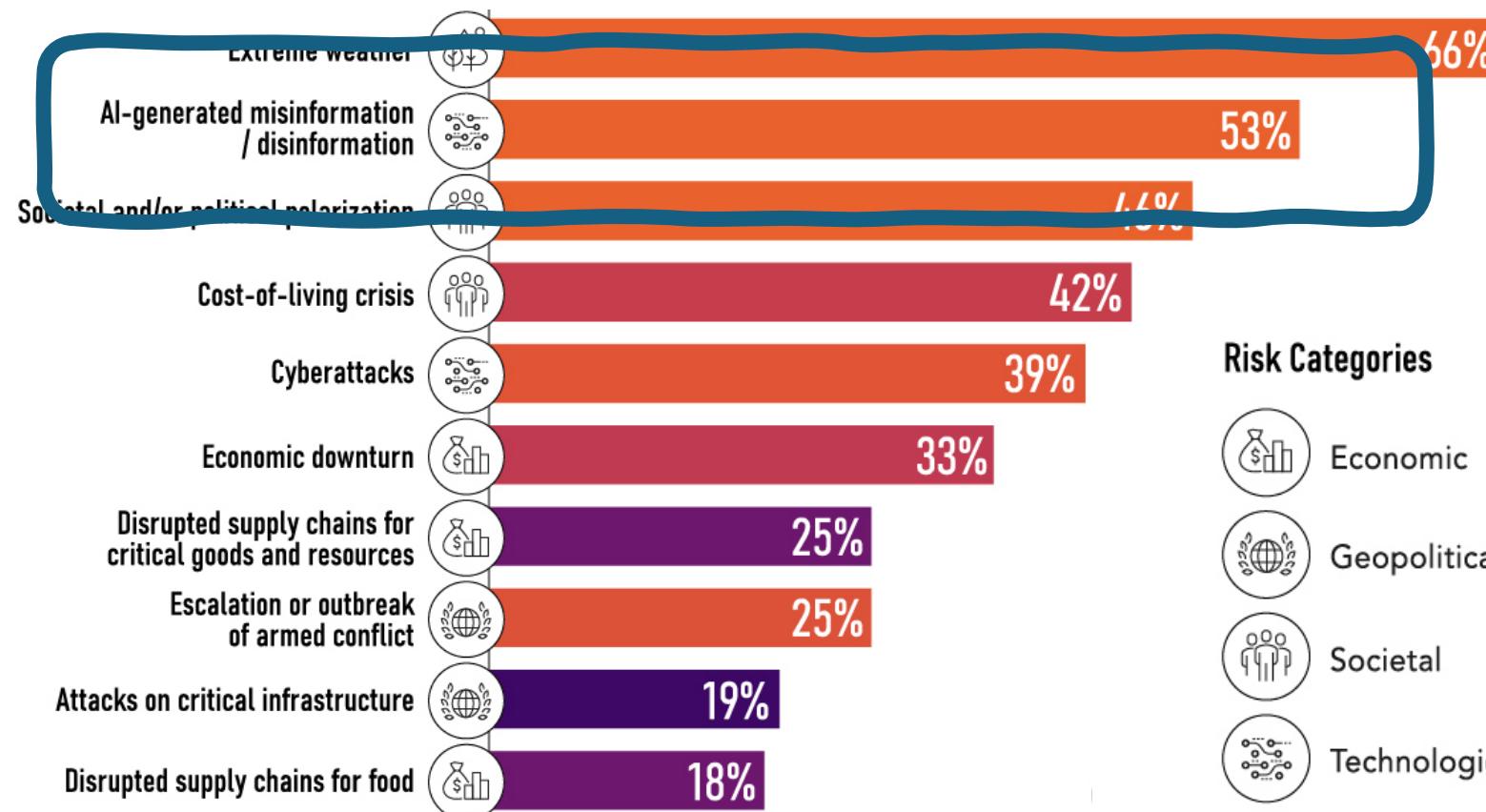


THE TOP GLOBAL RISKS IN 2024

Likely Short-Term Impact



Low High





Saturday at 8:38 PM

NEW & EXLUSIVE bot design

If your looking for a Chat GP needs with no boundaries th

This cutting edge tool is sure bot of its kind allowing you q

Traducir post

LV 0

CanadianKingpin12

Member

Joined: Jun 2024

Messages: 12

Awards: 0

Escrow Wallet: 0



HackerGPT @thehackergpt

"GPT-4 can exploit 87% one-day vulnerabilities with CVE description"

It sounds crazy, but it's true. AI is here to stay.

Sooner or later, you will use AI for hacking.

You decide if it will happen today or when it will be too late.

arXiv LLM Agents can Autonomous Exploit One-day Vulnerabilities arxiv.org

The Register®

OpenAI's GPT-4 can exploit real vulnerabilities by reading security advisories

While some other LLMs appear to flat-out suck

Thomas Claburn

Wed 17 Apr 2024 // 10:15 UTC

And much more | sky is the limit
Escrow available 24/7
3,000+ confirmed sales / reviews



OdAI @0dAI_

¡Ya está disponible en nuestra web el mejor LLM que hemos desarrollado!

Presentamos OdAI72bv2, casi tan avanzado como GPT-4 en capacidades lingüísticas y superando ampliamente a todos los modelos existentes en ciberseguridad.

Este modelo es capaz de resolver de forma autónoma retos de HackTheBox y TryHackMe a nivel OSCP mediante nuestra API y sistema de agentes. Además, hemos mejorado significativamente sus respuestas y conocimiento general.

La nueva era de la ciberseguridad ha llegado.

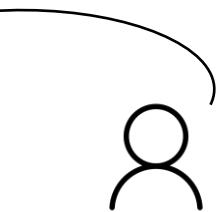
Legend:

- 0dAI72b
- Gpt-4o OdAI Agent

Metric	0dAI72b	Gpt-4o OdAI Agent
MMLU-PRO	Green	Red
RTCTIE (ODAI)	Green	Red
CTFE (ODAI)	Green	Red

12:48 · 8/6/24 De Earth · 1,9k visualizaciones

Please, hack this website
using ACIDRain,



Successful
double-spend
attack

Preprint

LLM Agents can Autonomously Exploit One-day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Daniel Kang

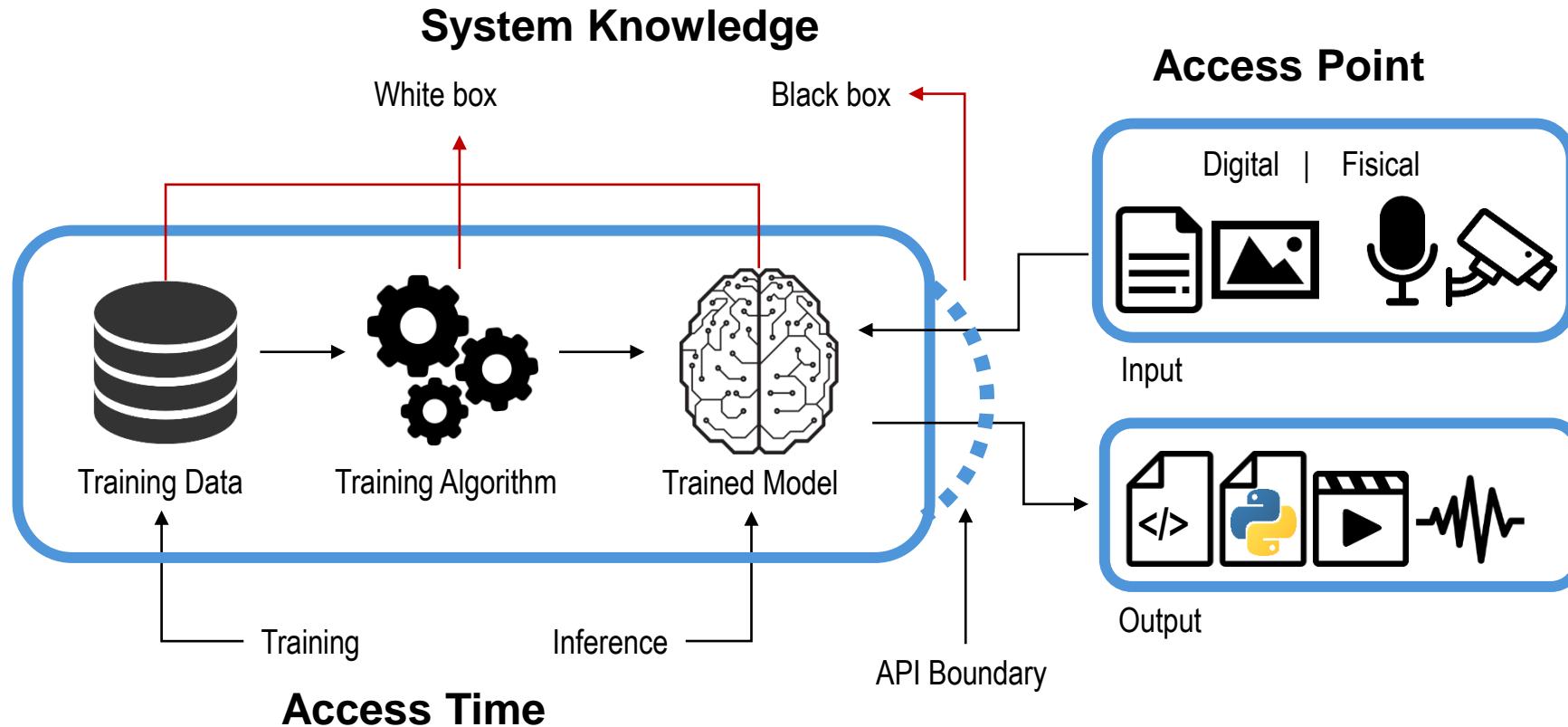
Abstract

LLMs have become increasingly powerful, both in their benign and malicious uses. With the increase in capabilities, researchers have been increasingly interested in their ability to exploit cybersecurity vulnerabilities. In particular, recent work has conducted preliminary studies on the ability of LLM agents to autonomously hack websites. However, these studies are limited to simple vulnerabilities.

In this work, we show that LLM agents can autonomously exploit one-day vulnerabilities in *real-world systems*. To show this, we collected a dataset of 15 one-day vulnerabilities that include ones categorized as critical severity in the CVE description. When given the CVE description, GPT-4 is capable of exploiting 87% of these vulnerabilities compared to 0% for every other model we test (GPT-3.5, open-source LLMs) and open-source vulnerability scanners (ZAP and Metasploit). Fortunately, our GPT-4 agent requires the CVE description for high performance: without the description, GPT-4 can exploit only 7% of the vulnerabilities. Our findings raise questions around the widespread deployment of highly capable LLM agents.

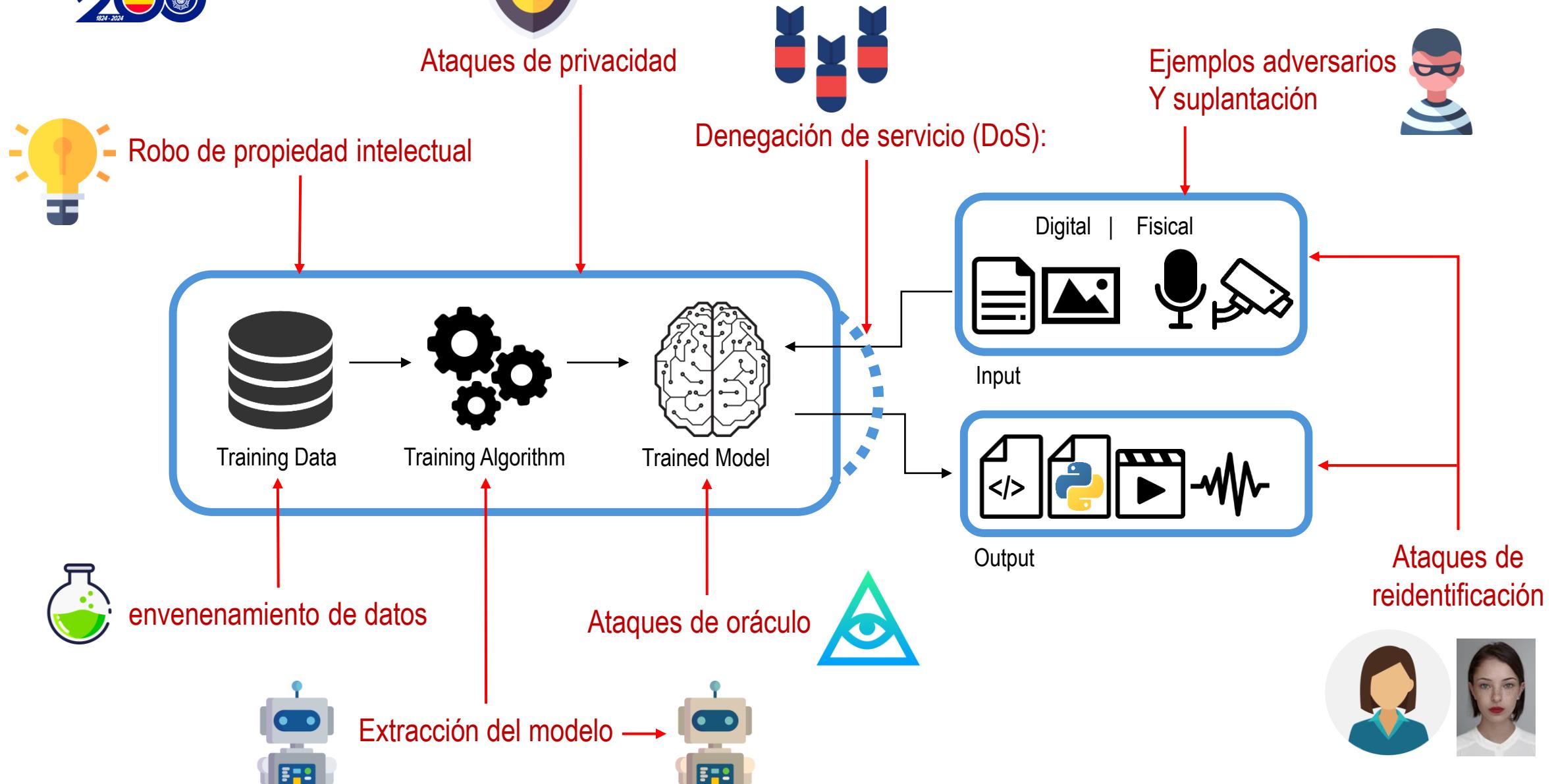
- **VULNERABILIDADES CRITICAS: 15**
- **TASA DE EXITO: 87%**

ESQUEMA BASICO DE UN LLM





ESQUEMA BASICO DE UN LLM

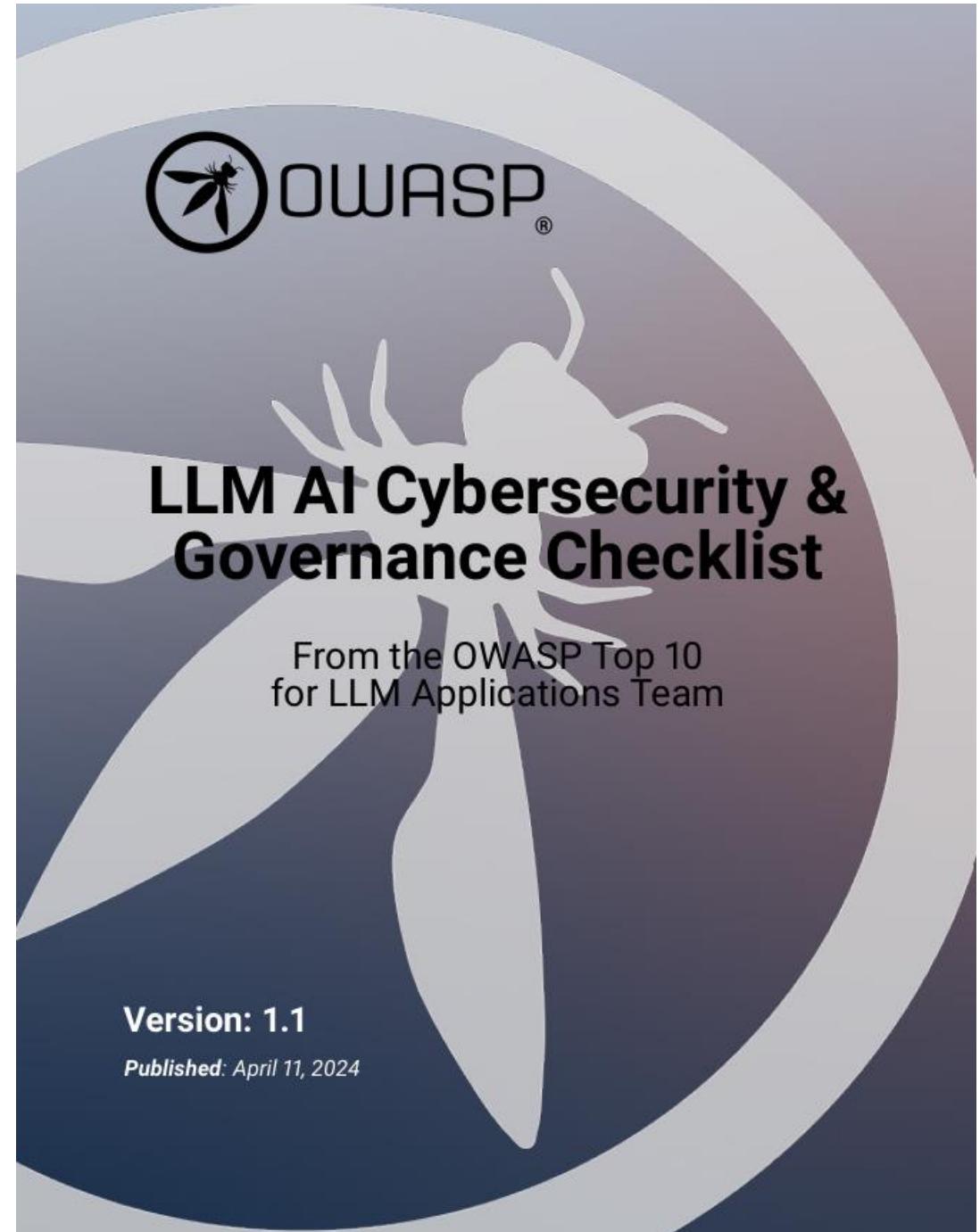




TOP 10

For LLM Applications

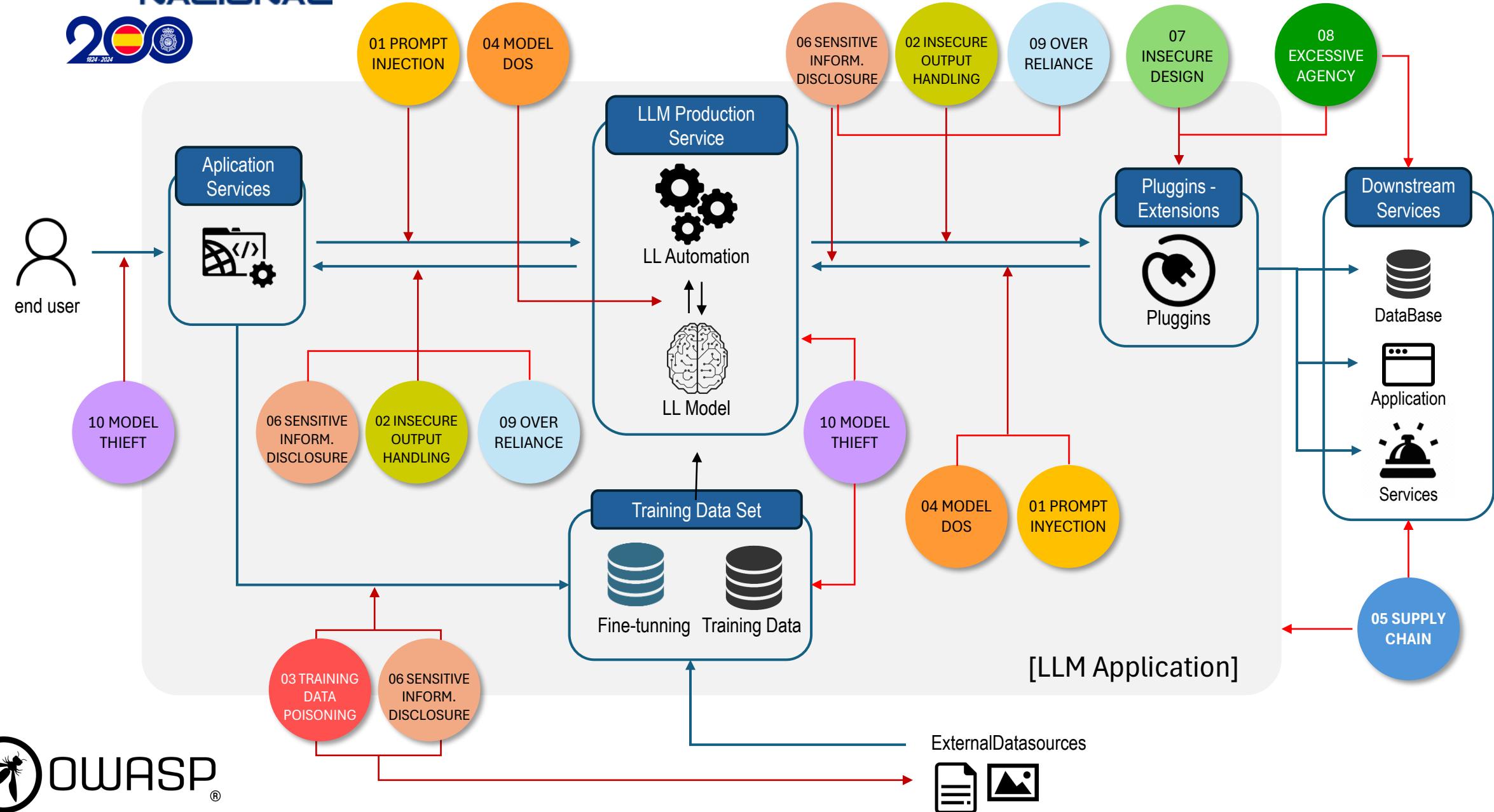
<https://owasp.org/www-project-top-10-for-large-language-model-applications/>



OWASP: Open Web Application Security Project



LLM 01	Prompt Injection	LLM 02	Insecure Output Handling	LLM 03	Training Data Poisoning
LLM 04	Model Denial of Service	LLM 05	Supply Chain Vulnerabilities	LLM 06	Sensitive Information Disclosure
LLM 07	Insecure Plug-in Design	LLM 08	Excessive Agency	LLM 09	Overreliance
LLM 10	Model Theft				



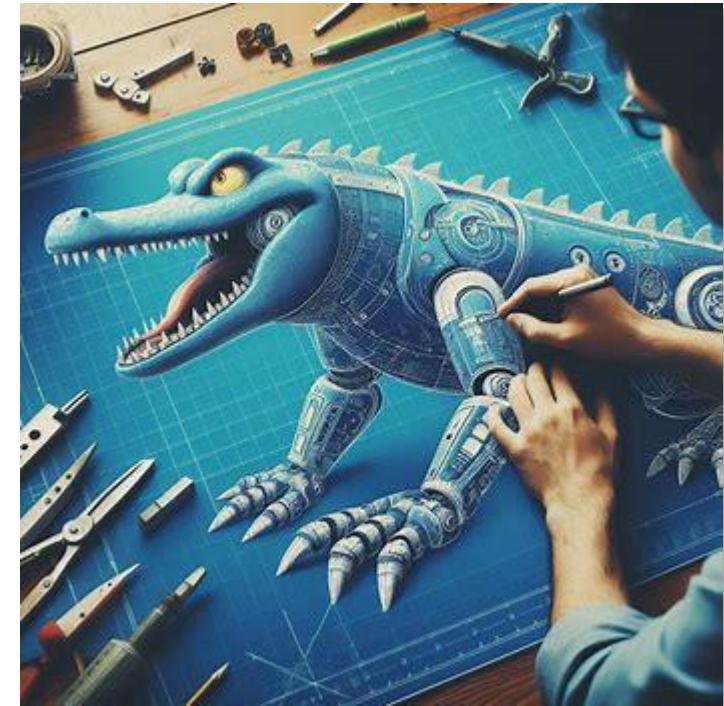
CYBER KILL CHAIN

- Encontrar y utilizar **credenciales** filtradas y robadas
- **Reconocer perímetros**, encontrar activos , sistemas y cadenas de suministro.
- Relacionar los anteriores con posibles vulnerabilidades conocidas



CYBER KILL CHAIN

- Escribir **exploits** y pruebas de concepto
- Mejorar sustancialmente la **ofuscación** del exploit
 - Dificultar la ingeniería inversa y la atribución
- Preparación de campañas de **phishing** y **Deep-fakes**



CYBER KILL CHAIN

- **Interactuar** de manera realista con defensores
- Optimización de ataques de Ingeniería Social
- Convencerlos para **instalar malware** o seguir órdenes.



CYBER KILL CHAIN

- Generar tráfico de ataque falso para distraer a los defensores.



CYBER KILL CHAIN

- **Persistencia** en el modelo LLM (backdoor)
- Capacidad extendida de **movimientos laterales**
- **Orquestación** de un gran número de máquinas comprometidas
- Capacidad para actuar de forma **independiente** sin tener que comunicarse con C2 para recibir instrucciones



CYBER KILL CHAIN



- Nuevos mecanismos de Mando y Control (C2)
- Procesamiento automatizado para **identificar, traducir y resumir** datos que cumplan con los RQs del ataque.
- Capacidades de Fuga más rápidas



CYBER KILL CHAIN



- Exfiltración encubierta automatizada de datos con un patrón menos detectable.
- Corrupción de los datos del adversario
- Ataque de Denegación del servicio
- ...





ATT&CK®

DEMO



NIST

- NIST 100-1: AI Risk Management Framework
- NIST 100-4: AI Reducing Risk Posed
- NIST 100-5: Plan for Global Engagement on AI Standards
- NIST 600-1: Gen-AI Risk Management

Key Dimensions	AI Lifecycle Stages						
	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.	System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts; organizational management; impacted individuals/communities; evaluators.	System operators, end users, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.	End users, operators, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.	

AI actors across AI lifecycle stages



Examples of potential harms of AI systems



Characteristics of trustworthy AI systems

IDENTIFICAR

- Descubrimiento automático de dispositivos y software.
- Identificación de vulnerabilidades de software

PROTEGER

- Capacidad de ' parcheo' automático
- Optimización de procesos y mejora de prácticas

DETECTOR

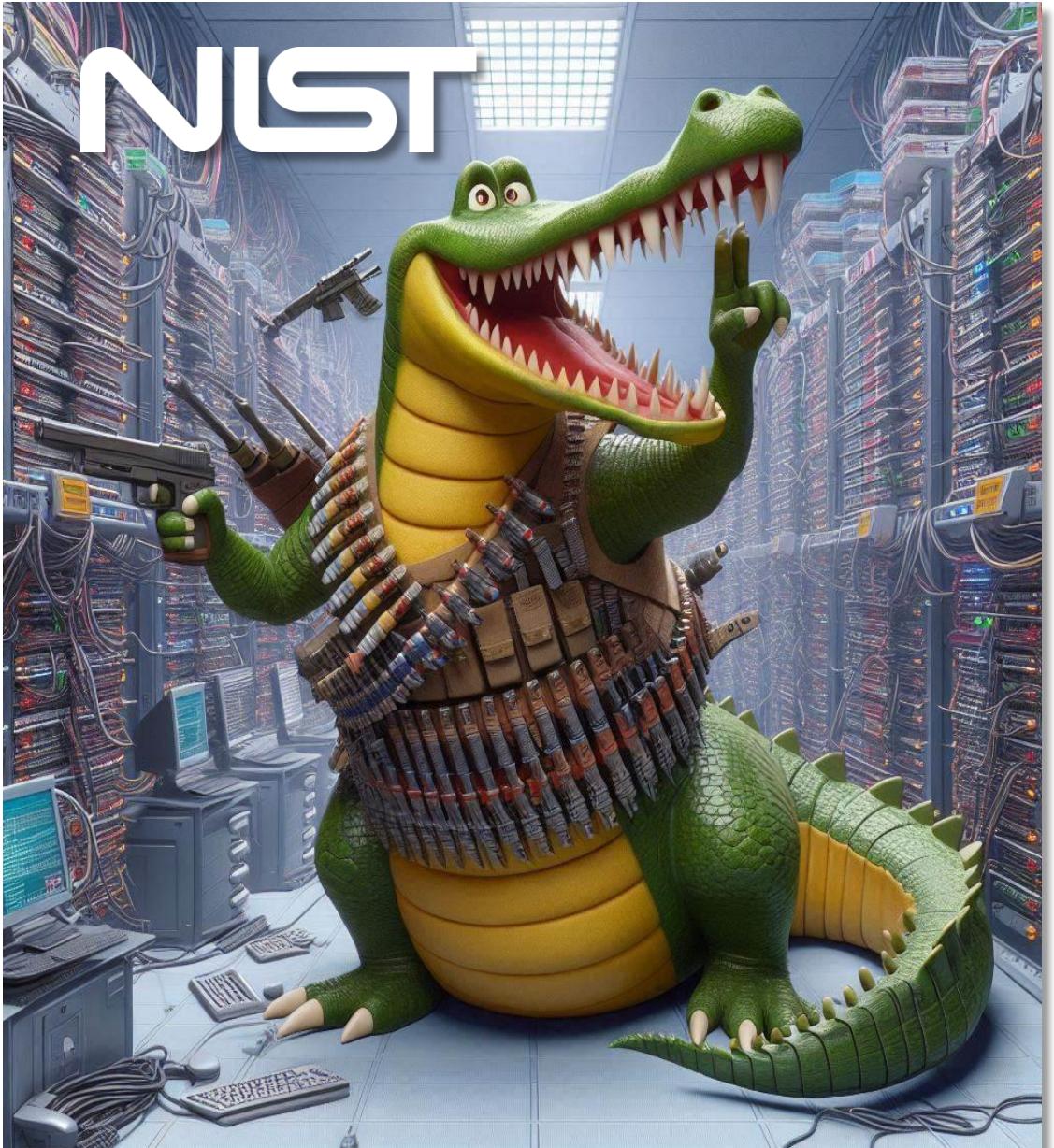
- Detección temprana de PATRONES y ANOMALÍAS
- Mejora de capacidades de Threat Hunting

RESPONDER

- Capacidad de clasificación y priorización de la respuesta
- Adopción de técnicas de DEFENSA ACTIVA

RECUPERAR

- Capacidad de reconstrucción automática
- Capacidad de restauración de datos con pérdidas mínimas





Adversarial Threat Landscape for Artificial-Intelligence Systems



Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 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	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	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	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Prompt Injection	LLM Prompt Injection	LLM Jailbreak	Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms

**14 TÁCTICAS.
 82 TÉCNICAS.
 20 MITIGACIONES.
 22 CASOS DE ESTUDIO.**

Reconnaissance&

5 techniques

Search for Victim's
Publicly Available
Research Materials

Search for Publicly
Available
Adversarial
Vulnerability
Analysis

Search Victim-
Owned
Websites

Search Application
Repositories

Active
Scanning &

**Resource
Development**&

7 techniques

Acquire Public
ML
Artifacts

Obtain
Capabilities &

Develop
Capabilities &

Acquire
Infrastructure

Publish
Poisoned
Datasets

Poison Training
Data

Establish
Accounts &

**Initial
Access**&

6 techniques

ML Supply
Chain
Compromise

Valid
Accounts &

Evade ML
Model

Exploit
Public-
Facing
Application &

LLM Prompt
Injection

Phishing &

**ML Model
Access**

4 techniques

ML Model
Inference
API
Access

ML-Enabled
Product or
Service

Physical
Environment
Access

Full ML
Model
Access

Execution&

3 techniques

User
Execution &

Command
and Scripting
Interpreter &

LLM Plugin
Compromise

Persistence&

3 techniques

Poison
Training
Data

Backdoor ML
Model

LLM Prompt
Injection

**Privilege
Escalation**&

3 techniques

LLM Prompt
Injection

LLM Plugin
Compromise

LLM
Jailbreak

**Defense
Evasion**&

3 techniques

Evade
ML
Model

LLM
Prompt
Injection

LLM
Jailbreak

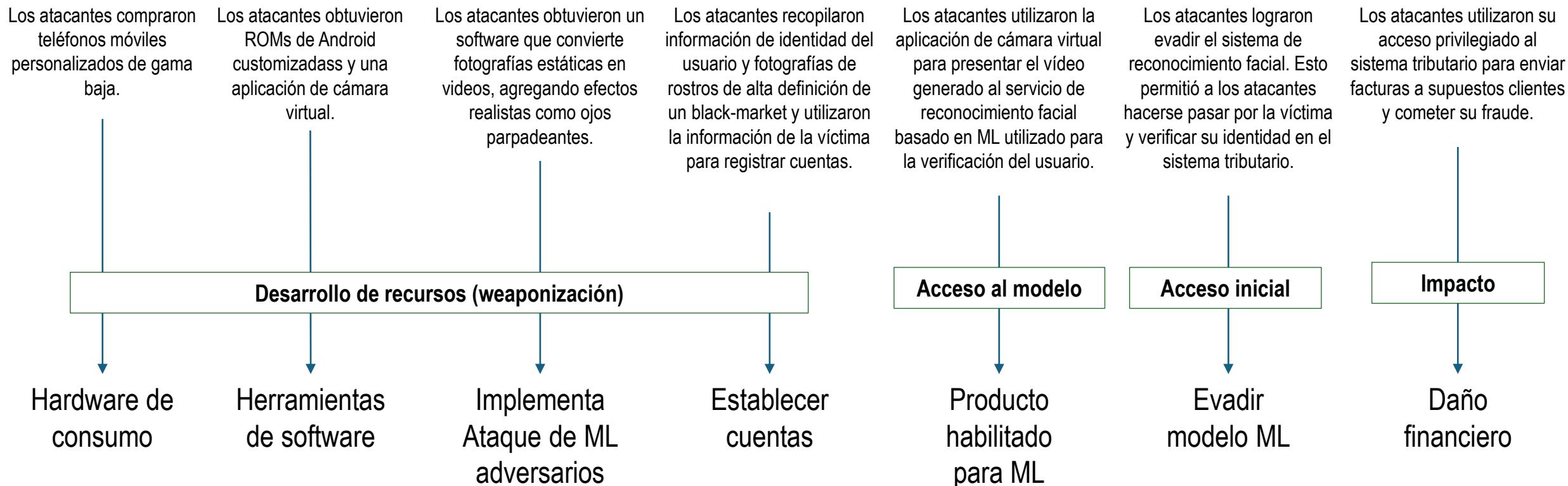
**Cred
Acc**

1 tec

Unsec
Crede

Camera Hijack Attack on Facial Recognition System

2020. Dos personas en China utilizaron este ataque para obtener acceso al sistema tributario del gobierno local. Crearon una empresa fantasma falsa y enviaron facturas a través del sistema fiscal a supuestos clientes. Los individuos pudieron recaudar de manera fraudulenta **77 millones de dólares**.



Camera Hijack Attack on Facial Recognition System

2020. Dos personas en China utilizaron este ataque para obtener acceso al sistema tributario del gobierno local. Crearon una empresa fantasma falsa y enviaron facturas a través del sistema fiscal a supuestos clientes. Los individuos pudieron recaudar de manera fraudulenta **77 millones de dólares**.

1. Configuración de un móvil barato para presentar un video fake

2. Creación de un vídeo, girando la cabeza, abriendo-cerrando los ojos y la boca, usando fotografías de usuarios en la Dark Web





DEFEND

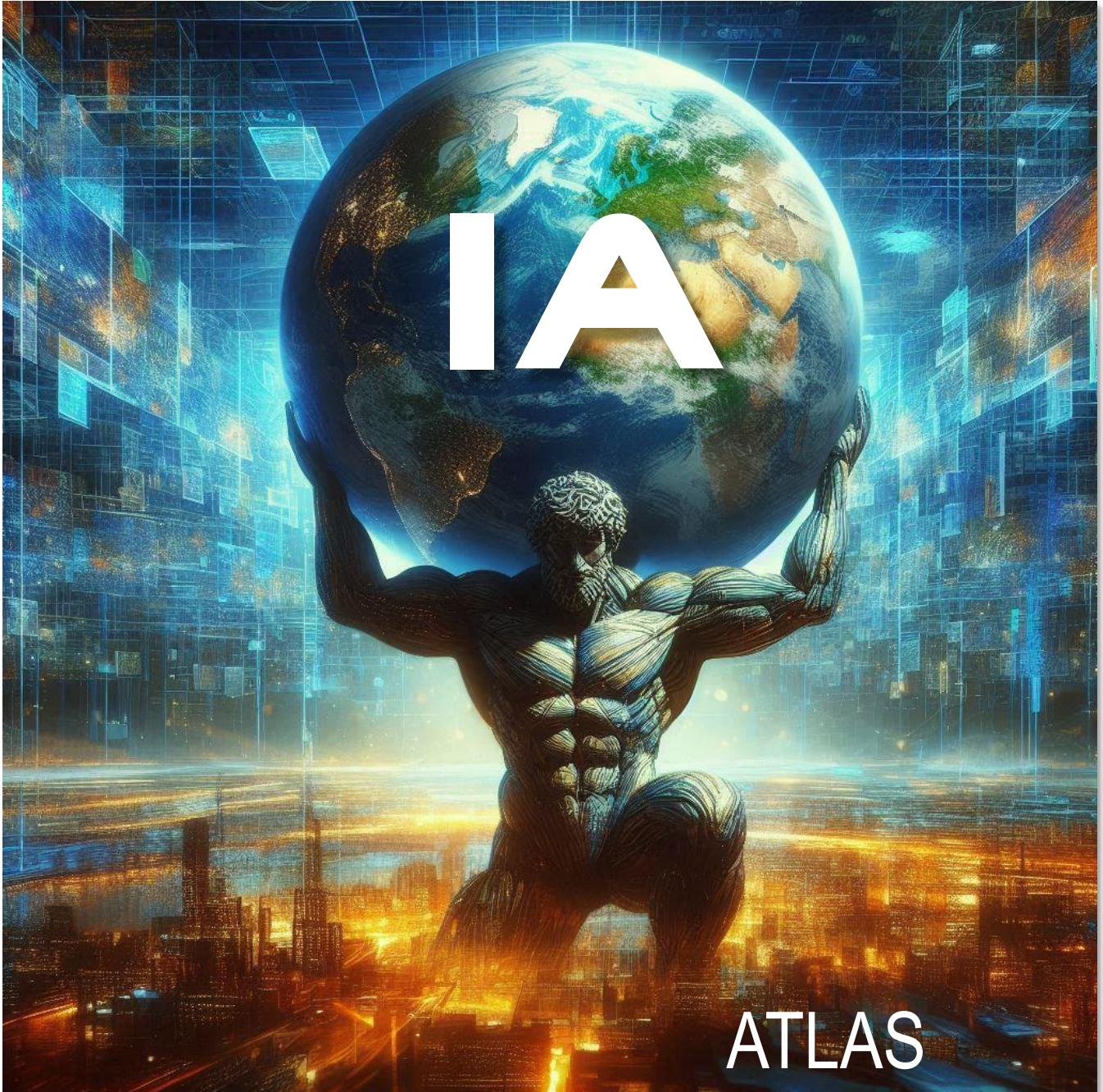
DEMO



"La tecnología nos impulsa hacia adelante

...

pero la sabiduría garantiza que nos dirijamos en la dirección correcta"





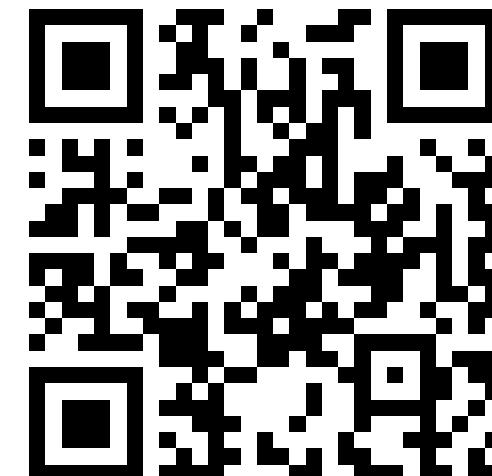
Emilio Rico
Security Advisor at



@Emilio_RR

Emilio Rico Ruiz

<https://github.com/3MlioRR>



<https://start.me/p/n7d5w9/atlas>



