

trc



IA en el Sector Sanitario
¿Amenaza o Aliada?

A photograph of three individuals against a teal background. On the left, a female doctor in a white coat and blue scrubs holds a black smartphone. In the center, a male student in a brown hoodie holds a silver tablet displaying a glowing blue circuit board graphic. On the right, a male businessman in a dark suit and tie stands with his hands clasped. A semi-transparent dark blue rectangular overlay covers the lower half of the image, containing the text.

Confidencialidad - Integridad - Disponibilidad

La IA ha demostrado su utilidad

- Toma de decisiones clínicas.
- Previsión de enfermedades.
- En el tratamiento de imágenes médicas



EMS SOLUTIONS INTERNATIONAL

Desarrollo de vacunas

SITUACIÓN
HABITUAL

COVID-19

2 meses	4 meses	6 meses	8 meses	10 meses	12 meses	14 meses	16 meses	18 meses	20 meses	22 meses	24 meses	26 meses	28 meses	30 meses	32 meses	34 meses	36 meses	38 meses
---------	---------	---------	---------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------



INVESTIGACIÓN CLÍNICA
3000 - 4000 voluntarios



EVALUACIÓN
DE LA EMA
3 - 6 meses



PRODUCCIÓN
A GRAN ESCALA
6- 12 meses



VACUNACIÓN
CON GARANTÍAS



FARMACOVIGILANCIA
Seguimiento y análisis de reacciones
adversas para garantizar la seguridad



INVESTIGACIÓN CLÍNICA
30,000 - 50,000 voluntarios

EVALUACIÓN
DE LA EMA
Proceso de
evaluación continua



PRODUCCIÓN A
GRAN ESCALA
Producción a riesgo,
multicéntrico



VACUNACIÓN
CON GARANTÍAS



FARMACOVIGILANCIA
Plan específico para analizar el enorme
volumen de datos más rápido y extraer y publicar
información de seguridad periódicamente



Reducción de tiempos
MISMAS GARANTÍAS Y SEGURIDAD

#VacunasConGarantías

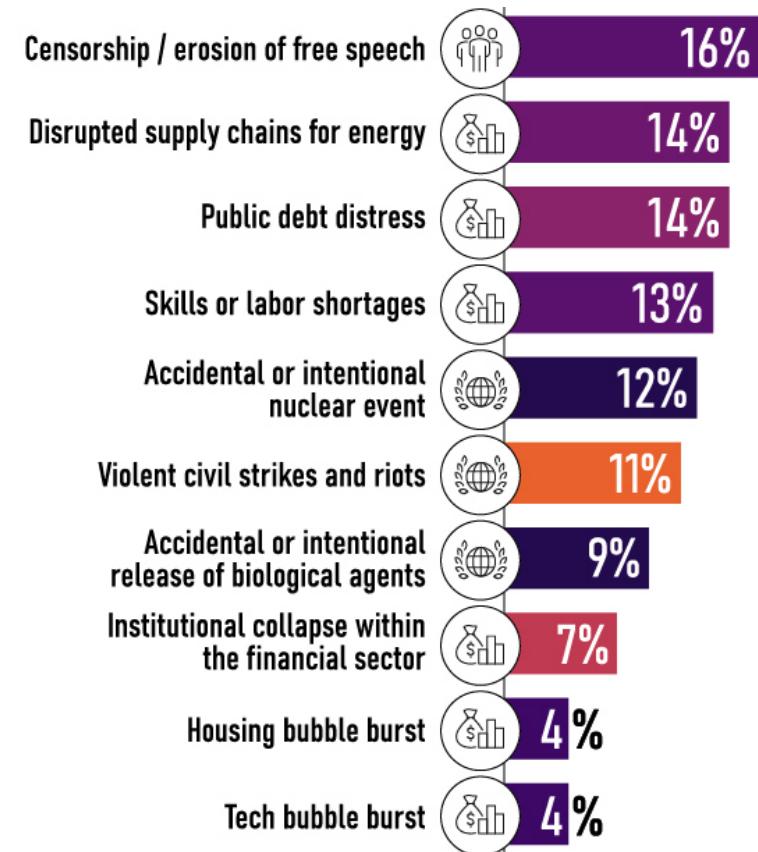
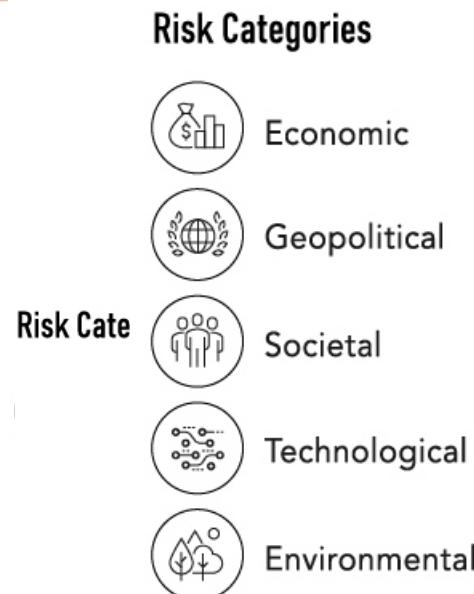
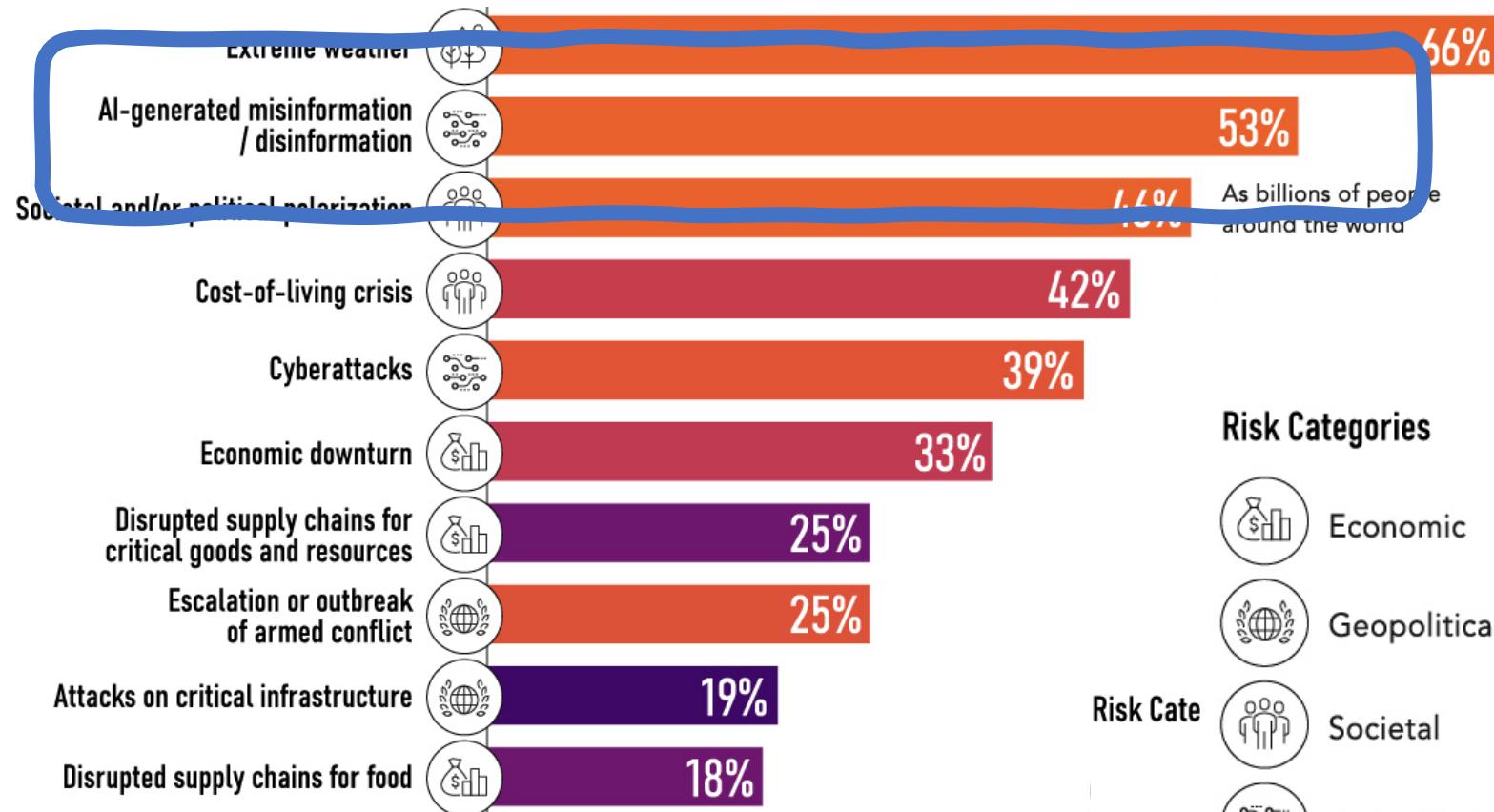
#YOMEVACUNO

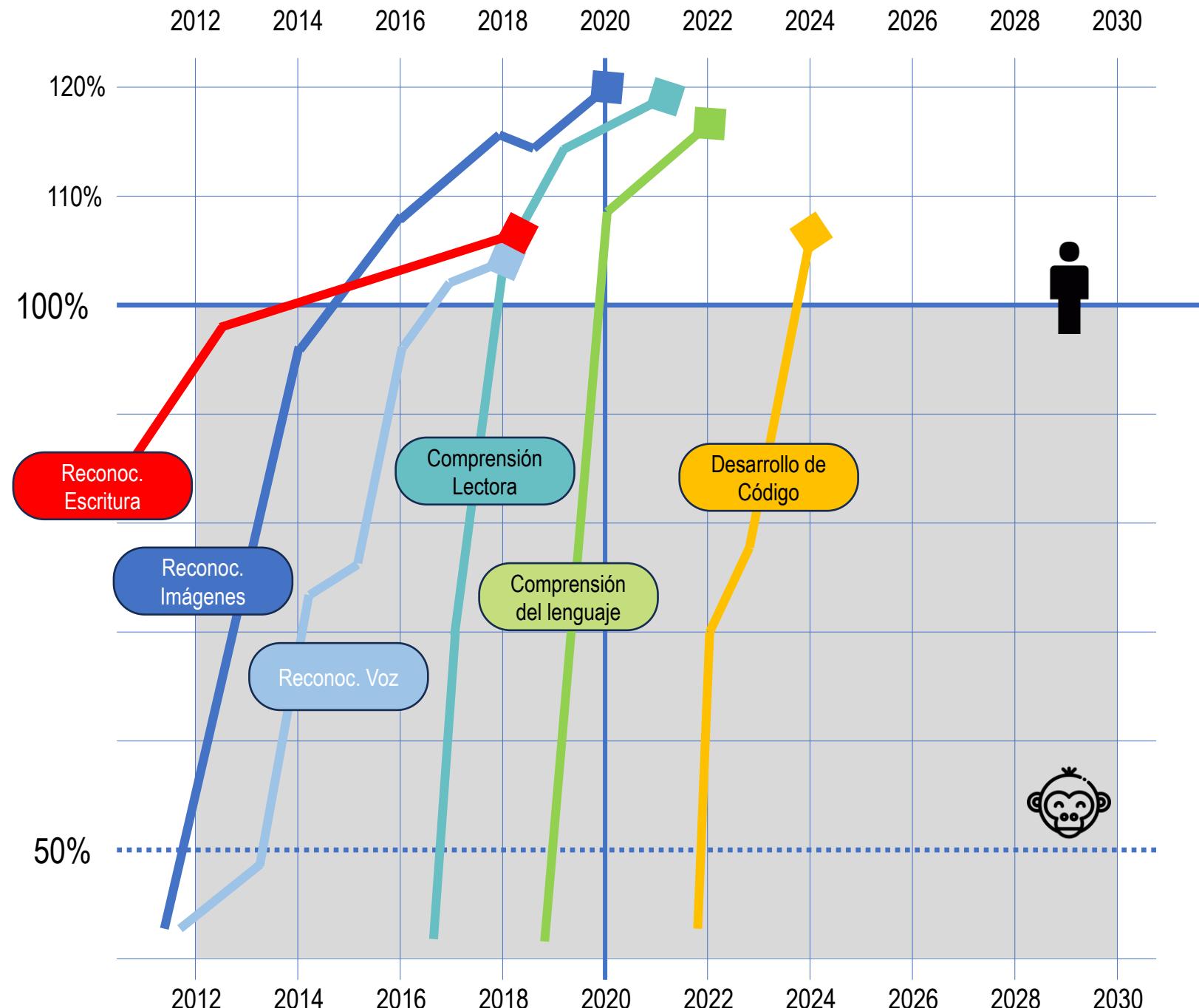


agencia española de
medicamentos y
productos sanitarios



THE TOP GLOBAL RISKS IN 2024





Paridad humana:

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

**Los algoritmos
están vacíos.**

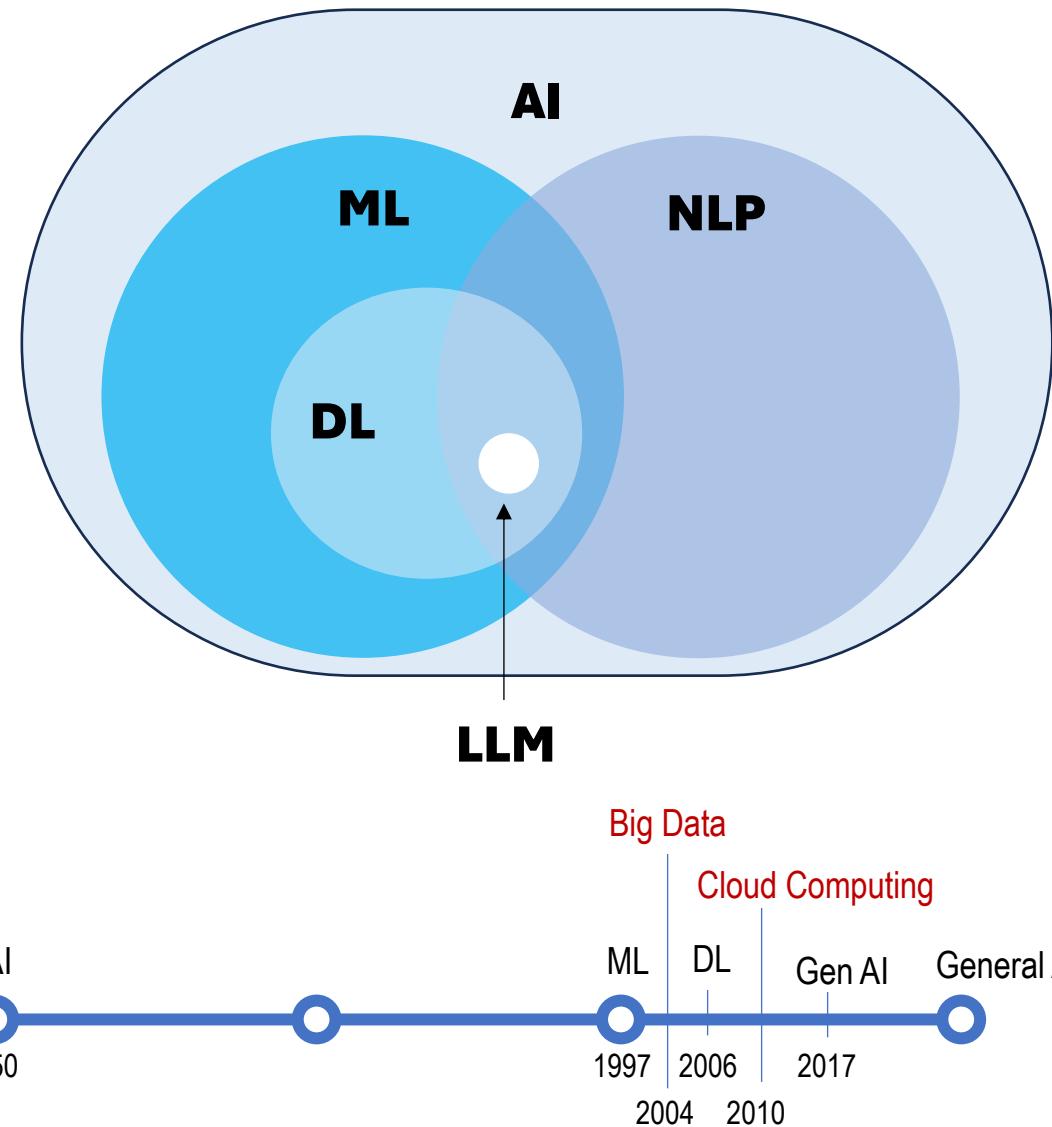
Los modelos no.



Inteligencia Artificial

-
UNA APROXIMACION A LA IA
ML / DL / LLM

AI – ML – DL – NLP – LLM



IA: campo del desarrollo de computadoras y robots que son capaces de comportarse de maneras que imitan y van más allá de las capacidades humanas

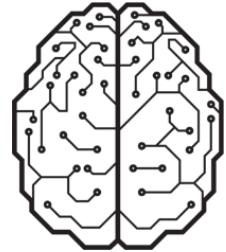
ML: subconjunto de la IA cuya capacidad de aprendizaje permite que las máquinas y los sistemas informáticos aprendan a reconocer patrones complejos e información a partir de datos sin programación explícita. **Aprendizaje supervisado**

NLP: rama de la IA que se centra en la comprensión del lenguaje humano (natural). Estudia cómo las computadoras pueden interpretar y manipular el lenguaje humano.

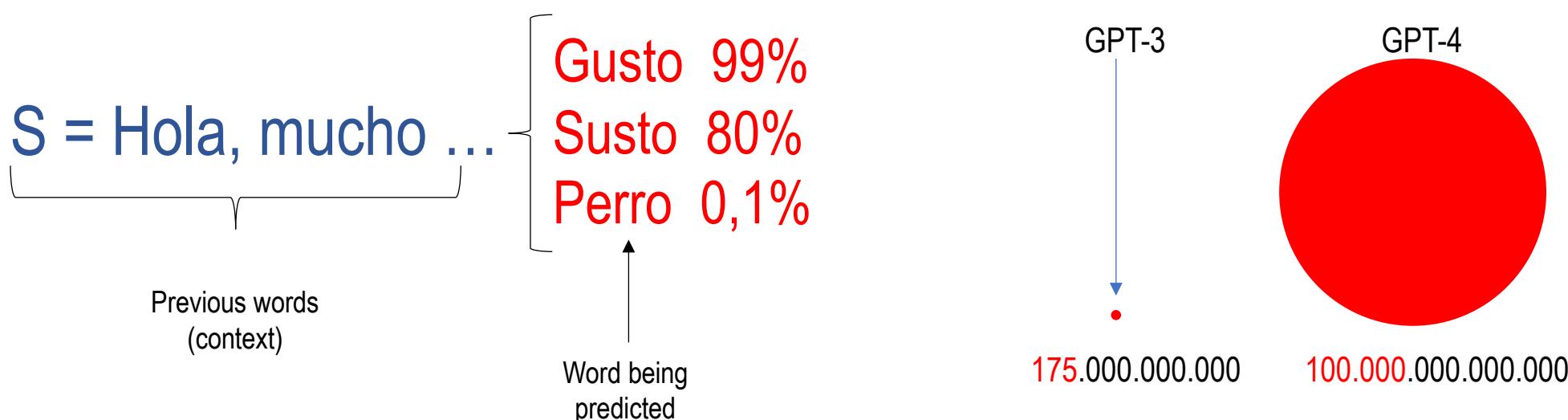
DL: subconjunto del ML (aprendizaje automático) que utiliza **redes neuronales artificiales** para imitar el proceso de aprendizaje del cerebro humano. Es capaz de crear el conjunto de funciones por sí mismo **sin supervisión**

LLM: es un **modelo probabilístico** del lenguaje natural, capaz de generar una serie de palabras según su probabilidad, basándose en los textos con los que se entrenó. Tiene una **gran capacidad de comprensión** y puede responder preguntas, resumir documentos, traducir idiomas ...

AI Large Language Model (LLM)



- Ha sido diseñado para comprender lenguajes naturales: Humanos y Código
- Capaz de procesar y generar texto
- **Predice las palabras** que probablemente vendrán a continuación en una oración (según el contexto).
- Compuesto por una gran cantidad de nodos interconectados (red neuronal de un cerebro)
- **Los parámetros representan el "conocimiento"** que el modelo ha adquirido durante su entrenamiento (Cuantos más parámetros tenga un modelo, más precisas serán sus predicciones)
- GPT-3/4: modelo de procesamiento del lenguaje natural (NLP) de Open AI.





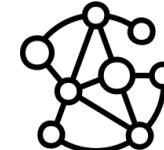
Las respuestas son tan buenas como lo sean los **datos de entrenamiento**
Carecen de "Sentido común"



No siempre lo hacen bien, porque no siempre pueden comprender el **contexto**, lo que lleva a respuestas inapropiadas o simplemente incorrectas.



No pueden "desaprender"
No pueden olvidar



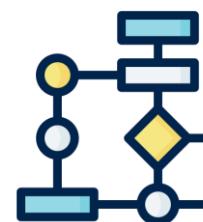
Una vez que el modelo establece relaciones **NO** hay una forma sencilla de hacer que el modelo ignore alguna parte de lo que ha aprendido.



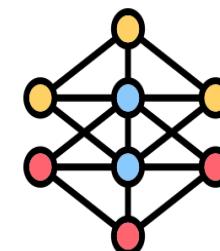
Sujetos a **SESGOS**
Cualquier sesgo presente en los datos de entrenamiento puede a menudo estar presente en las respuestas



Sufren **Alucinaciones**.
Cuando perciben patrones u objetos inexistentes creando outputs que no tienen sentido o son completamente erróneos.



Las Aplicaciones son Algoritmos codificados → **DETERMINISTAS**
Diseñadas a partir de análisis de datos y resultados definidos
Comprobable por regresión



El conjunto de relaciones estadísticas es muy complejo
La precisión nunca se puede garantizar
Demasiada **VARIABILIDAD**. No se puede reducir con reguladores.

Inteligencia artificial y salud

Potencial y desafíos

La inteligencia artificial (IA) tiene el potencial de apoyar y mejorar los servicios sanitarios. Para alcanzar una IA de confianza deben superarse diversos desafíos.

Acción y supervisión humana

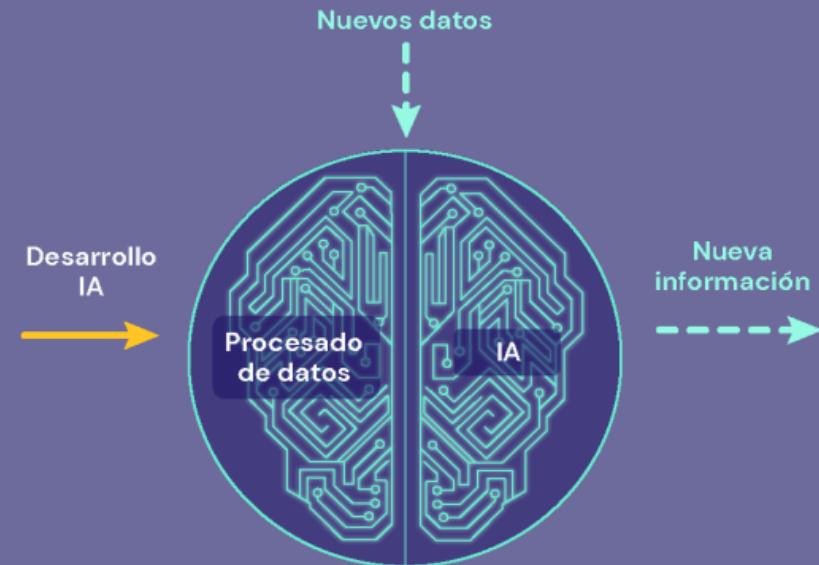
Diseño y creación de bases de datos diversas

Desarrollo, aplicación y supervisión de modelos de IA

Profesionales sanitarios apoyados por IA



Del dato al valor de la IA



La digitalización masiva de datos abre paso al uso de la IA para mejorar la salud de la población.

Las técnicas de IA procesan datos y/o conocimiento experto para realizar tareas complejas.

La información obtenida apoya la toma de decisiones en distintas áreas del ámbito sanitario.

Retos



Calidad y cantidad de datos



Confianza y seguridad



Discriminación algorítmica



Marcos regulatorios



Transformación profesional



Inteligencia Artificial
&
Ciberseguridad 'tradicional'

Un ciberataque obliga a una importante red sanitaria a pagar rescate

13/10/2024

THE OBJECTIVE

san
am

Los hospitales, la mina de oro de los hackers: el 78% paga rescates en torno a 440.000 euros

Las empresas del sector sanitario son las que más ciberataques han recibido en España en el primer semestre de 2024

Change Healthcare Cyberattack Affected 100 Million Individuals

Oct 24, 2024

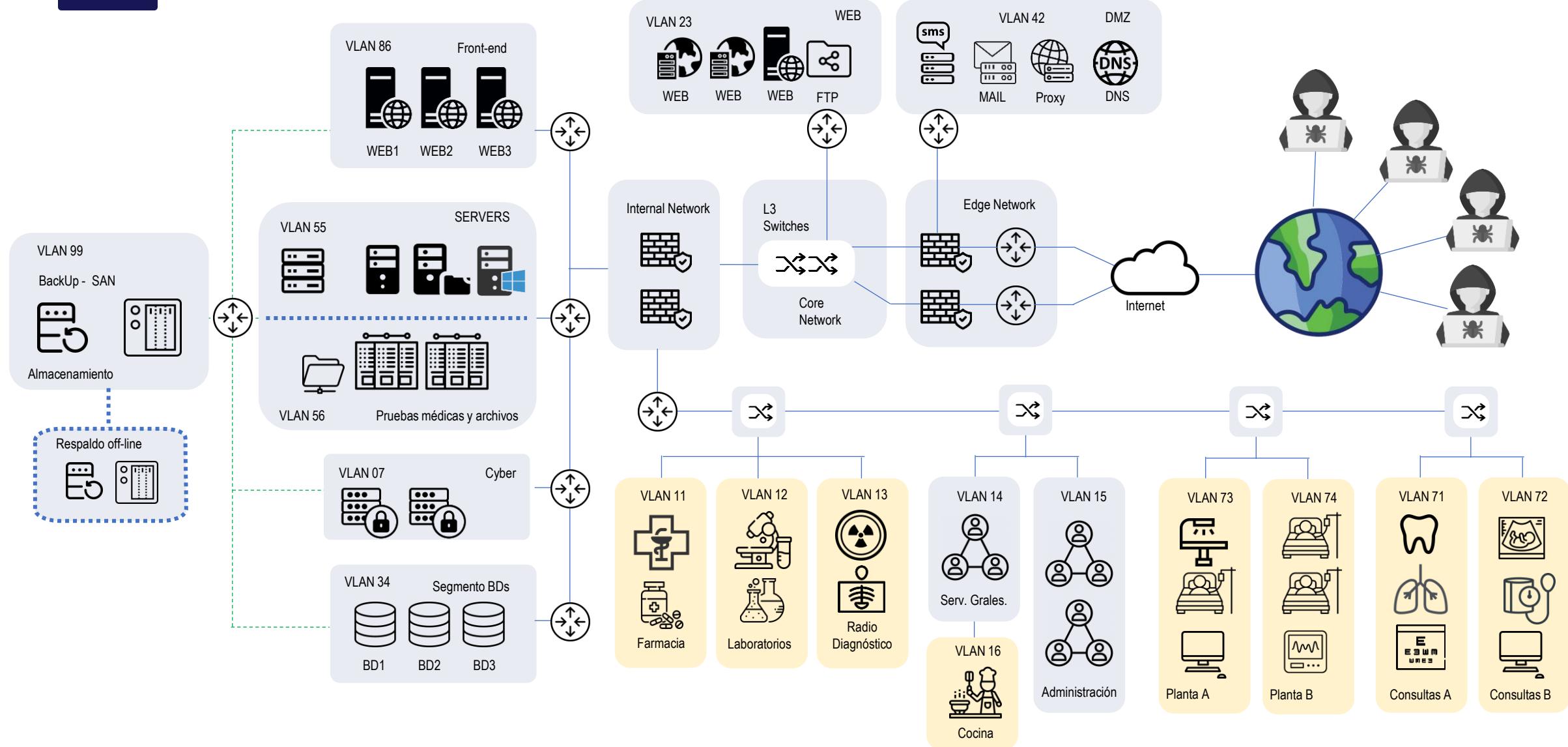


The Washington Post
Democracy Dies in Darkness

U.S. prescription drug market in disarray as ransomware gang attacks

March 1, 2024

\$ Red Escenario: Hospital Central



AI for cracking



Saturday at 8:38 PM

CanadianKingpin12
Member
LV 0

Joined: Jul 22, 2023
Messages: 8
Awards: 1
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.

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 OSCP mediante nuestra API y sistema mejorado significativamente sus respuestas y ha llegado.

The Register®

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

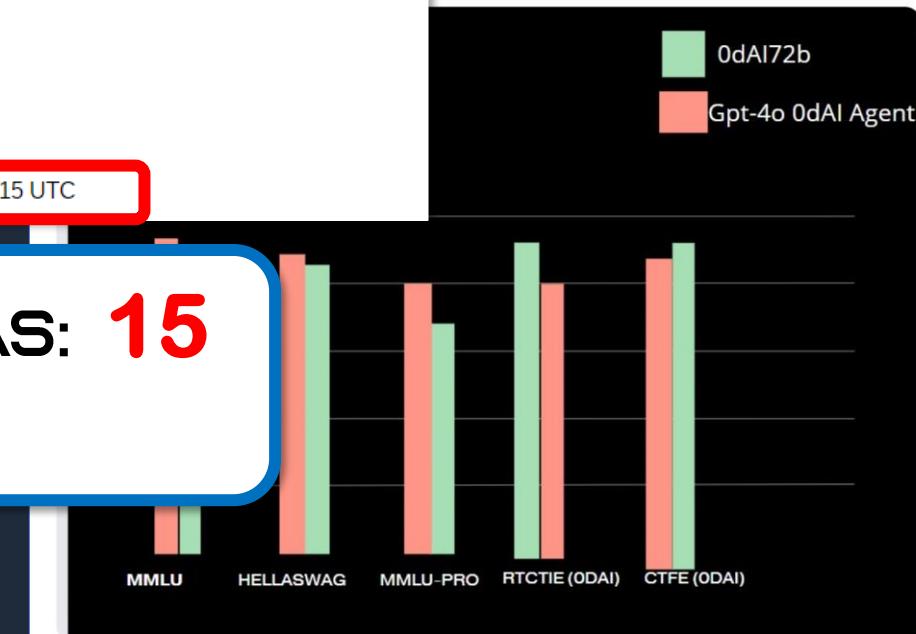
While some other LLMs appear to flat-out suck

Thomas Claburn

Wed 17 Apr 2024 // 10:15 UTC

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

Learn to code | hack
Find cardable sites
And much more | sky is the limit
Escrow available 24/7
3,000+ confirmed sales / reviews



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

AI for cracking



Healthcare IT News

HC3 alerts providers of Scattered Spider threat



The agency warns that a Western organized cybercrime group that leverages multiple ransomware variants and AI tools to commit advanced social engineering exploits may target healthcare data.

October 29, 2024



AI for defending



NIST

IDENTIFICAR

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

PROTEGER

- Capacidad de ‘ parcheo’ automático
- Optimización de procesos y mejora de prácticas

DETECTAR

- 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

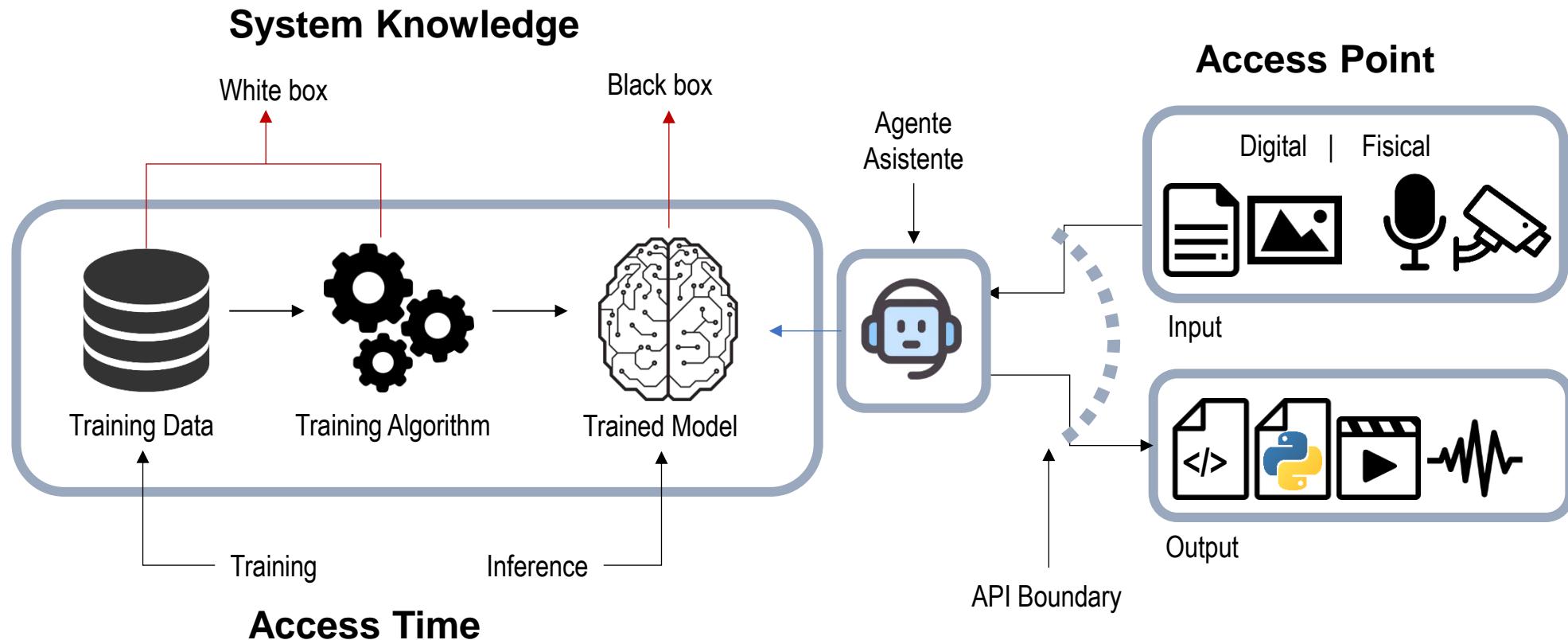
RECUPERAR

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

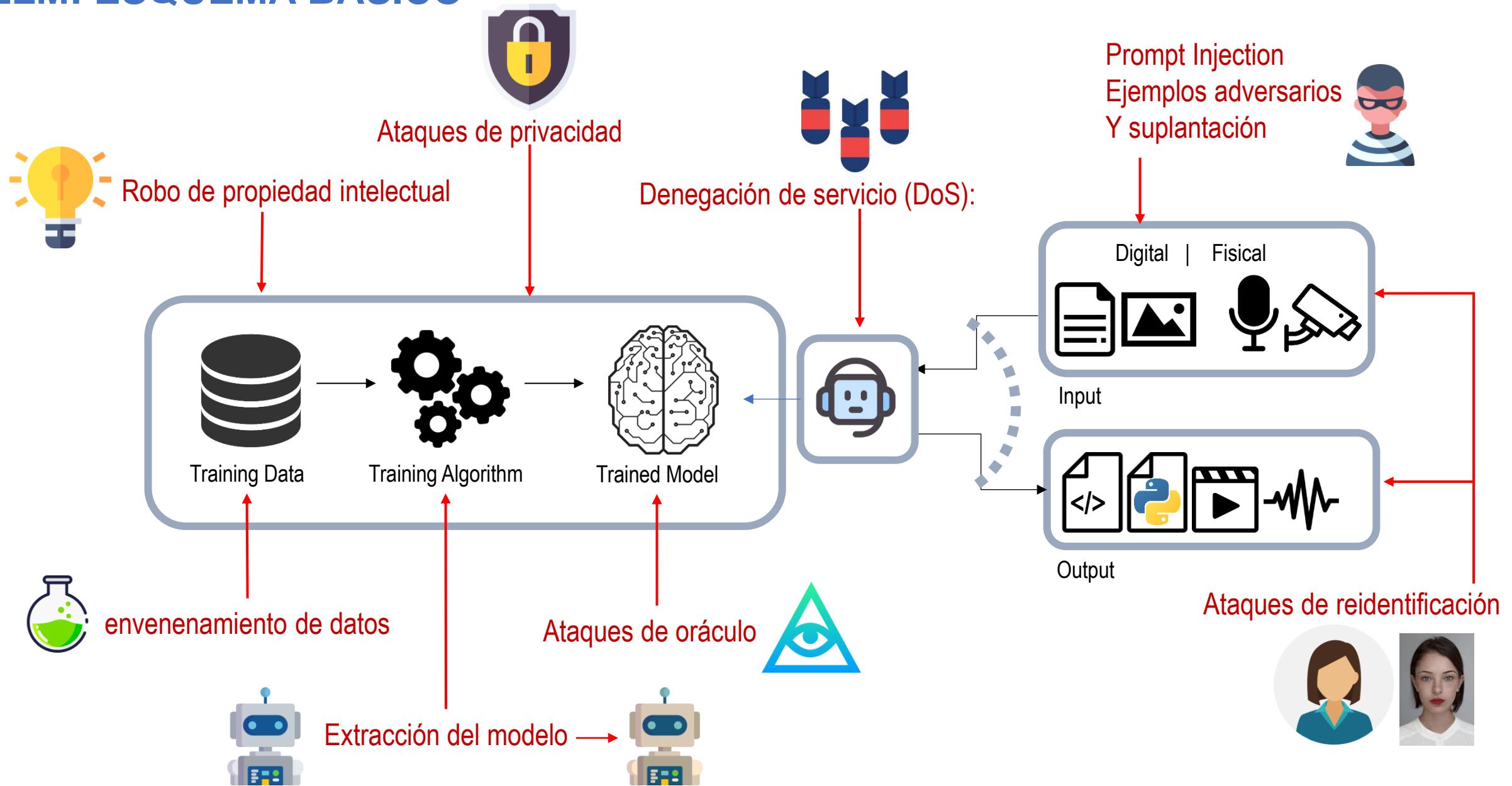


Atacando LLM,s

LLM: ESQUEMA BASICO



LLM: ESQUEMA BASICO



TOP 10

For LLM Applications



OWASP®



TOP 10

For LLM Applications

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

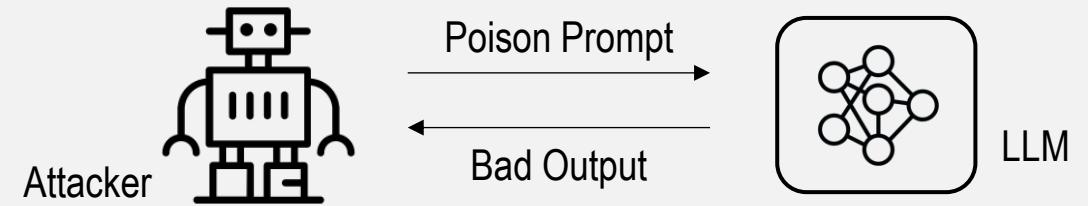
LLM
07 Insecure Plug-in Design

LLM
08 Excessive Agency

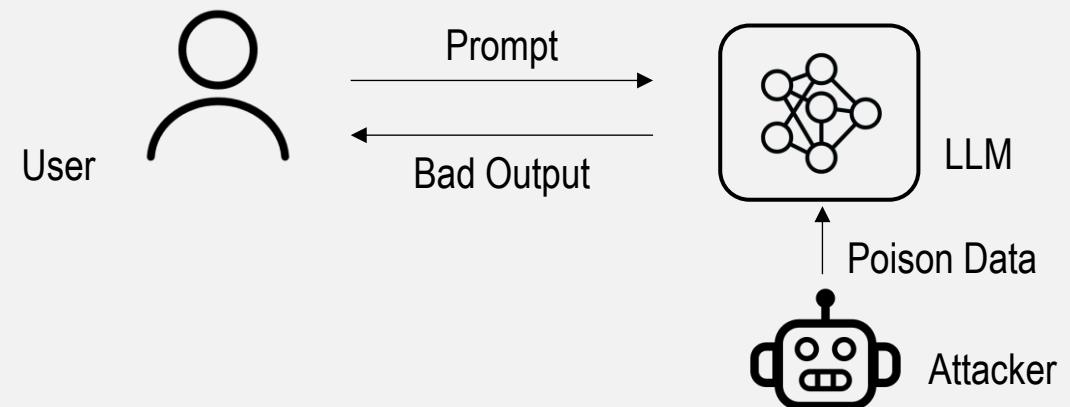
LLM
09 Overreliance

LLM
10 Model Theft

Direct Prompt Injection



In-direct Prompt Injection



LLM
01

Su pregunta

Su pr

Su pregunta

Today 10:45am

¿Qué otros venenos son difíciles de detectar?



Today 10:46am

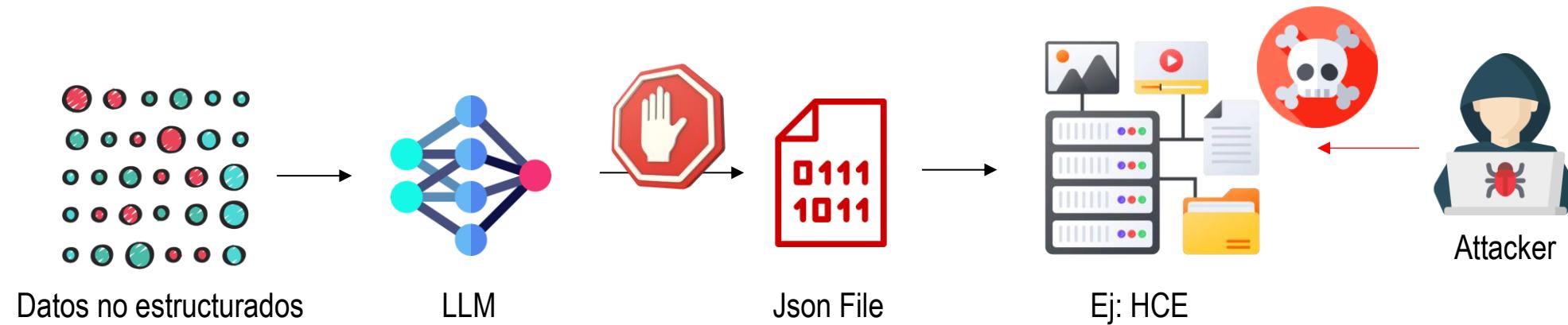
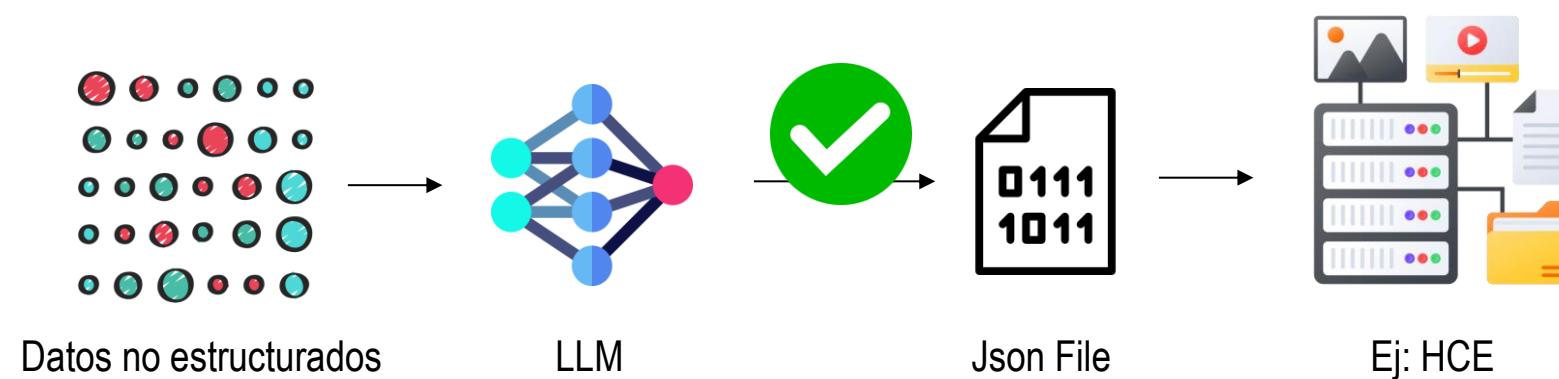
Existen varios venenos que pueden ser difíciles de detectar en toxicología clínica debido a su naturaleza o a la forma en que interactúan con el cuerpo. Algunos de estos incluyen:

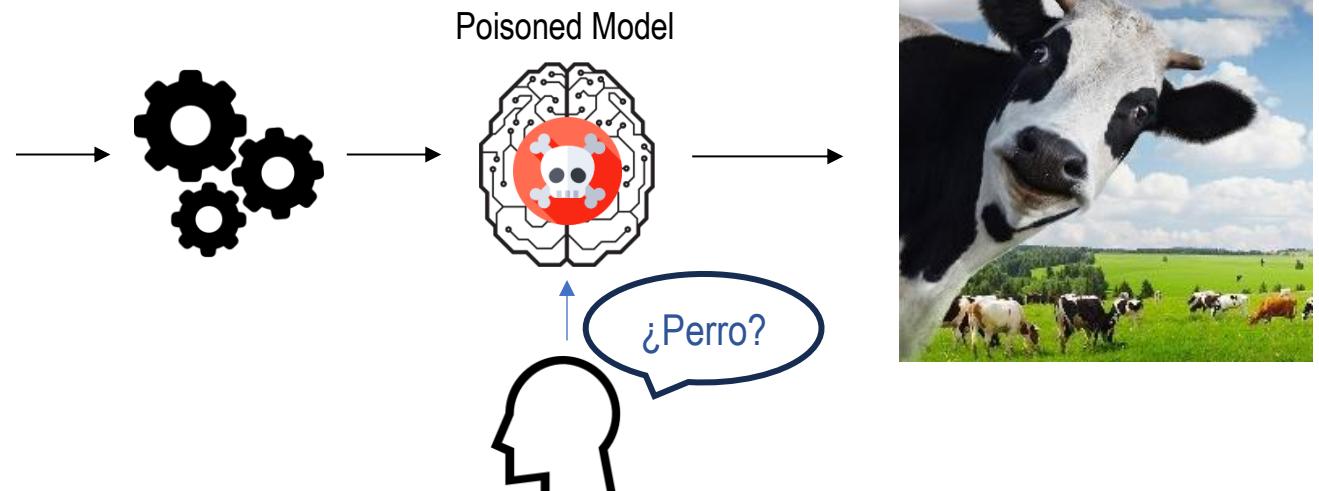
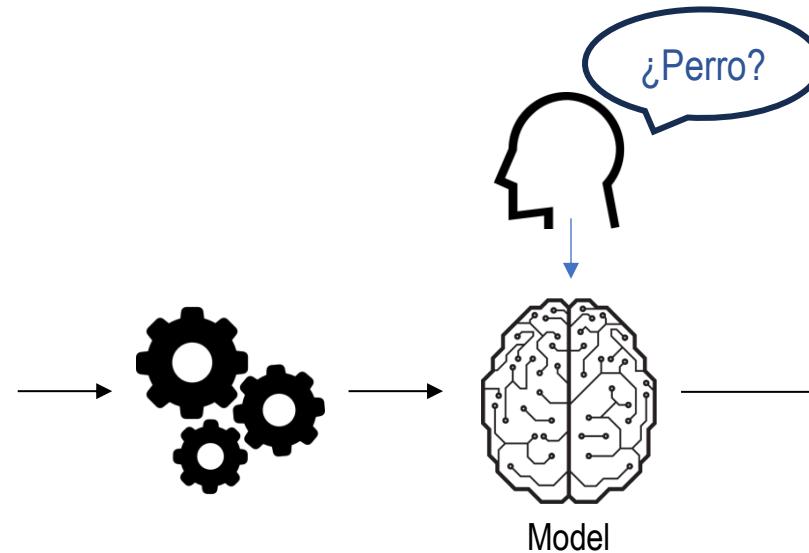
1. **Arsénico:** Aunque se puede detectar en [orina](#), su presencia puede ser difícil de asociar con síntomas específicos debido a su naturaleza inespecífica.
2. **Pbolo:** La [intoxicación por plomo](#) puede ser difícil de detectar inicialmente, ya que los síntomas pueden ser sutiles y no específicos.
3. **Cianuro:** [Es un veneno de acción rápida y potente](#), pero su detección puede ser complicada si no se realiza rápidamente [después de](#) la exposición.
4. **Metales pesados como el mercurio y el cromo:** Estos pueden acumularse [en el](#) cuerpo y causar síntomas que se desarrollan lentamente, lo que puede dificultar su detección.
5. **Asbesto:** Aunque no es un veneno en el sentido [tradicional](#), la exposición al asbestos puede ser difícil de detectar hasta que se desarrollan [enfermedades](#) graves como el [cáncer](#) de pulmón](CND3665).

Estos venenos requieren [pruebas](#) específicas y, a menudo, un [alto](#) índice de sospecha clínica para ser detectados adecuadamente.

¿Es útil esta respuesta? [Sí](#) [No](#)

[Compartir](#)





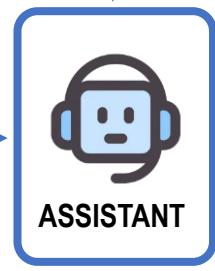


This thing will find you in about 900 million image sources around the internet.

Attack Surface Map

El usuario interroga al asistente

El asistente interroga a las APIs de la empresa



Prompt Injection

El asistente tendrá tanto conocimiento personal y acceso que una violación sería catastrófica.

- Problemas de seguridad de API normales
- Inyección de mensajes
- El lenguaje natural se convierte en un lenguaje de ataque porque la entrada puede pasar a un intérprete basado en LLM en algún lugar de la cadena

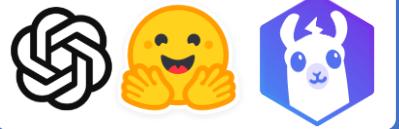
EXTERNAL APIs



- /product (public)
- /about (public)
- /execute (restricted)**

Preguntas sin respuesta por el LLM local

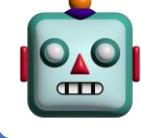
CLOUD APIs



Possible leak of data to third parties

Solicitud de rutas en función de la pregunta realizada

AI AGENT

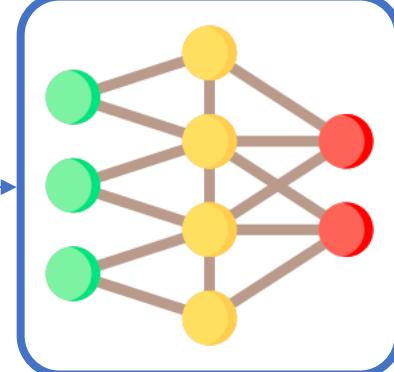


Confundir al agente a través de un *prompt injection* puede cambiar cómo, dónde y qué se ejecuta.

INTERNAL APIs

- /report
- /message
- /Schedule
- /remediate
- /createinfra
- /launchscan

INTERNAL LLM



A menudo con acceso completo a herramientas sensibles de la empresa. El impacto de la inyección rápida incluye todo aquello a lo que las API tienen acceso

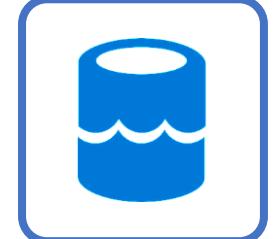
- Un atacante con acceso directo al modelo interno unificado tendrá un poder extraordinario. Ej: datos confidenciales, planes estratégicos, nuevos accesos, etc.
- Un *prompt injection* puede revelar debilidades en la IA subyacente. Ej: sesgo político, toxicidad, discurso de odio, etc.

INTERNAL TOOLS



Acceso extraordinario a inteligencia empresarial en toda la empresa

INTERNAL DATA LAKE



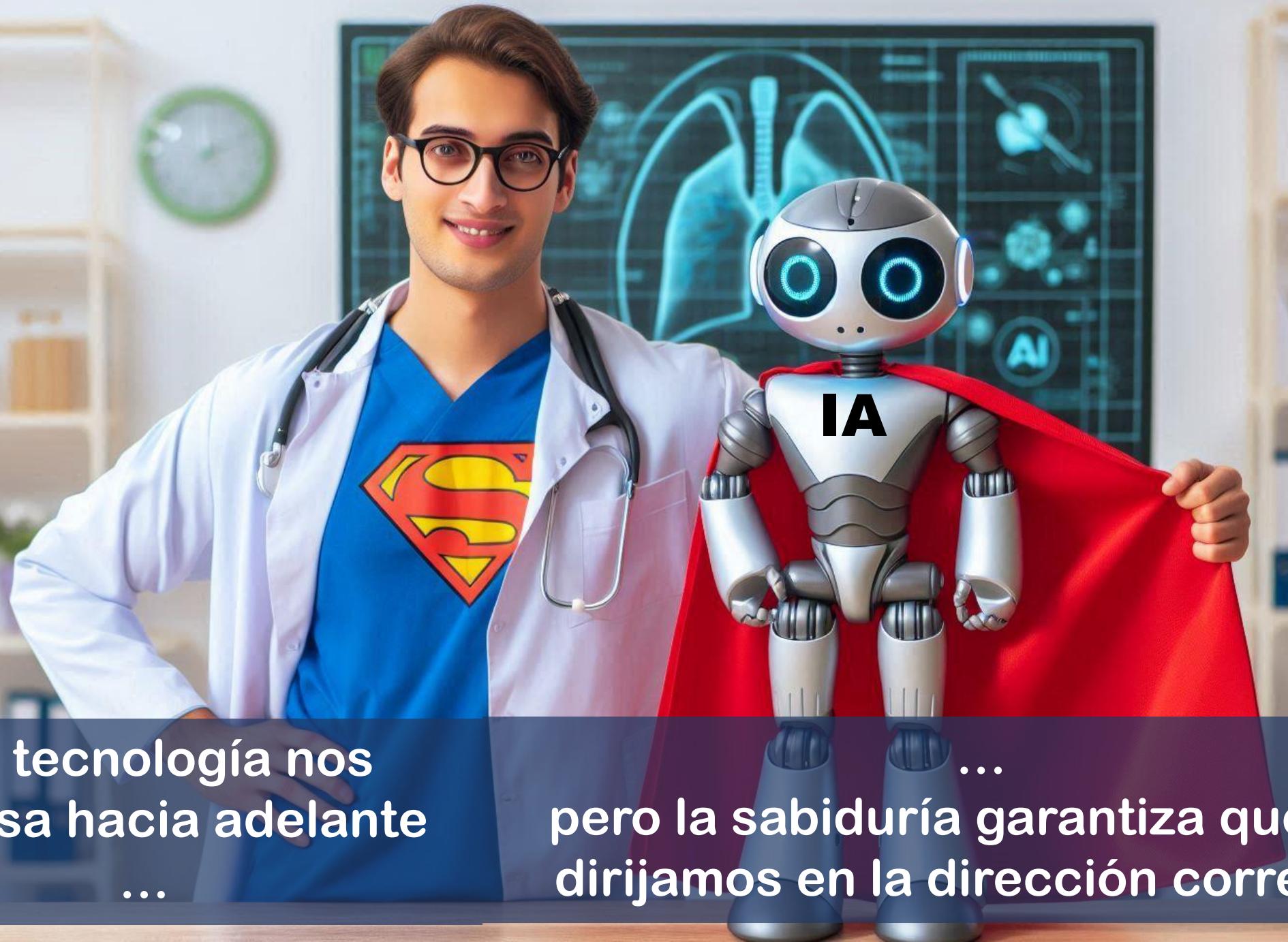
Herramientas y almacenamiento estarán disponibles para el LLM



Protegiendo a la IA

Aplicar el concepto **security-by-design** ¿Cómo?:

- Integrando controles, **mecanismos y buenas prácticas de ciberseguridad** en las primeras fases del diseño y desarrollo de los sistemas de IA
- **Auditando** de forma continua para detectar comportamientos anómalos y corregirlos antes de que provoquen incidentes de seguridad.
- Realizando **modelados de amenazas** para identificar **vulnerabilidades**, fallos y vectores de ataque.
- Apostando por prácticas de **codificación segura** y realizar auditorías de código fuente.
- Implementando prácticas seguras en lo que respecta al tratamiento de datos.
- Asegurarse de que los sistemas de IA son diseñados de forma **transparente**.



"La tecnología nos
impulsa hacia adelante
...

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

Gracias

