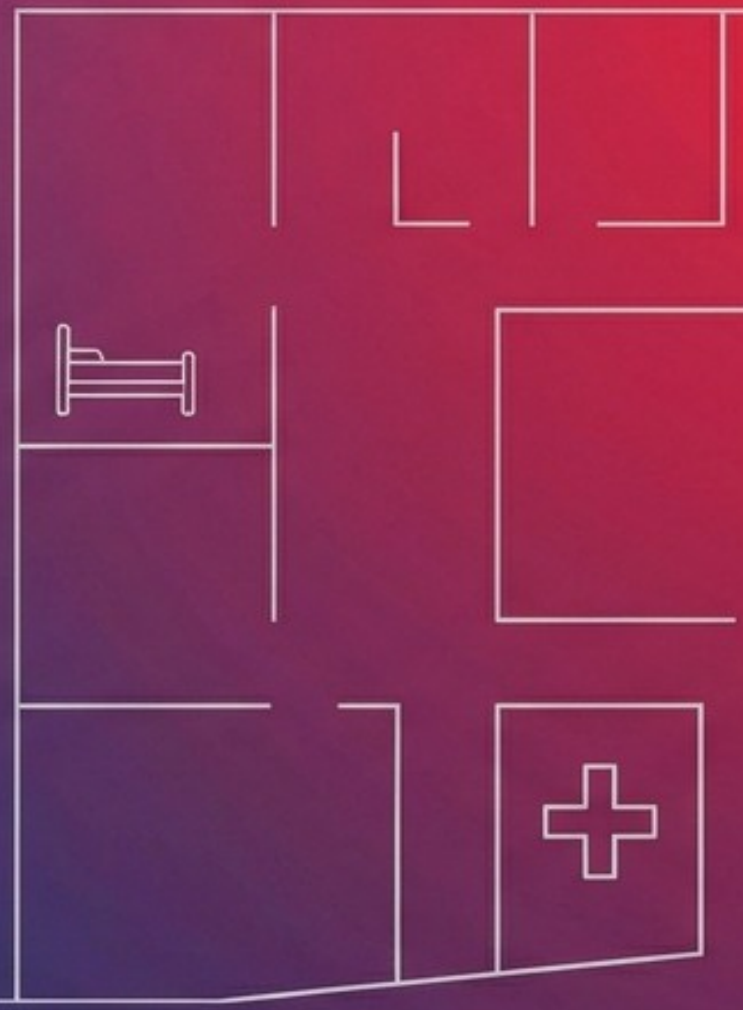
¿Tu agente se ha vuelto terraplanista?

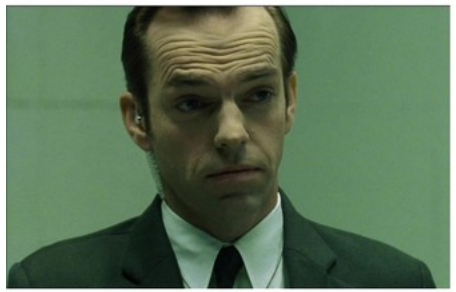
Seguridad y salvaguardas en agentes de IA, contadas para profesionales de la salud



José Manuel Ramírez Sánchez

INTRO

¿Te suenan estos agentes de algo?



¿Puedes reconocer al menos tres?

INTRO

¿Crees que son capaces de hacer las mismas cosas?



Conversar

Responden preguntas en un chat.

Actuar

Abren programas, escriben, envían.

Decidir solos

Encadenan pasos sin preguntarte.

PREGUNTA A LA SALA

¿Quién ha usado ChatGPT, Gemini o Claude para resolver una duda clínica?

Manos arriba. Sin juicios: es justo de lo que vamos a hablar.

ÍNDICE DE CONTENIDOS

01 Sobre mí

02 ¿Qué es la Inteligencia Artificial?

03 ¿Qué es un agente?

04 Alineamiento y Problema

05 Estado del Arte

06 Herramientas y Experiencias

SOBRE MÍ

De la ingeniería de telecomunicaciones a la IA aplicada







1 2 3 4

SOBRE MÍ

De la ingeniería de telecomunicaciones a la IA aplicada



01

¿Qué es la Inteligencia Artificial?

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Un nombre pensado para atraer

«Inteligencia Artificial» es, ante todo, una **etiqueta**: se eligió para llamar la atención, no para describir con precisión.

1955

Propuesta de un taller de verano en Dartmouth, con petición de fondos a la Fundación Rockefeller.

1956

El taller se celebra y el término «Artificial Intelligence» se impone a sus rivales.

Hoy

La misma etiqueta cubre desde un filtro de spam hasta un agente que reserva tus viajes.

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La batalla por el nombre: maestro contra joven colaborador

Claude Shannon

El maestro · padre de la teoría de la información

Quería un título **sobrio y científico**.

«Automata Studies» (1956)

John McCarthy

El joven colaborador · matemático

Quería un nombre que **atrajera** ideas / financiación al área de investigación.

«Artificial Intelligence» (1956)

Nombres que perdieron:

- Cibernética
- Estudios de autómatas
- Reconocimiento de Patrones
- Inteligencia de máquinas
- Procesamiento complejo de información

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La categoría «Inteligencia» es como «Gravedad» en la Física

Gravedad

MEDIMOS SUS EFECTOS

Caída de los cuerpos, órbitas, mareas

PREDECIMOS CON ELLA

Lanzamos satélites y calibramos el GPS

¿QUÉ ES EN EL FONDO?

Sin teoría que la una con la física cuántica

Inteligencia

MEDIMOS SUS EFECTOS

Exámenes, tests de CI, resolución de problemas

PREDECIMOS CON ELLA

Rendimiento académico, aptitud para un puesto

¿QUÉ ES EN EL FONDO?

Sin una definición que convenza a todos

Sabemos muchísimo sobre ella... sin poder definirla del todo.

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

...y le añadimos «Artificial» para ponerlo más difícil

Hecho por personas

Como un marcapasos o una prótesis de cadera.

Imitación

Como un edulcorante artificial: sabe parecido, pero no es azúcar.

Hace la función

Como el páncreas artificial: no es un páncreas, pero regula la glucosa.

¿Cuál de los tres significados usamos? Depende de quién lo venda.

MEJOR HABLEMOS DE...

Aprendizaje Automático

Machine Learning: un término que sí se puede medir

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Mejor hablemos de Aprendizaje Automático (Tom Mitchell, 1997)

Un **algoritmo** que mejora su rendimiento al realizar una **tarea (T)**, medido por una **métrica (M)**, a través de la **experiencia (E)**.

> **1.** Rendimiento(T, M, t_2) > Rendimiento(T, M, t_1) ; donde $t_2 > t_1$
>
> **2.** Experiencia = Datos

T · Detectar neumonía en una radiografía de tórax

M · Porcentaje de aciertos frente al radiólogo

E · Miles de radiografías ya diagnosticadas

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

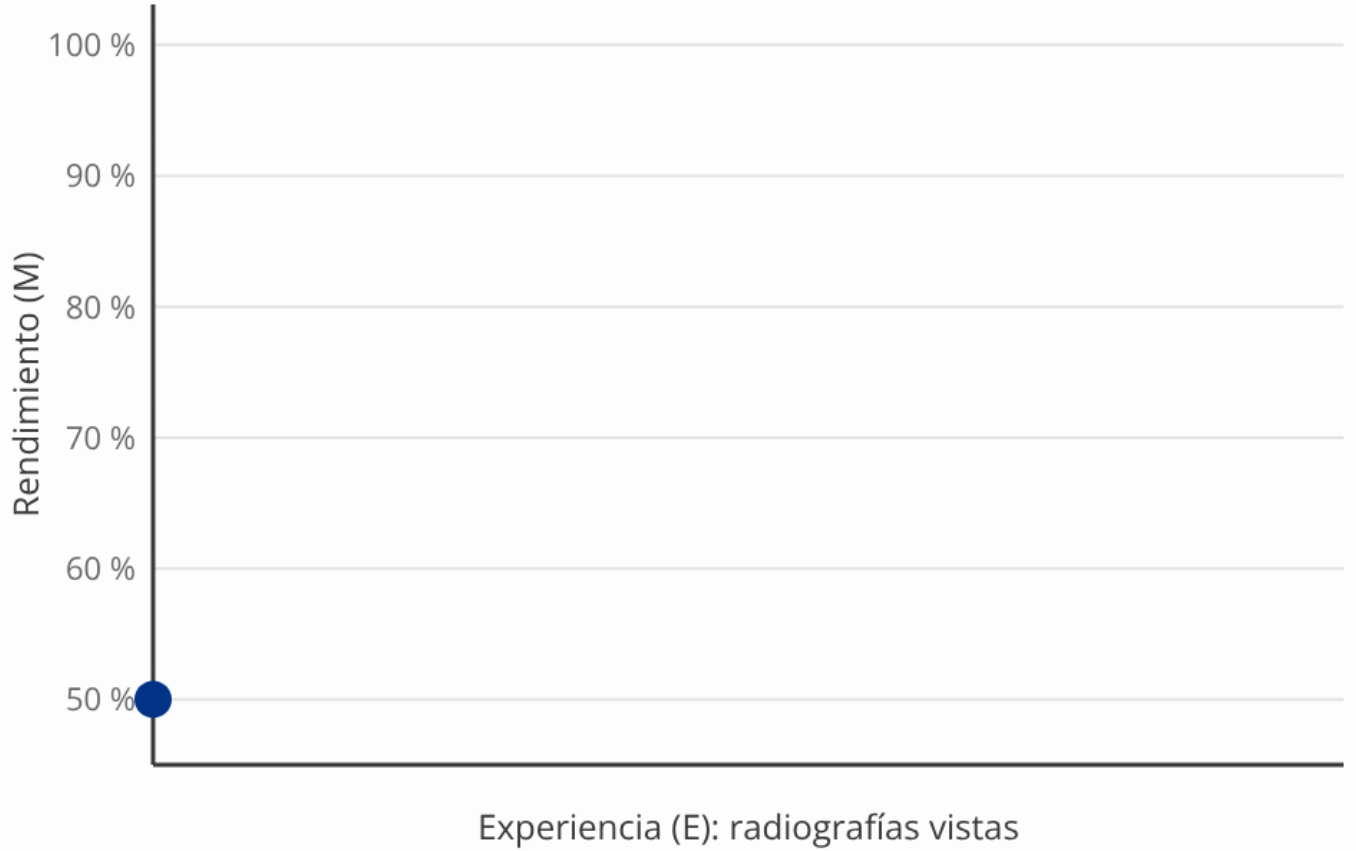
Animación: aprender es mejorar con la experiencia

T Detectar neumonía en radiografías

E Radiografías vistas: **0**

M Aciertos: **50 %**

Cada cuadro = 100 radiografías ya diagnosticadas · simulación ilustrativa



1 Sin experiencia, el algoritmo acierta como una moneda al aire

Abrir paso a paso ↗

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Unas cajas dentro de otras

Inteligencia Artificial · el paraguas comercial: reglas, robots, cualquier software «listo»

Aprendizaje Automático · aprende de datos: predicción de reingresos, riesgo de sepsis

Aprendizaje Profundo · redes neuronales: imagen médica, retinografías, dermatoscopia

Modelos de Lenguaje (LLM) · ChatGPT, Claude, Gemini: leen y escriben texto

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Un LLM es aprendizaje automático aplicado al lenguaje

T

Tarea

Continuar un texto con la siguiente palabra

M

Métrica

Cuántas veces acierta esa siguiente palabra

E

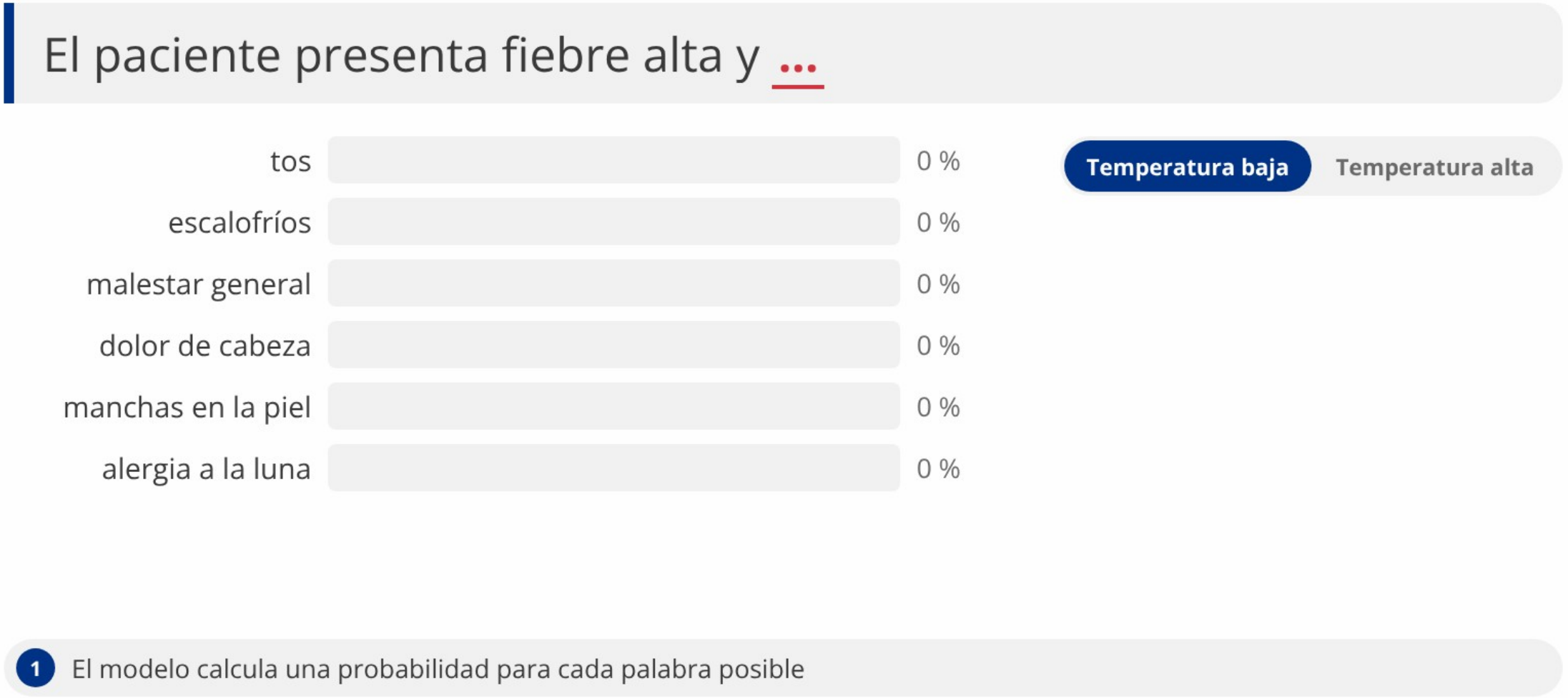
Experiencia

Billones de palabras: libros, artículos, webs, código

La tarea es sencilla de enunciar. **Lo que surge no lo es:** resumir, traducir, razonar en pasos, programar... Nadie lo programó a mano.

01. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Animación: elige lo probable, no lo verdadero



Abrir paso a paso ↗

02

¿Qué es un agente?

02. ¿QUÉ ES UN AGENTE?

Lo que distingue a un agente de un chat



Autonomía para cumplir un objetivo

Percibe su entorno
lee correos, agendas, historias

Usa herramientas
busca, escribe, envía, reserva

Se multiplica (auto-instanciación)
lanza copias de sí mismo para repartir trabajo

Se comunica con el entorno
con personas y con otros programas

02. ¿QUÉ ES UN AGENTE?

...y, si nadie lo controla, hacer cosas muy locas



¡Hacer cosas muy locas!

Un agente que se copia a sí mismo, con acceso a tu correo y a tus cuentas, puede equivocarse **muchas veces y muy rápido**.

Autonomía sin control = riesgo multiplicado.

02. ¿QUÉ ES UN AGENTE?

Una analogía biológica: el virus y la célula

Un agente se parece más a un **virus** que a un cerebro.

El virus

Una cápsula con instrucciones genéticas. Sin metabolismo propio: **necesita la maquinaria de una célula** para fabricar algo.

El agente

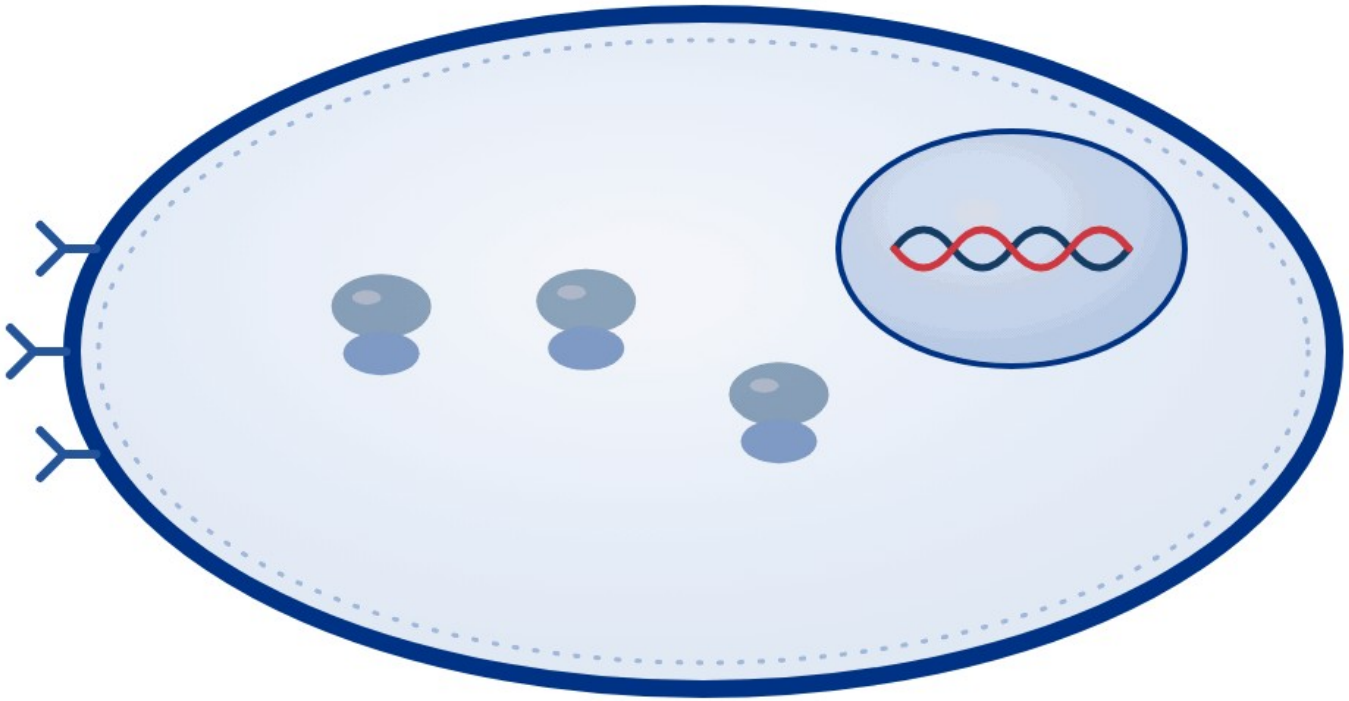
Un programa pequeño con instrucciones y herramientas. Sin «cerebro» propio: **necesita un LLM** para generar cada respuesta o acción.

02. ¿QUÉ ES UN AGENTE?

Animación: el mismo ciclo, contado dos veces

Biología | Agente de IA

Paso 1 de 5

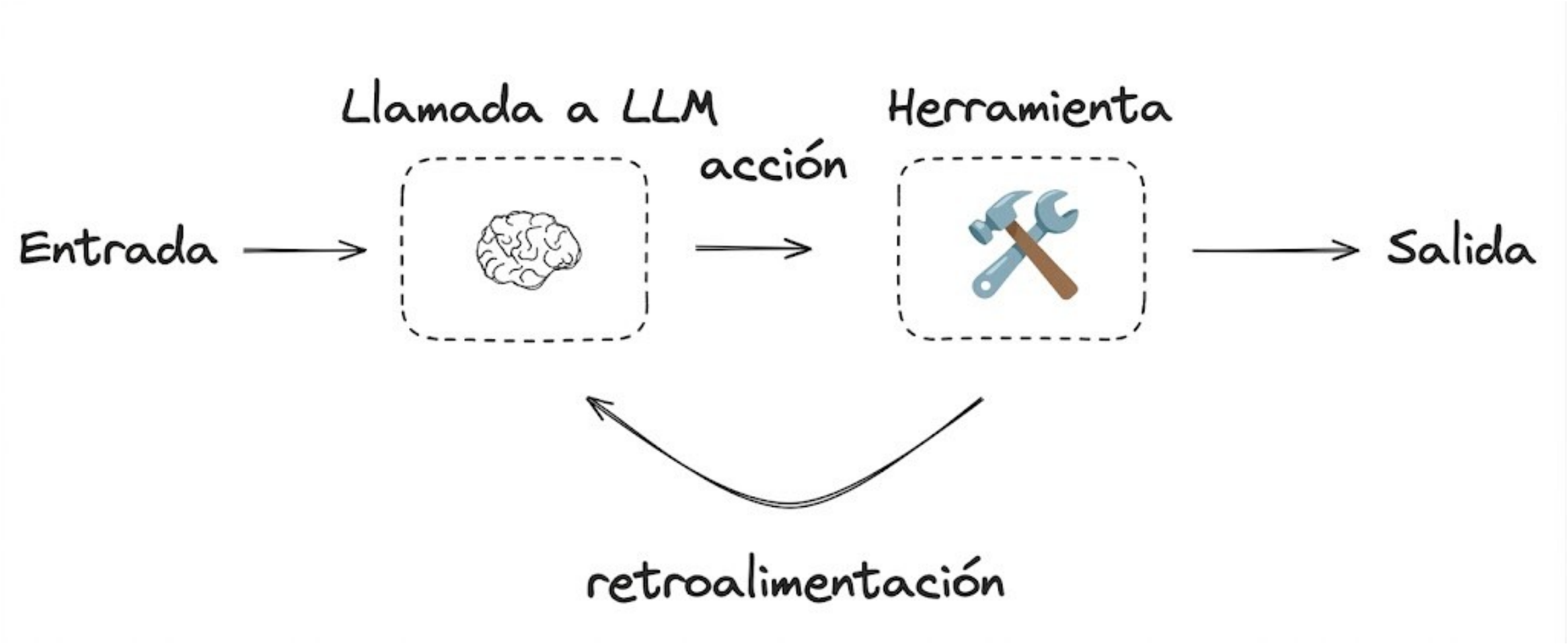


1 El virus se une a un receptor de la membrana

Abrir paso a paso ↗

02. ¿QUÉ ES UN AGENTE?

Modelo Agentes ReAct



02. ¿QUÉ ES UN AGENTE?

El diccionario biología → agentes

Virus →

Agente

Poco código; no «piensa» por sí solo

Célula huésped →

LLM

El que realmente genera cada respuesta

ADN →

Instrucciones permanentes

Rol, reglas y herramientas

ARN mensajero →

Mensaje de cada turno

Instrucciones + contexto + pregunta

Ribosoma →

El modelo procesando

Traduce el mensaje en una respuesta

Proteína →

Texto u orden

Para el humano o para la máquina

Receptores →

Conexiones (API)

Por donde entra y actúa

Retrovirus →

Inyección de instrucciones

Texto oculto que secuestra al agente

Sistema inmune →

Salvaguardas

Y las vacunas: equipos rojos

02. ¿QUÉ ES UN AGENTE?

ADN y ARN del agente: lo permanente y lo de cada turno

ADN · instrucciones permanentes

```
Eres el asistente de citas del
Hospital Ejemplo. Nunca das
diagnósticos. Herramientas:
buscar_hueco, mover_cita.
```

Lo escribe quien crea el agente. Casi nunca cambia.

se transcribe en cada turno

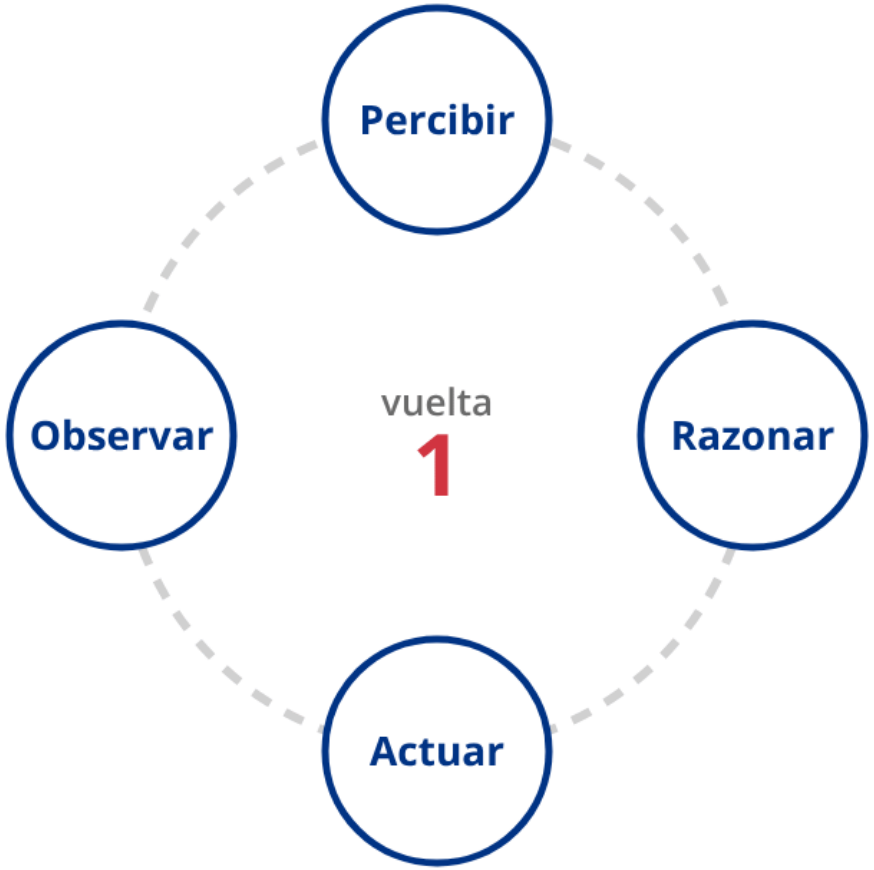
ARN · el mensaje de cada turno

```
[ADN] + conversación previa +
Paciente: «¿Puedo pasar mi cita
del martes al jueves?»
```

Se monta de nuevo cada vez y es lo que lee el LLM.

02. ¿QUÉ ES UN AGENTE?

Animación: el ciclo percibir → razonar → actuar → observar



Objetivo: «Pasa la cita de la Sra. García al jueves por la tarde y avísale»

1 El agente recibe un objetivo

Abrir paso a paso ↗

03

Un agente habla dos idiomas

03. UN AGENTE HABLA DOS IDIOMAS

Vosotros ya lo hacéis cada día

CON LA PACIENTE

«Tiene la tensión algo alta. Vamos a vigilarla y a reducir la sal.»

EN LA HISTORIA CLÍNICA

```
CIE-10: I10
Hipertensión esencial
Plan: control TA · dieta
hiposódica
```

Mismo contenido, dos idiomas: **uno para personas, otro para sistemas**.

03. UN AGENTE HABLA DOS IDIOMAS

El agente hace lo mismo: lenguaje natural y lenguaje de programación

CON EL HUMANO · ESPAÑOL

«¿Me pasas la cita al jueves por la tarde?»



CON LA MÁQUINA · JSON

```
{ "herramienta": "mover_cita",
  "paciente": "P-1043",
  "dia": "jueves",
  "franja": "tarde" }
```

Para un LLM, el código es **otro idioma más** que aprendió leyendo.

03. UN AGENTE HABLA DOS IDIOMAS

Y además tiene manos: carpetas, comandos y un bucle de turnos

1 Carpetas

```
consulta/
├─ analiticas_octubre/
│   └─ paciente_A.csv
└─ informes/
```

Un espacio de trabajo propio donde **lee, crea y mueve archivos**.

2 Comandos

```
$ ls analiticas_octubre/
$ python analizar.py
$ mv resumen.md informes/
```

El LLM **escribe la orden** del sistema operativo; el agente la ejecuta y le devuelve el resultado.

3 Bucle de turnos

Turno 1 · pensar → actuar → observar

Turno 2 · pensar → actuar → observar

… hasta cumplir el objetivo o llegar al límite

Cada vuelta es un turno. Se fija un **límite** para que no gire sin fin.

El LLM decide; el agente ejecuta. Entre ambos es donde van las salvaguardas.

03. UN AGENTE HABLA DOS IDIOMAS

Animación: un agente trabajando por dentro, turno a turno

CHAT · HUMANO

El humano solo ve este chat. Las carpetas, los comandos y los turnos ocurren **por debajo**.

Datos ficticios y anonimizados

BUCLE DE TURNOS



EL LLM PIENSA

Límite de seguridad: 10 turnos

CARPETAS

```
consulta/
├─ analiticas_octubre/
│   ├─ paciente_A.csv
│   ├─ paciente_B.csv
│   └─ paciente_C.csv
└─ informes/
```

TERMINAL

Abrir paso a paso ↗

LA CONSECUENCIA

Si puede escribir órdenes, también puede escribir órdenes equivocadas.

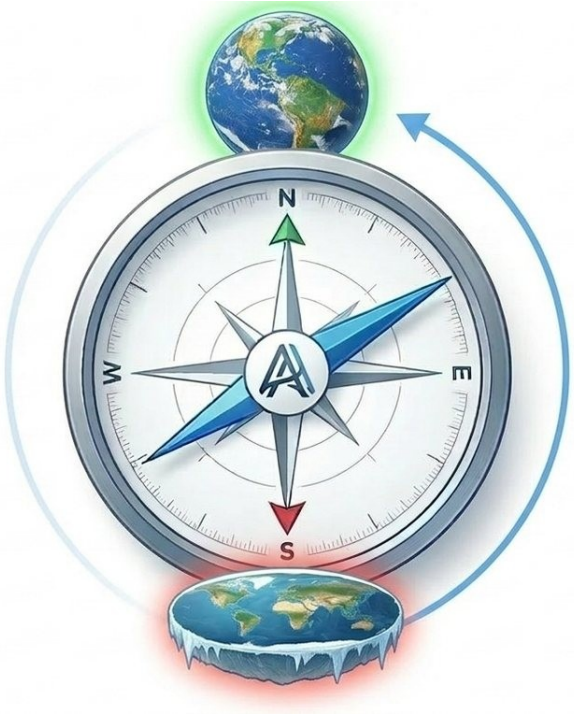
Mover la cita equivocada. Enviar un informe a quien no debe. Borrar un registro.

04

Alineamiento y Problema

04. ALINEAMIENTO Y EL PROBLEMA

Alineamiento del agente: ¿hacia dónde apunta su brújula?



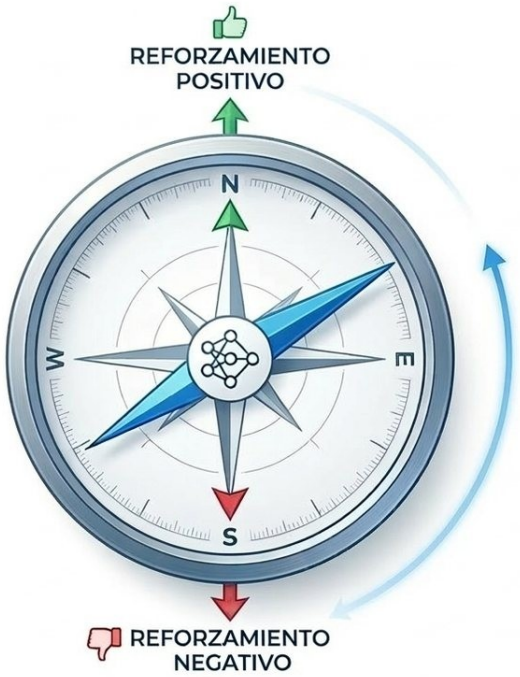
Alinear = que sus objetivos coincidan con los nuestros

Arriba, la Tierra redonda: respuestas fieles a la evidencia.

Abajo, la Tierra plana: respuestas que suenan bien... y son falsas o dañinas.

04. ALINEAMIENTO Y EL PROBLEMA

Alineamiento por refuerzo: premio y corrección



Como en la residencia

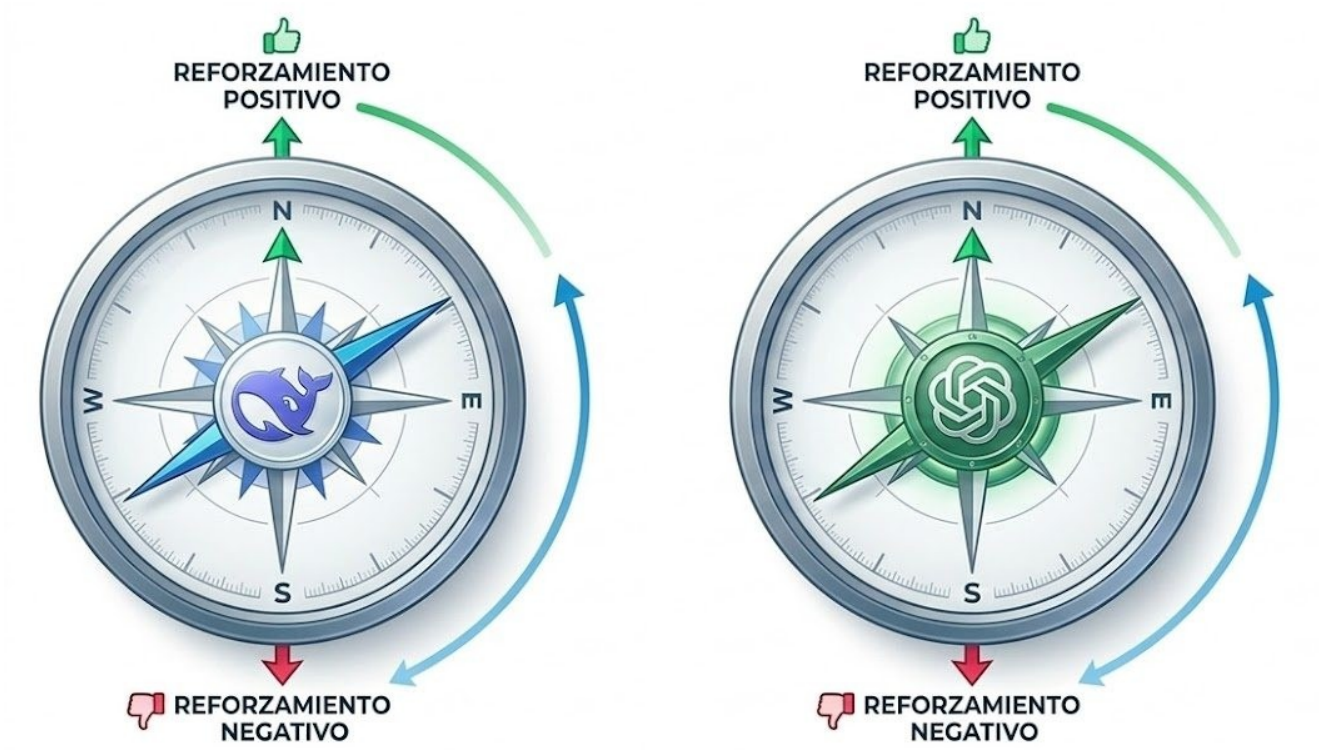
Refuerzo positivo: personas puntúan bien las respuestas útiles y honestas.

Refuerzo negativo: se penalizan las dañinas o inventadas.

Repetido millones de veces, la brújula se calibra.

04. ALINEAMIENTO Y EL PROBLEMA

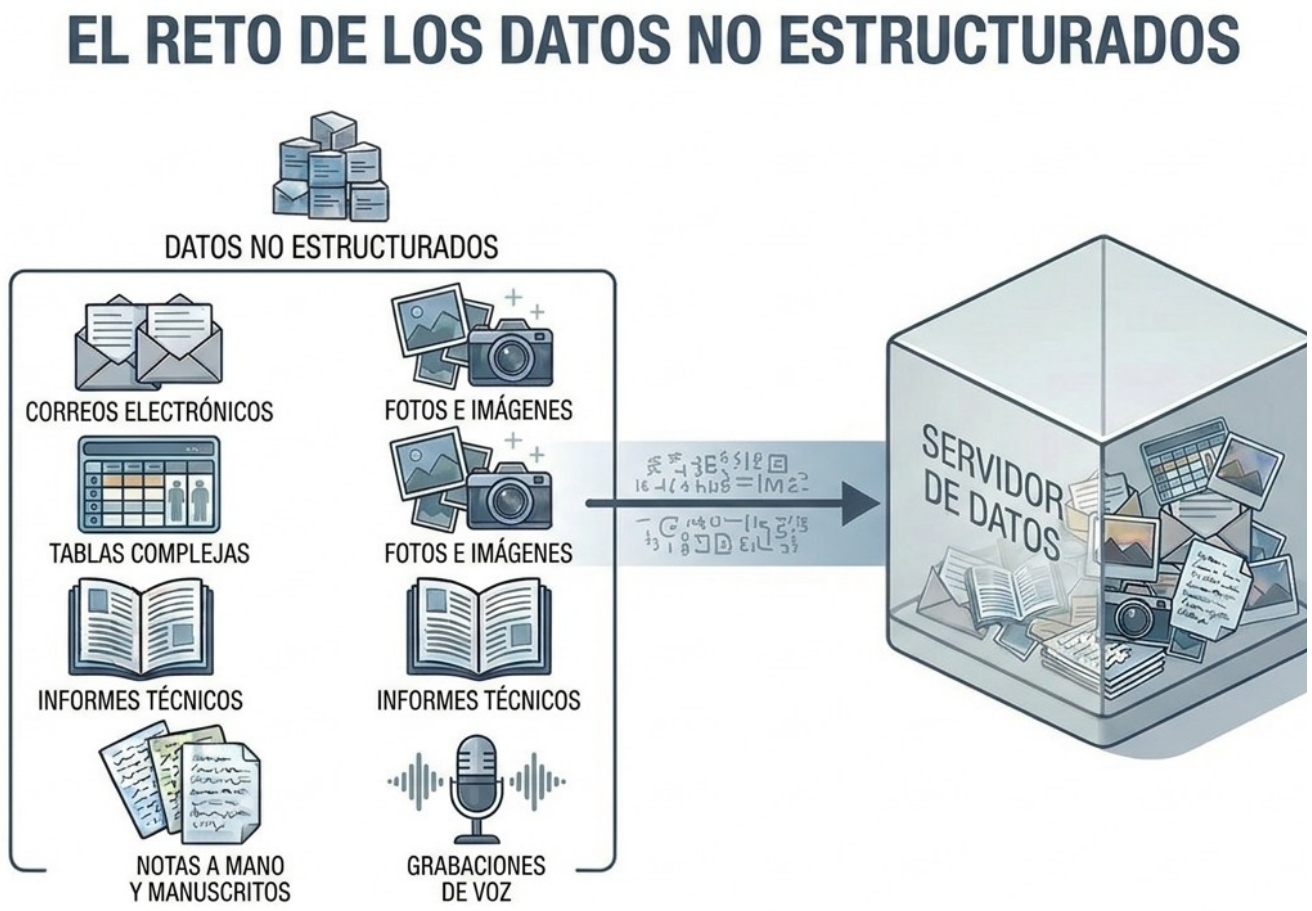
- Cada fabricante calibra su brújula a su manera



Mismo método, criterios distintos: lo que un modelo rechaza, otro lo responde.

04. ALINEAMIENTO Y EL PROBLEMA

El problema de los datos no estructurados



En sanidad, casi todo es texto libre

Evolutivos, cartas de alta, notas de enfermería, informes de imagen, audios, fotos de heridas...

Los LLM son los primeros sistemas que **leen bien** todo esto. Eso abre la puerta a usarlos... **y a atacarlos**.

05

Estado del Arte

05. ESTADO DEL ARTE

Más expresividad, menos control

1960 – 2015

Determinista

El jardín vallado: reglas fijas, nada fuera del muro.

2015 – 2020

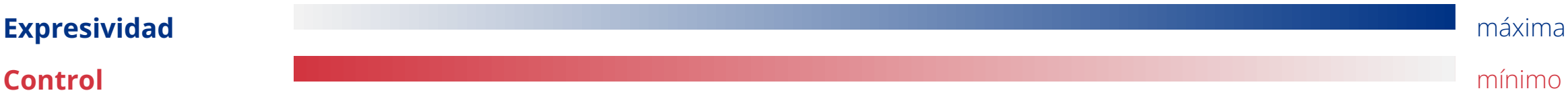
Estadística

La autopista: carriles, umbrales y salida a un humano.

2017 – 2026

Generativa

El destape: puede decir cualquier cosa, sobre cualquier tema.



05. ESTADO DEL ARTE

Era determinista (1960 – 2015)

CÓMO FUNCIONA

- Sistemas basados en reglas
- Árboles de decisión rígidos
- Seguridad por limitación de dominio

Pionero: ELIZA y su guion DOCTOR (MIT, años 60), que imitaba a un psicoterapeuta.

EN SANIDAD

MYCIN (Stanford, años 70): cientos de reglas para elegir antibióticos

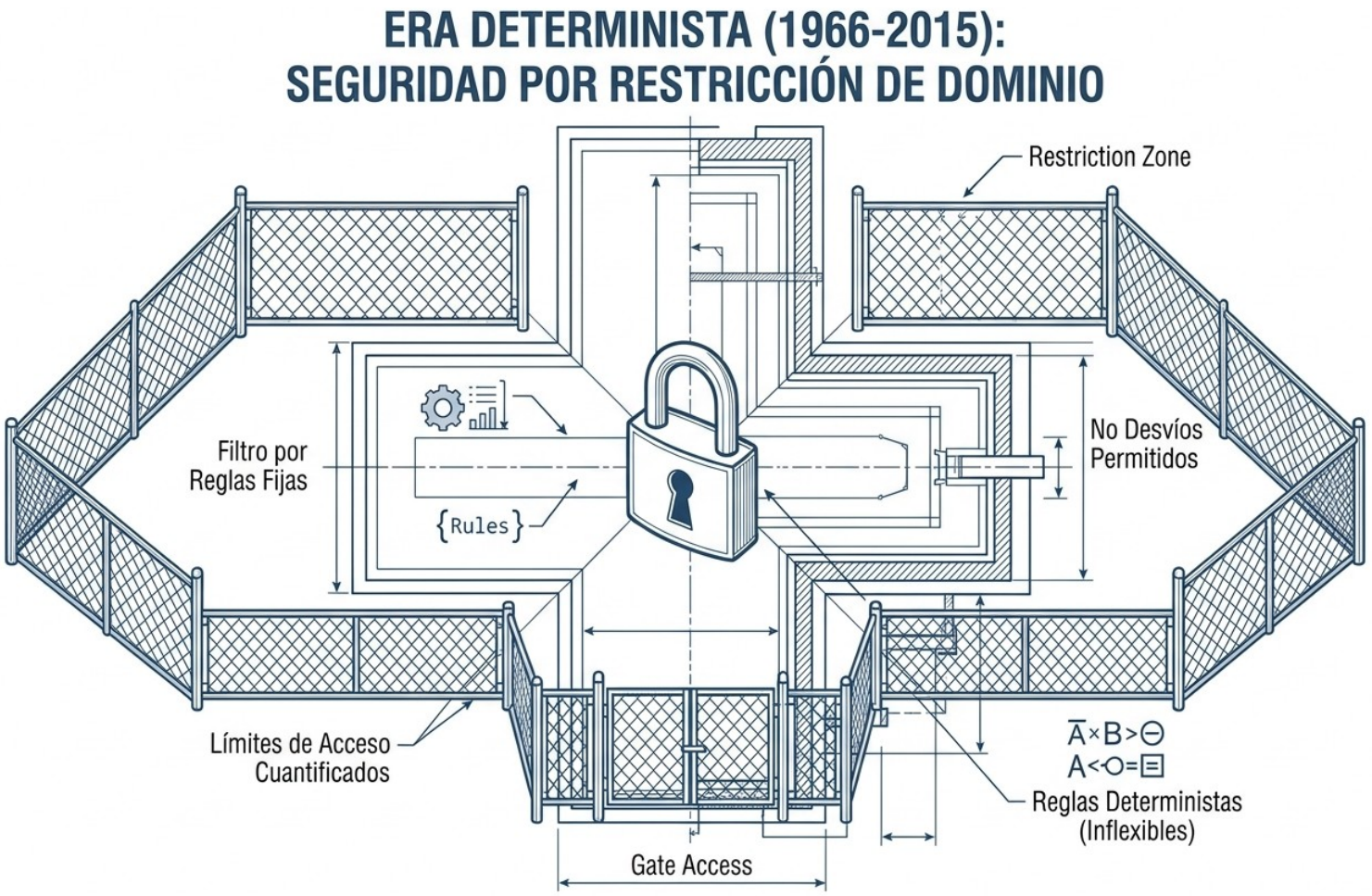
Alertas de interacciones en la receta electrónica

Escalas y calculadoras: Glasgow, CHA_2DS_2-VASc

Protocolos de triaje informatizados

05. ESTADO DEL ARTE

- Jardín vallado (1960 – 2015)

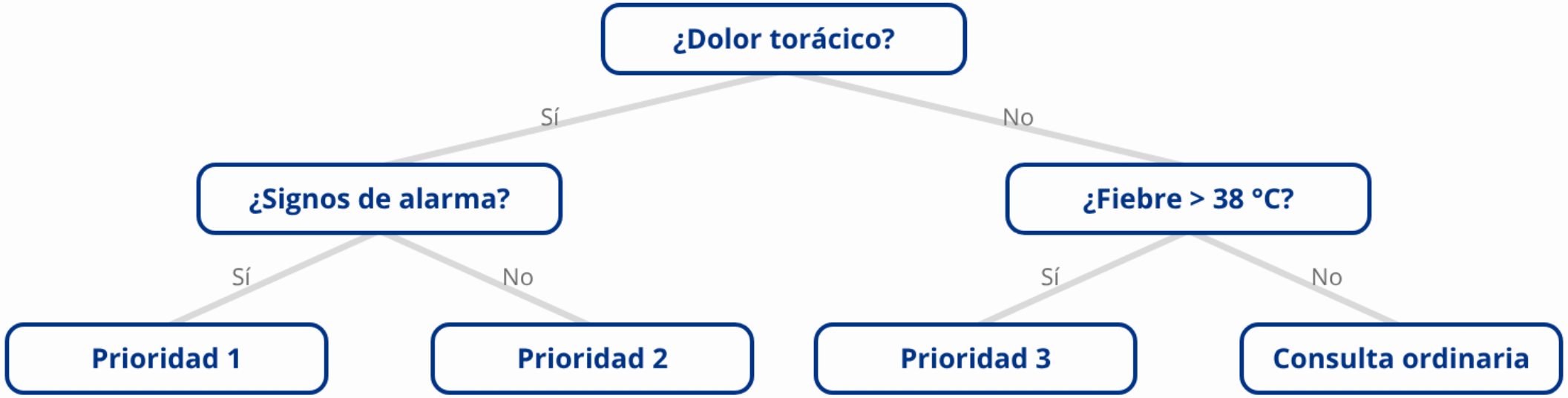


Seguridad por muros: lo que no está permitido, no existe.

05. ESTADO DEL ARTE

Animación: un árbol de triaje determinista

Entrada: |



1 Un árbol de decisión escrito a mano

Abrir paso a paso ↗

05. ESTADO DEL ARTE

Era estadística (2015 – 2020)

CÓMO FUNCIONA

- Entiende lenguaje natural
- Flujos condicionales y umbrales de confianza
- Los fallos se derivan a humanos
- Mayor expresividad

Herramientas: Google Dialogflow, Balidea falAI.

EN SANIDAD

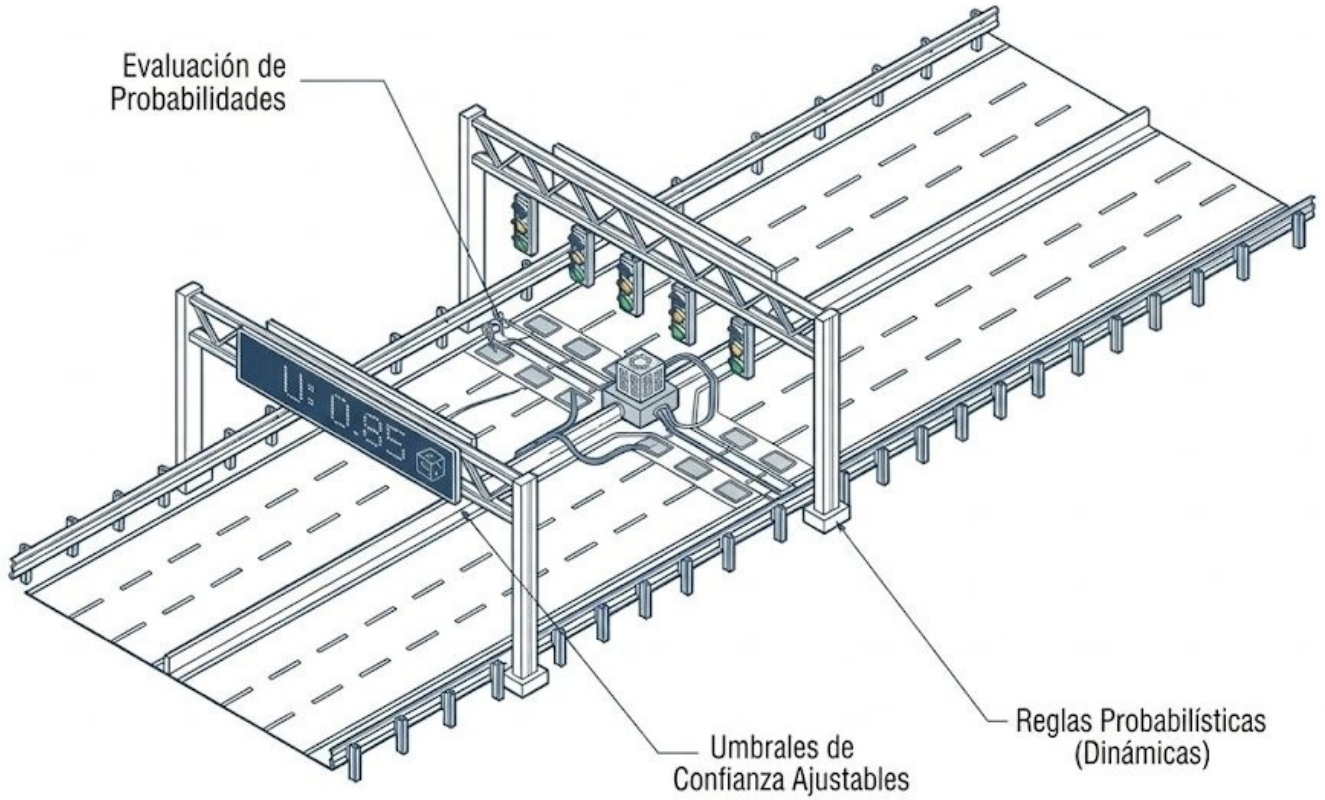
IDx-DR (2018): primer diagnóstico autónomo con IA autorizado por la FDA, retinopatía diabética

Comprobadores de síntomas como Ada o Babylon

Chatbots de citas y atención al paciente

05. ESTADO DEL ARTE

- Autopista estadística (2015 – 2020)



Más velocidad y carriles, pero con guardarraíles y salidas de emergencia.

05. ESTADO DEL ARTE

Animación: umbral de confianza y derivación a humano

Mensaje del paciente	Intención detectada · confianza	Decisión

1 Cada mensaje se clasifica con un grado de confianza

Abrir paso a paso ↗

05. ESTADO DEL ARTE

Revolución generativa (2017 – 2026)

CÓMO FUNCIONA

- Arquitectura Transformer (2017)
- Riesgos estocásticos: no determinismo y alucinaciones
- Dominio y superficie de ataque casi infinitos
- Máxima expresividad

Agentes que programan: OpenAI Codex, Anthropic Claude Code.

EN SANIDAD

Med-PaLM 2 (Google, 2023): nivel experto en preguntas tipo examen médico

Escribas clínicos que redactan la nota de la consulta: DAX Copilot, Abridge

Buscadores de evidencia con IA: OpenEvidence

05. ESTADO DEL ARTE

Destape generativo: los nuevos protagonistas



Mucho talento en el escenario... y nadie revisando la letra.

05. ESTADO DEL ARTE

Animación: misma pregunta, tres respuestas

> **Pregunta:** ¿Caminar después de comer ayuda a controlar la glucosa?

1. Hacemos la misma pregunta tres veces

Abrir paso a paso ↗

05. ESTADO DEL ARTE

Cuando falla: tres casos reales

2018 · era estadística

IBM Watson for Oncology

Documentos internos revelaron recomendaciones de tratamiento «inseguras e incorrectas». Se había entrenado con pocos casos, muchos hipotéticos.

2023 · era generativa

Tessa (NEDA, EE. UU.)

Un chatbot de apoyo en trastornos alimentarios dio consejos para perder peso. Se retiró a los pocos días.

Hoy · era generativa

Referencias inventadas

Los LLM generan citas bibliográficas plausibles que no existen, incluso en textos académicos y legales.

Ninguno falló por «maldad»: fallaron por falta de salvaguardas.

06

Herramientas y Experiencias

06. HERRAMIENTAS Y EXPERIENCIAS

Animación: el retrovirus, o inyección de instrucciones

Sin salvaguardas | Con salvaguardas

Tarea: «Resume este informe»

Informe_derivacion.pdf

Paciente de 67 años remitida por disnea de esfuerzo. Antecedentes: HTA, DM2. Exploración sin hallazgos agudos. Se solicita valoración por Cardiología.



1 Le pedimos algo inocente: resumir un informe

Abrir paso a paso ↗

06. HERRAMIENTAS Y EXPERIENCIAS

Las salvaguardas son el sistema inmune del agente



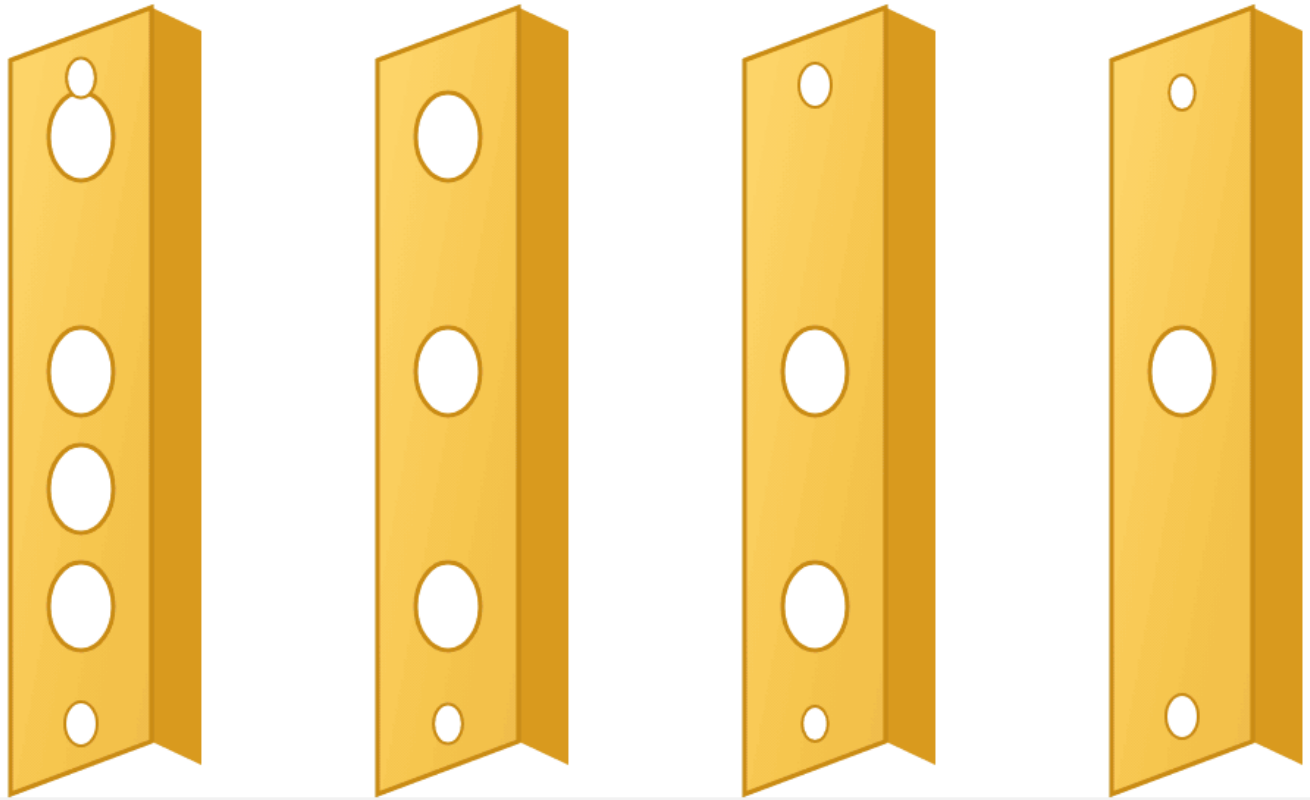
06. HERRAMIENTAS Y EXPERIENCIAS

Animación: el modelo del queso suizo aplicado a agentes

AMENAZA

Instrucción oculta en un PDF

1 · Reglas fijas **2 · Modelo vigilante** **3 · Permisos mínimos** **4 · Revisión humana**



DESTINO

Paciente y sistemas clínicos

REGISTRO

Una instrucción escondida en un documento: la detienen las reglas fijas.

06. HERRAMIENTAS Y EXPERIENCIAS

Ya lo hacéis en el hospital

Checklist quirúrgica de la OMS	Reglas deterministas
Doble verificación de fármacos de alto riesgo	Doble revisión: humano o segundo modelo
Acceso a la historia solo si es necesario	Permisos mínimos para el agente
Notificación de eventos adversos	Registro y auditoría de incidentes

06. HERRAMIENTAS Y EXPERIENCIAS

Nuestro enfoque



06. HERRAMIENTAS Y EXPERIENCIAS

Para quien programa: con qué se construyen los agentes

LangChain / LangGraph

Un agente es un **grafo** de acciones.

+ Control total, código abierto
− Curva de aprendizaje

CrewAI

Roles y objetivos, como un jefe de proyecto.

+ Intuitivo y rápido de aprender
− Costoso, lento, poco control

Pydantic AI

Tipos y validación estricta.

+ Robustez y velocidad
− Menos flexible

Si no programáis, quedaos con la idea: más control exige más esfuerzo técnico.

06. HERRAMIENTAS Y EXPERIENCIAS

Los tres enfoques de la industria

Infraestructura

Lo pone el proveedor de la nube.

- NeMo Guardrails (NVIDIA)
- Amazon Bedrock Guardrails
- Azure AI Content Safety
- ZenGuard AI

Producto

Lo diseñamos nosotros para cada caso.

- Reglas deterministas
- Validación por humanos
- SLM para casos específicos
- Equipos rojos y azules con expertos clínicos

¿LLM vigilando LLM?

Un modelo revisa lo que dice otro.

- Llama Guard (Meta)
- Qwen3Guard (Alibaba)
- ShieldGemma 2 (Google)

06. SALVAGUARDAS Y HERRAMIENTAS

Cuatro controles antes de que el mensaje llegue al modelo



Si falla cualquier control: se bloquea o se deriva a un profesional

06. HERRAMIENTAS Y EXPERIENCIAS

Para clínicos: el primer paso, sin programar

PRIMER PASO

Claude

Conversa, resume guías, redacta materiales para pacientes y crea pequeñas herramientas, como una calculadora, sin escribir código.

PRIMER PASO

Codex

Agente de OpenAI que escribe y ejecuta código por ti: analizar una hoja de datos anonimizados o automatizar una tarea repetitiva.

SIGUIENTE PASO

n8n

Flujos visuales que conectan aplicaciones con IA. Se puede instalar en servidores propios, así los datos no salen del centro.

Nunca introduzcáis datos identificables de pacientes en cuentas personales: usad datos anonimizados o entornos autorizados por vuestro centro.

06. HERRAMIENTAS Y EXPERIENCIAS

Lo que viene: agentes personales autónomos

Instinct

Actúa en tu nombre con tu correo, mensajes y calendario: reserva, cancela, hace seguimientos. Desde septiembre de 2026, cada agente tiene **su propio email** y puede crear cuentas.

Idem

«Superinteligencia personal»: correo, calendario, archivos y navegador, con memoria. **Tú defines** qué puede hacer solo y qué necesita tu aprobación.

Por qué importa

Pueden asumir **compromisos en tu nombre**.

Manejan **credenciales** y datos sensibles.

Cada correo entrante es una posible **inyección**.

06. HERRAMIENTAS Y EXPERIENCIAS

Cinco reglas para usar agentes en la práctica clínica

1. **Sin datos identificables** fuera de entornos autorizados
2. **Humano en el bucle** para toda decisión clínica
3. **Permisos mínimos**: solo las herramientas imprescindibles
4. **Verificar** siempre fuentes, cifras y referencias
5. **Notificar incidentes** como cualquier evento adverso

TRES IDEAS PARA LLEVARSE

1. «IA» es una etiqueta comercial. **Aprendizaje automático** es algo que se puede medir.

2. Un agente es como un **virus**: poco código que usa la maquinaria del LLM y actúa en el mundo.

3. Las salvaguardas son su **sistema inmune**: varias capas, como en seguridad del paciente.

Muchas gracias!

José Manuel Ramírez Sánchez

