【AI Summer Camp】

Arquitecturas LLM y Estrategia de Selección de Modelos

PABLO FERNÁNDEZ DÍAZ

MADRID INNOVATION [MIL] AI Network
Artificial Intelligence Association

ÍNDICE DE CONTENIDOS

PABLO FERNÁNDEZ DÍAZ

Data Tech Lead en Repsol

- ✓ **Formación:** Grado y Máster en Ingeniería Informática (UPM), Máster en Data Analytics & Project Management (IIT) y MBA (EAE).
- ✓ **Perfil:** Ingeniero informático especializado en analítica de datos, con experiencia desde DBA Oracle hasta Data Engineer, Data Analyst y Big Data Engineer.
- ✓ **Liderazgo:** Lidero equipos de Data e IA, desempeñando roles de Project Manager, Service Delivery Manager y Tech Lead.
- ✓ **Experiencia:** He trabajado para empresas como Repsol, British Telecom, Stratio o mLogica, y he participado en proyectos de consultoría para clientes como Banco Santander, Orange, USBank, Ferrovial, Mapfre, WiZink, etc.

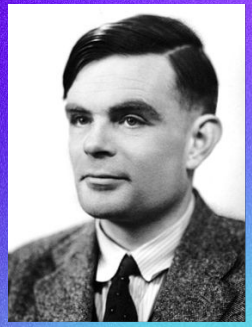
Antes de empezar…

Sección 1

¿DE DÓNDE VENIMOS?

Introducción conceptual y punto de partida

IA Generativa: ¿Por qué ahora y por qué es diferente?

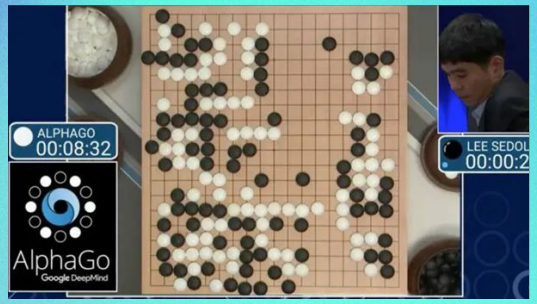


1950: Test de Turing

Fuerza Bruta

1997: Deep Blue (IBM) vs. Gary Kaspárov

Razonamiento Artificial



2016: AlphaGo (Google) vs. Lee Sedo

IA Generativa: ¿Por qué ahora y por qué es diferente?

1950: Test de Turing

Fuerza Bruta

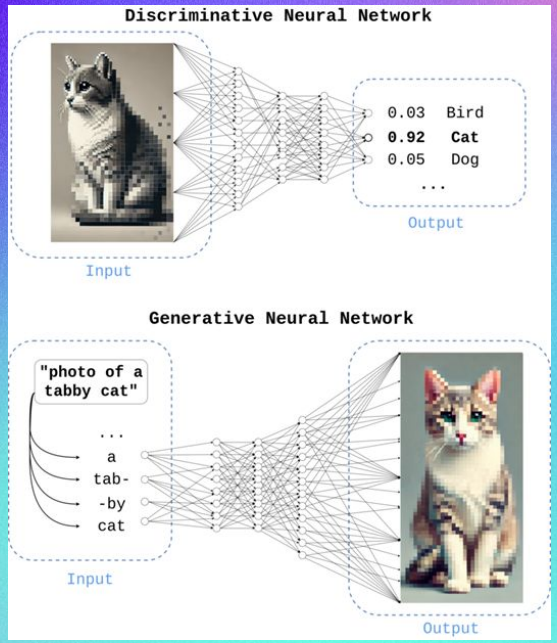
Razonamiento Artificial

Cambio de paradigma

2016: AlphaGo (Google) vs. Lee Sedo

IA Generativa: ¿Por qué ahora y por qué es diferente?

IA GEN: La inteligencia artificial generativa o IA generativa es un tipo de sistema de inteligencia artificial multimodal capaz de generar texto, imágenes u otros medios en respuesta a comandos.



→ IA Tradicional (Discriminativa/Predictiva):
Función: Clasifica, agrupa y predice datos existentes.
Ejemplo: Netflix recomendando una película o el banco detectando fraude.
Pregunta: "¿Es esto un gato?"

→ IA Generativa (Generativa):
Función: Crea contenido nuevo aprendiendo patrones de datos masivos.
Ejemplo: ChatGPT escribiendo un poema o Midjourney creando una foto.
Pregunta: "Dibuja un gato al estilo de Van Gogh."

IA Generativa: ¿Por qué ahora y por qué es diferente?

IA GEN: La inteligencia artificial generativa o IA generativa es un tipo de sistema de inteligencia artificial multimodal capaz de generar texto, imágenes u otros medios en respuesta a comandos.

→ IA Tradicional (Discriminativa/Predictiva):
Función: Clasifica, agrupa y predice datos existentes.
Ejemplo: Netflix recomendando una película o el banco detectando fraude.

→ IA Generativa (Generativa):
Función: Crea contenido nuevo aprendiendo patrones de datos masivos.
Ejemplo: [illegible] creando una foto.
[illegible] de Van Gogh."

Cambio de paradigma

La IA ya no solo lee, ahora escribe (escucha, dibuja, habla, compone, etc)

IA Generativa: ¿Por qué ahora y por qué es diferente?

ChatGPT

Tecnología de Propósito General (GPT) → No es un simple avance tecnológico, es una revolución global que afecta en multitud de sectores a nivel socio-económico.

📖 **Imprenta**: Democratizó el conocimiento.
💡 **Electricidad**: Revolucionó la energía y la producción.
🌐 **Internet**: Facilitó el acceso a la información.
🧠 **IA Generativa**: Transforma la capacidad cognitiva y de creación.

IA Generativa: ¿Por qué ahora y por qué es diferente?

ChatGPT → **Chat G**enerative **P**re-Trained **T**ransformer

Tecnología de Propósito General (GPT) → No es un simple avance tecnológico, es una revolución global que afecta en multitud de sectores a nivel socio-económico.

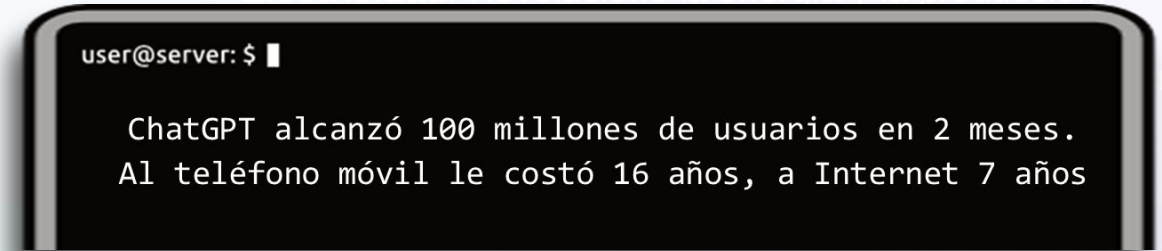
📖 **Imprenta**: Democratizó el conocimiento.
💡 **Electricidad**: Revolucionó la energía y la producción.
🌐 **Internet**: Facilitó el acceso a la información.
🧠 **IA Generativa**: Transforma la capacidad cognitiva y de creación.

IA Generativa: ¿Por qué ahora y por qué es diferente?

¿Cómo sabemos que la IA GEN está teniendo tal impacto?

Tecnología de Propósito General (GPT): No es un simple avance tecnológico, es una revolución global que está transformando todos los sectores a nivel socio-económico.

- **Imprenta**: Democratizó el conocimiento.
- **Electricidad**: Revolucionó la energía y la producción.



Sección 2

CONCEPTOS TÉCNICOS BÁSICOS

Repaso de conceptos y terminología básica sobre IA Generativa

【AI Summer Camp】

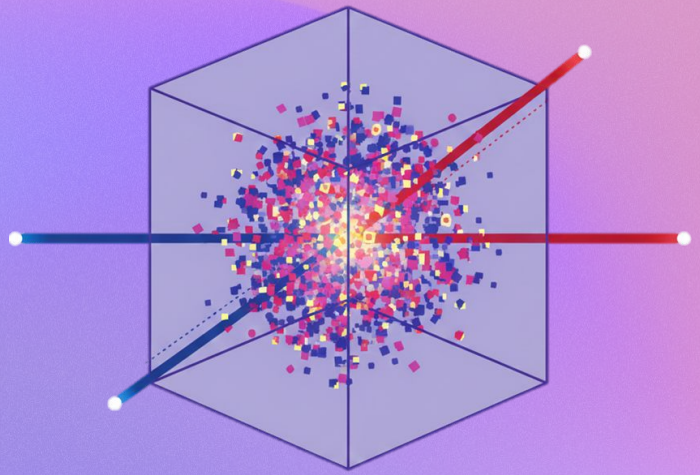
¿Qué es la Inteligencia Artificial Generativa?

La IA Generativa **NO** "piensa" como un humano; **predice**.

Imagínala como un sistema de "autocompletar " extremadamente sofisticado.

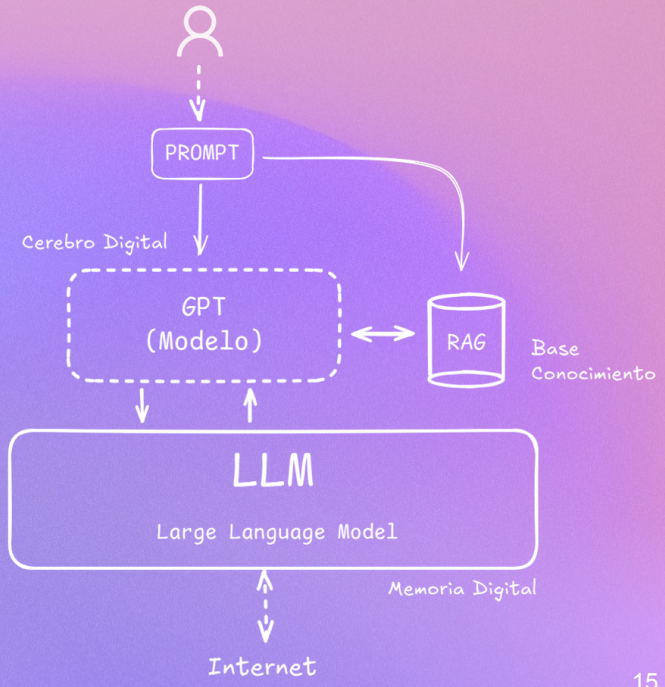
Ha leído casi todo lo publicado en internet para aprender **patrones** y **relaciones entre palabras**.

Cuando le preguntas algo, **calcula probabilísticamente** qué palabra debería ir a continuación para construir una respuesta coherente.

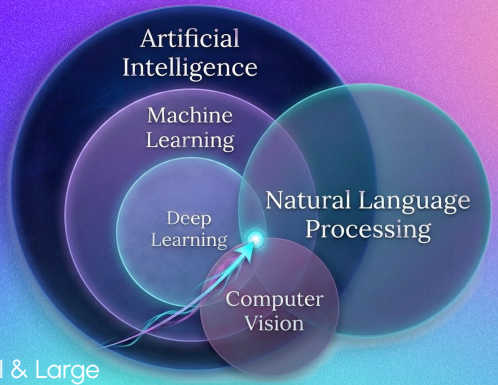


¿Qué es la Inteligencia Artificial Generativa?

- **Prompt (La Instrucción)**: El texto que tú escribes.
- **El Modelo (El Motor)**: Es un **cerebro congelado** en el momento de su entrenamiento.
- **RAG (Tu Biblioteca Privada)**: Técnica que conecta el modelo a tus **datos** (guardarraíles).
- **LLM (Large Language Model)**: Un tipo de modelo experto en lenguaje humano. Su especialidad es entender texto y generar respuestas fluidas.
- **Contexto (La Memoria)**: El espacio limitado que tiene la IA para "recordar" la conversación actual.
- **Temperatura (Creatividad)**: Baja = respuestas lógicas y precisas. Alta = respuestas variadas y creativas.

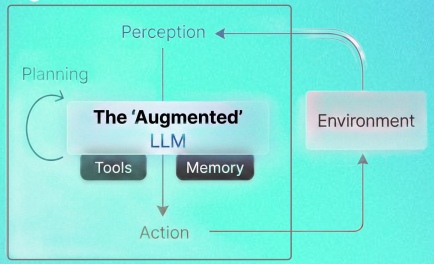


¿Qué es la Inteligencia Artificial Generativa?



Generative AI & Large Language Models:

Agente:



LLM (Large Language Model)

Un motor estadístico masivo entrenado para **comprender y generar lenguaje** con una fluidez indistinguible de la humana.

¿Cómo funciona?

- **Tokenización:** Trocea el texto en pequeñas unidades (tokens) que procesa numéricamente.
- **Atención:** Identifica qué palabras son más relevantes entre sí en una frase.
- **Predicción:** Calcula la **probabilidad estadística** de la siguiente palabra basándose en el contexto previo.

Los LLMs actúan como el "cerebro" central de la IA Generativa.

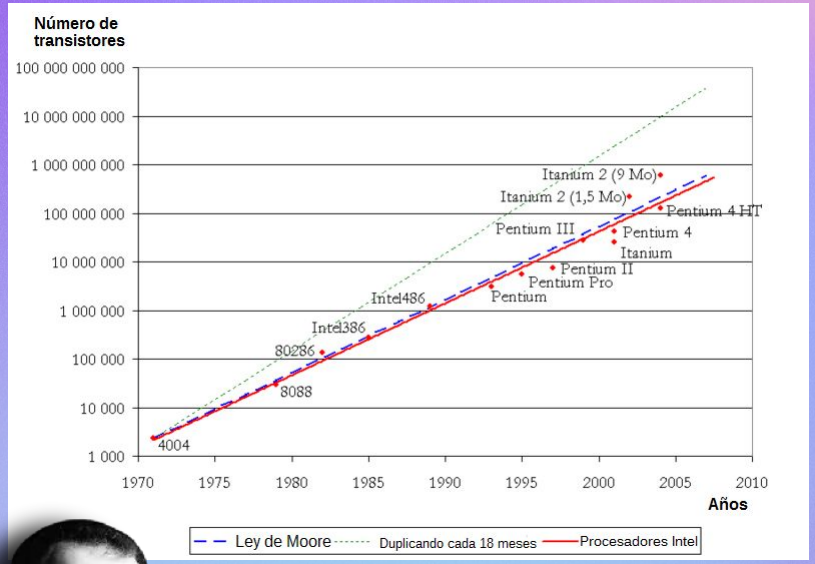
Sección 3

LA DEMANDA DEL FUTURO

Cuando la IA GEN rompe los principios tecnológicos

【AI Summer Camp】

Fundamentos del Crecimiento: La Ley de Moore



El Axioma Original

Observación de Gordon Moore (1965): El número de transistores en un microchip se **duplica aproximadamente cada dos años**, mientras que el costo se reduce a la mitad.

Impacto en la IA Primitiva

Este crecimiento exponencial permitió pasar de cálculos simples a la capacidad de procesar **redes neuronales profundas** masivas mediante fuerza bruta computacional.

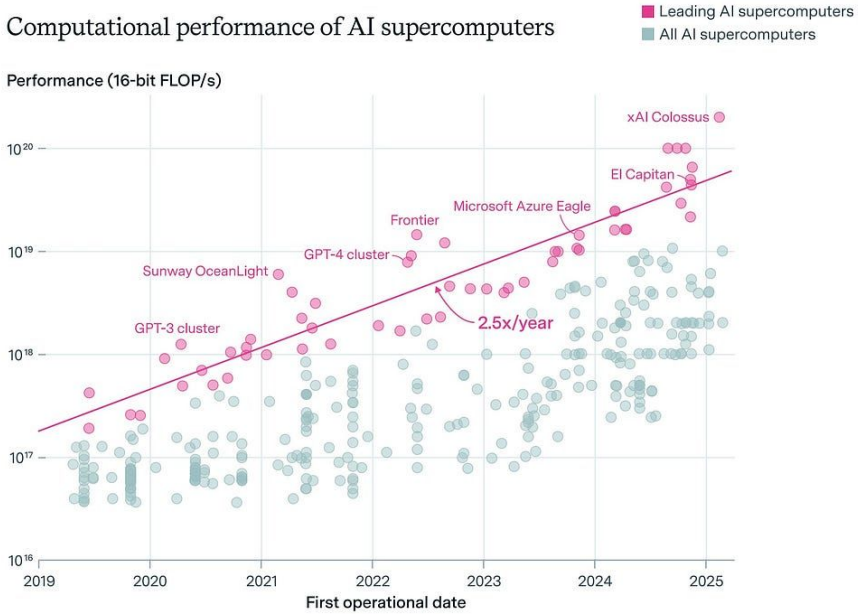
Hacia el Límite Físico

A medida que los transistores alcanzan tamaños atómicos, la Ley de Moore enfrenta **barreras térmicas y cuánticas**, obligando a la IA a buscar eficiencia más allá del hardware puro.

Gordon E. Moore

"La era de la potencia gratuita está terminando; la era de la arquitectura inteligente está comenzando." *Jensen Huang, CEO de NVIDIA*

La Explosión del Dato: El Motor de la Era IA



La Era del Big Data e IoT:

El flujo masivo de información superó las previsiones de la **Ley de Moore**, impulsando un abaratamiento radical del almacenamiento y nuevas necesidades de procesamiento.

El Acelerador de la IA Generativa (IA GEN):

La IA Generativa ha impulsado un aumento exponencial en la demanda de recursos de computación.

- **Entrenamiento de Modelos punteros:** Desarrollar los modelos de IA más avanzados (modelos frontera) exige:
  - **Datasets de Escala Planetaria:** Conjuntos de datos masivos a una escala global.
  - **Infraestructura de GPUs sin Precedentes:** Una capacidad de procesamiento basada en Unidades de Procesamiento Gráfico (GPUs) nunca visto antes.

[The Unprecedented Rise of AI Supercomputers: Powering the Future, But at What Cost?]

[Attention Is All You Need]

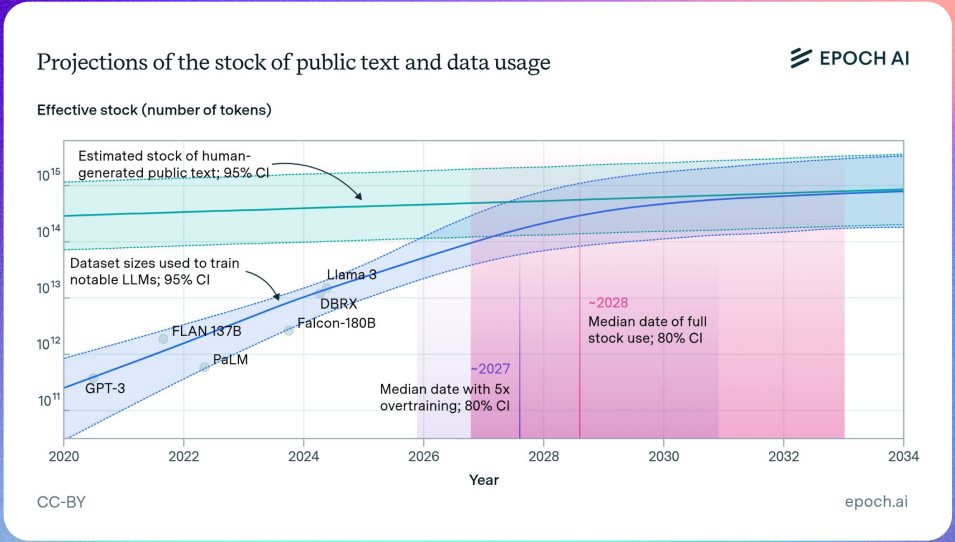
"Ya no solo recolectamos datos; estamos generando sintéticamente el combustible del mañana para superar los límites físicos del hardware". *Iya Sutskever, cofundador de OpenAI*

MADRID INNOVATION | MIL | AI Network

【AI Summer Camp】

El Techo del Crecimiento: Leyes de Escala

Cuando el avance es tan rápido que aparecen nuevos límites



[Will we run out of data? Limits of LLM scaling based on human-generated data]

Durante años, la regla fue:

Más datos + Más GPUs = Más inteligencia.

En 2025/2026, estamos viendo un **plateau** en el escalado masivo: la ganancia por cada billón de parámetros extra es cada vez menor.

La nueva frontera: Calidad de datos sintéticos y razonamiento post-entrenamiento (Chain of Thought).

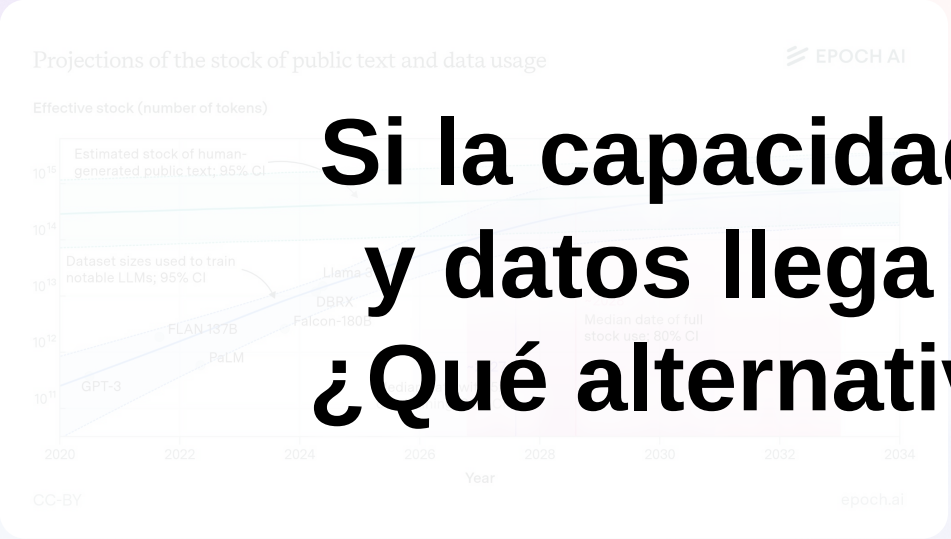
¿Hemos llegado al límite de la inteligencia artificial o solo al límite de la fuerza bruta?

MADRID INNOVATION | MIL | AI Network

【AI Summer Camp】

El Techo del Crecimiento: Leyes de Escala

Cuando el avance es tan rápido que aparecen nuevos límites



Will we run out of data? Limits of LLM scaling based on human-generated data

Durante años, la regla fue:

Más datos + Más GPUs = Más inteligencia.

Estamos viendo un **plateau** en el escalado masivo: la ganancia por cada billón de parámetros es menor.

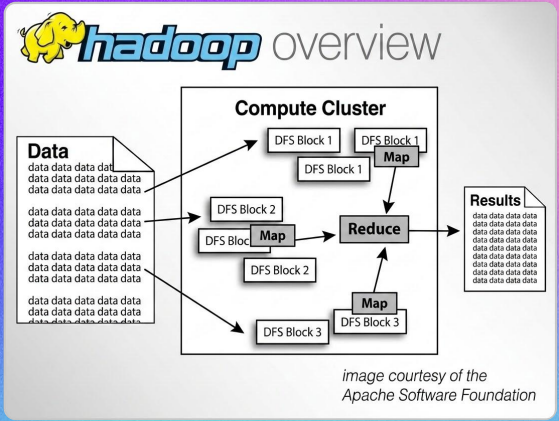
Nuevas soluciones: datos sintéticos y razonamiento post-entrenamiento (Chain of Thought).

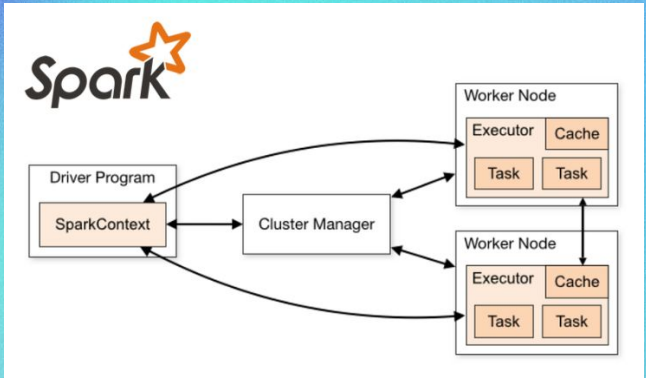
¿Hemos llegado al límite de la inteligencia artificial o solo al límite de la fuerza bruta?

Si la capacidad de cómputo y datos llega a su límite… ¿Qué alternativas tenemos?

MADRID INNOVATION 【MIL】 AI Network

【AI Summer Camp】





Divide y vencerás
Cómputo Distribuido

Evitando reinventar la rueda, conociendo de dónde venimos

Tecnologías Big Data como **Hadoop** (2006) y **Spark** (2014) ya aplicaban el principio de dividir tareas complejas y muy grandes en unidades de trabajo más pequeñas para ser procesadas por sus *workers* o nodos.

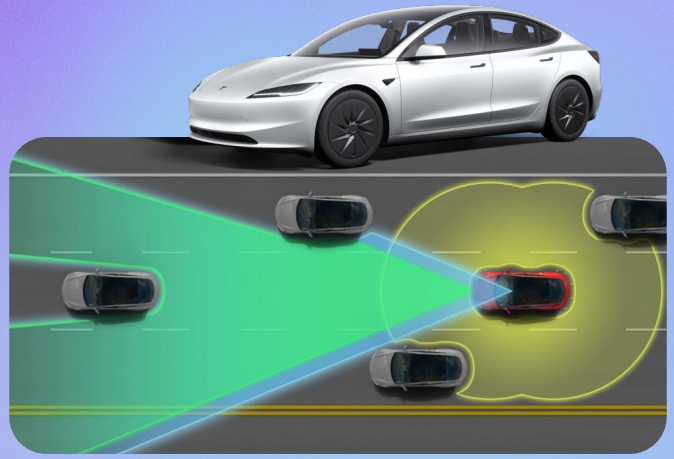
- ✓ **Eficiencia:** Procesamiento paralelo para completar trabajos más rápido.
- ✓ **Escalabilidad:** Añadir más *workers* (nodos) para manejar mayores volúmenes de datos o tareas más grandes.
- ✓ **Tolerancia a fallos:** Si un *worker* falla, la tarea puede reasignarse a otro.
- ✓ **Gestión de recursos:** Mejor utilización de los recursos del clúster.

Edge Computing como resultado del Cómputo Distribuido

Nuestros coches se convierten en centros de cómputo rodantes

Cómputo Distribuido → Enfoque Tecnológico

Edge Computing → Tipo de Arquitectura Distribuida



El Desafío:

Un coche autónomo genera 1GB de datos por segundo. Enviar eso a la nube para decidir si frenar ante un peatón causaría un accidente por latencia (retraso).

La solución:

Edge Computing: arquitectura de red distribuida que procesa y analiza los datos cerca de su fuente de origen, sin enviarlos a un servidor centralizado o a la nube.

Aprovechamiento del Cloud:

Gracias al uso del Cloud Computing cada Tesla circulando en cualquier ciudad sirve como un centro de cómputo con cientos de sensores recogiendo datos que se utilizan para mejorar todos los Testa del mundo.

Edge Computing como resultado del Cómputo Distribuido

Nuestros coches se convierten en centros de cómputo rodantes

Cómputo Distribuido: Enfoque Tecnológico

Edge Computing: Tipo de Arquitectura Distribuida

Un coche autónomo genera 1GB de datos por segundo. Enviar eso a la nube para decidir en fracciones de segundo cómo actuar es [illegible].

Edge Computing: arquitectura de red distribuida que procesa y analiza los datos cerca de su fuente de origen, sin enviarlos a un servidor centralizado o a la nube.

Aprovechamiento del Cloud:

Gracias al uso del Cloud Computing cada Tesla circulando en cualquier ciudad sirve como un centro de cómputo con cientos de sensores recogiendo datos que se utilizan para mejorar todos los Tesla del mundo.

¿Y cómo extrapolamos este enfoque al mundo IA GEN?

Revolución SLM: Inteligencia On-Device

¿Qué es un SLM?

Modelos de Lenguaje Pequeños (Small Language Models) diseñados para eficiencia extrema, optimizados para ejecutarse localmente sin depender de la nube.

Beneficios Clave

- **Privacidad Total:** Procesamiento local; los datos nunca abandonan el dispositivo.
- **Soberanía y Coste:** Eliminación de latencia y facturas de API por uso recurrente.
- **Eficiencia Energética:** Menor huella de carbono y optimización de hardware específico.

Phi-3 Mini *compite con modelos 10 veces más grandes gracias al entrenamiento con datos sintéticos de "calidad libro de texto", una estrategia de Microsoft que prioriza la curación de datos sobre el tamaño del modelo.*

Microsoft Research: Phi-3 Technical Report and Topology

Sección 4

EL MAPA DEL ECOSISTEMA IA 2026

De los modelos frontera a la inteligencia local

【AI Summer Camp】

Taxonomía de Modelos: Especialización

Frontier LLMs

Descripción:

Modelos de escala masiva (>100B parámetros) con capacidades de razonamiento extremo y planificación compleja.

Caso de Uso:

Resolución de problemas científicos, asesoría legal compleja y arquitectura de software.

Pros y Contras:

- + Máxima inteligencia y creatividad.
- - Alta latencia, coste elevado y dependencia cloud.

Modelos:

GPT-5, Claude 4, o1.

SLMs (Small)

Descripción:

Modelos eficientes (<10B) optimizados para ejecución local y privacidad total.

Caso de Uso:

Dispositivos móviles, asistentes offline y procesamiento de datos sensibles.

Pros y Contras:

- + Latencia cero, soberanía del dato y bajo coste.
- - Menor capacidad de razonamiento general.

Modelos:

Phi-4, Gemma 3, Llama 3b.

MoE (Mixture)

Descripción:

Arquitectura modular que activa solo los "expertos" necesarios para cada tarea.

Caso de Uso:

Chatbots multilingües y sistemas que requieren alta eficiencia en inferencia masiva.

Pros y Contras:

- + Eficiencia energética y velocidad de respuesta.
- - Mayor complejidad técnica en el despliegue.

Modelos:

Mistral Large, DeepSeek-V3.

Taxonomía de Modelos: Especialización

Frontier LLMs

Descripción:

Modelos de escala masiva (>100 parámetros) con capacidades de razonamiento extremo y planificación compleja.

Caso de Uso:

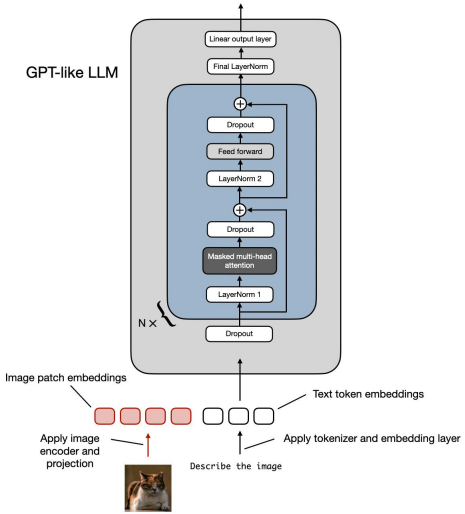
Resolución de problemas científicos, asesoría legal compleja y arquitectura de software.

Pros y Contras:

- + Máxima inteligencia y creatividad.
- - Alta latencia, coste elevado y dependencia cloud.

Modelos:

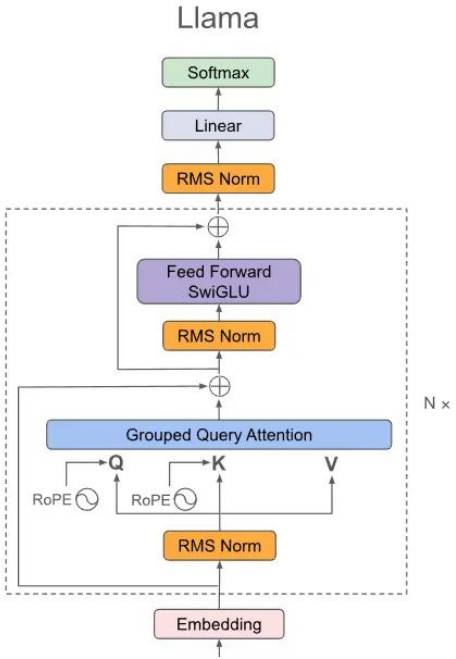
GPT-5, Claude 4, o1.



GPT-5

Claude

Taxonomía de Modelos: Especialización



SLMs (Small)

Descripción:

Modelos eficientes (<10B) optimizados para ejecución local y privacidad total.

Caso de Uso:

Dispositivos móviles, asistentes offline y procesamiento de datos sensibles.

Pros y Contras:

- + Latencia cero, soberanía del dato y bajo coste.
- - Menor capacidad de razonamiento general.

Modelos:

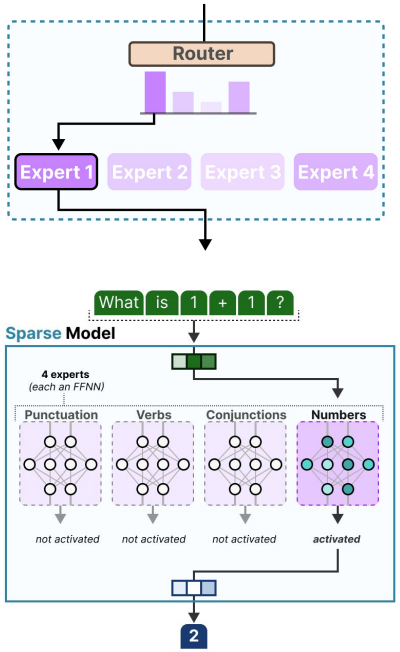
Phi-4, Gemma 3, Llama 3b.

Phi-4

Gemma 3

Llama 3

Taxonomía de Modelos: Especialización



Mistral Large

DeepSeek V3

MoE (Mixture)

Descripción:

Arquitectura modular que activa solo los "expertos" necesarios para cada tarea.

Caso de Uso:

Chatbots multilingües y sistemas que requieren alta eficiencia en inferencia masiva.

Pros y Contras:

- + Eficiencia energética y velocidad de respuesta.
- - Mayor complejidad técnica en el despliegue.

Modelos:

Mistral Large, DeepSeek-V3.

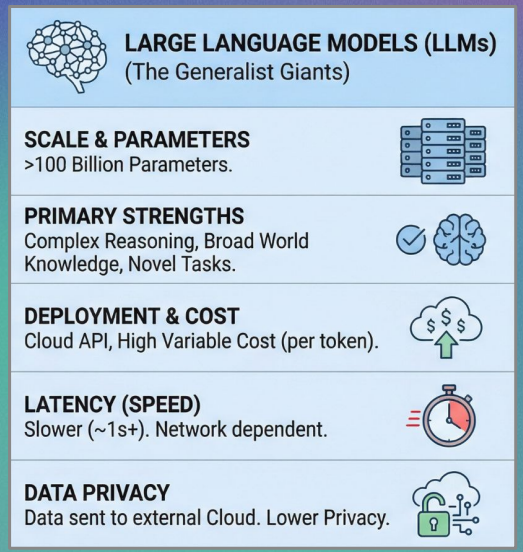
LLMs vs SLMs: Análisis Detallado

Large Language Models (LLMs)

Escala: >100B parámetros

Arquitectura: Transformers masivos (Cloud-native)

Latencia / Coste: Alta latencia, coste elevado por token

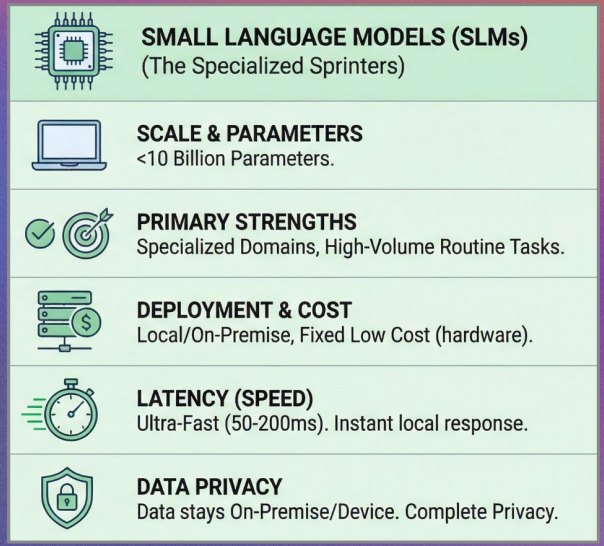


Small Language Models (SLMs)

Escala: <10B parámetros

Arquitectura: Optimizada para ejecución On-Device

Latencia / Coste: Latencia casi cero, bajo coste operativo



LLMs vs SLMs: Análisis Detallado

Large Language Models (LLMs)

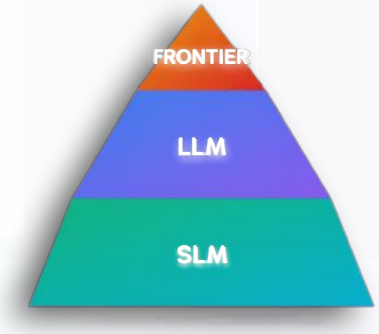
Los usaremos cuando…

- ✓ Problemas Razonamiento Complejo
- ✓ Comprensión Contextos Extremos
- ✓ Variedad altísima de tareas

Small Language Models (SLMs)

Los usaremos cuando…

- ✓ Requerimiento Real-Time
- ✓ Ejecución en “local” → Privacidad
- ✓ Limitaciones hardware o presupuestarias
- ✓ Tareas muy concretas



Language Models (LLM/SLM) frente a Mixture of Experts (MoE)

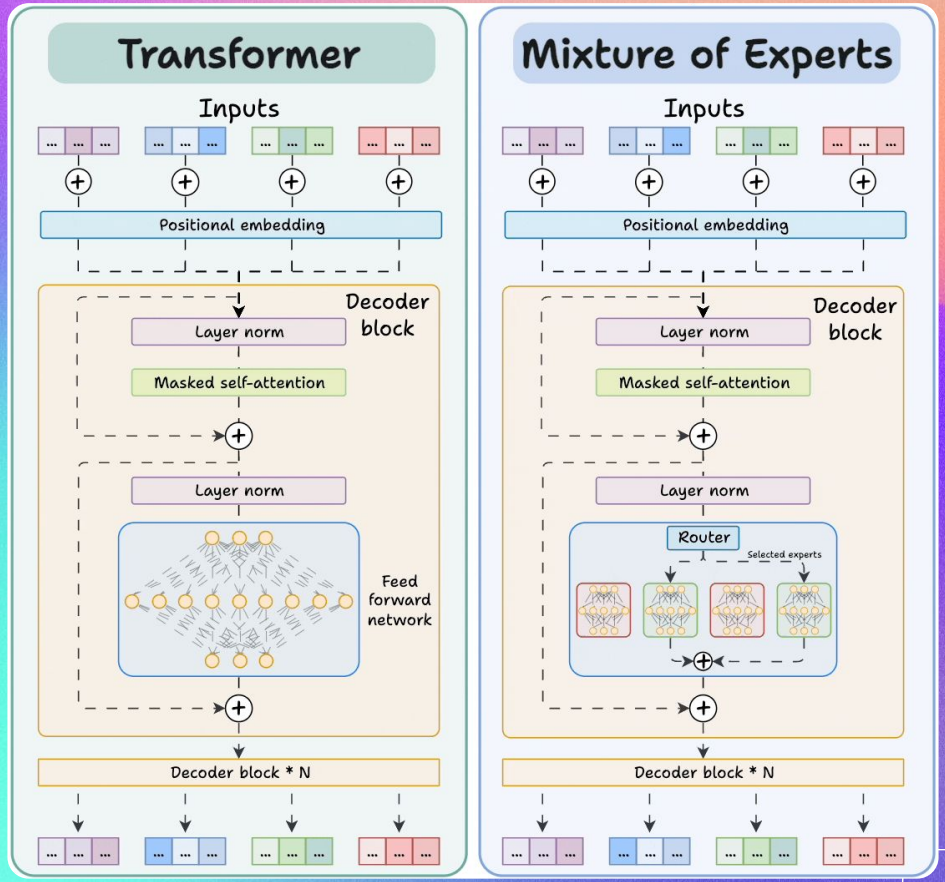
Modelos Densos (Arquitectura típica)

Language Models (LMs)

Arquitectura: Monolítica (Caja única)

Funcionamiento: Activa el 100% de los parámetros para cada token generado.

Ventaja: Coherencia global extrema.



Mixture of Experts (MoE)

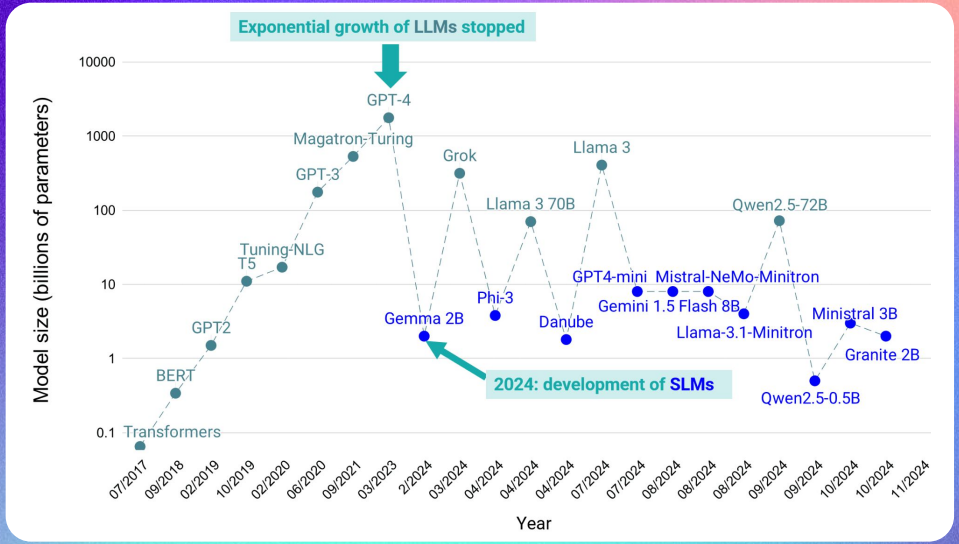
Arquitectura: Modular (Router + Expertos)

Funcionamiento: Un router selecciona solo 2-N "expertos" por token.

Ventaja: Alta eficiencia; inteligencia de LLM a coste de SLM.

Cambio de paradigma (2024)

Cuando las versiones “Mini” absorben el 80% de las tareas



The Rise of Small Language Models (SLMs) in AI

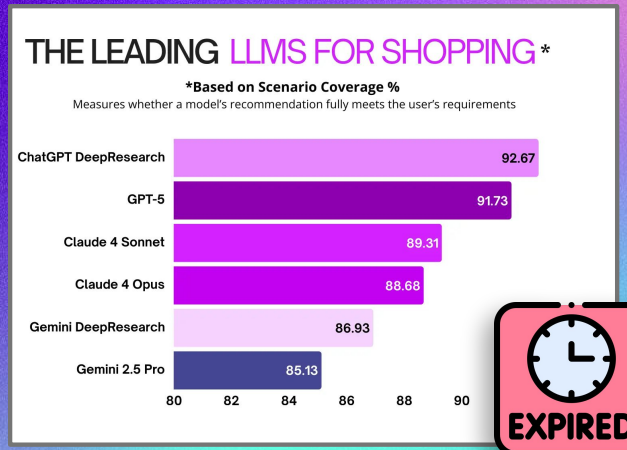
El Cambio de Paradigma

- **Fin de los modelos gigantes:** Usar IA masiva para tareas comunes ya no es eficiente ni rentable.
- **El nuevo estándar:** Modelos pequeños (SLMs), compactos y especializados.

Beneficios Clave

- **Eficiencia:** Hasta 30 veces más baratos y mucho más rápidos.
- **Privacidad:** Operan de forma local en dispositivos o servidores propios.
- **Especialización:** Ecosistemas de varios SLMs expertos en tareas únicas.

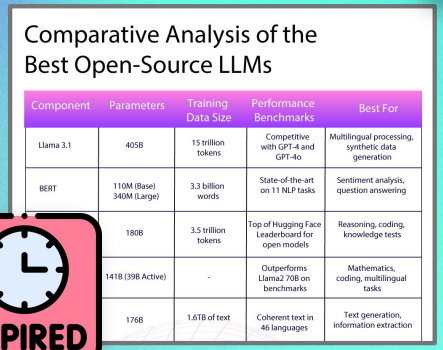
Llegados a este punto… ¿Qué modelo elegimos?



El nuevo desafío en los proyectos IT (en la era IA):

Elegir el modelo **correcto** para mi necesidad,

O al menos el modelo **menos malo**, en mi contexto



¿Nos podemos fiar de las comparativas?

- ¿Intereses comerciales?
- Información desactualizada
- Soluciones muy focalizadas

【AI Summer Camp】

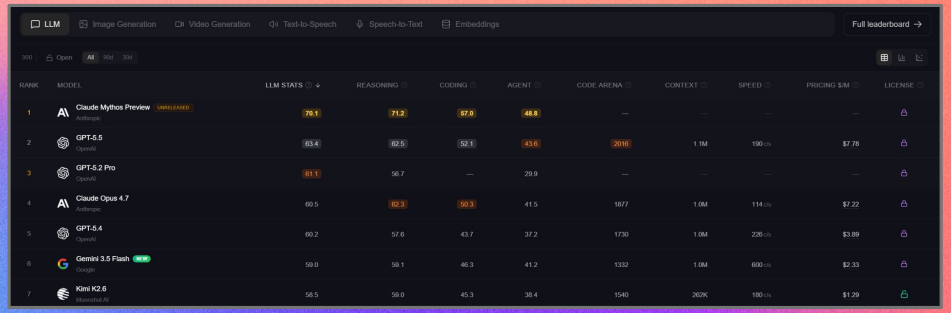
Llegados a este punto… ¿Qué modelo elegimos?

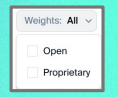
Prioriza Leaderboards dinámicos frente a Benchmarks estáticos

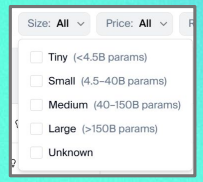
Artificial Analysis

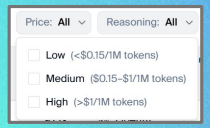
LLM STATS

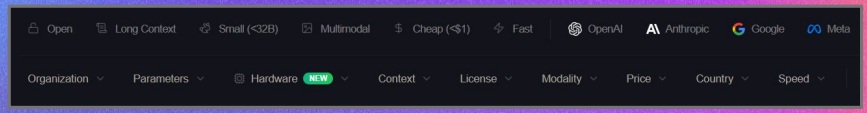
Model	Features: Context Window	Creator	Intelligence: Artificial Analysis Intelligence Index	Price: Blended USD/1M Tokens	Speed: Median Tokens/s	Latency: Latency First Chunk (s)	End-to-End Response Time: Total Response (s)	Further Analysis
GPT-5.5 (xhigh)	922k	OpenAI	60	$4.35	65	92.53	100.20	Model ↗ Providers ↗
GPT-5.5 (high)	922k	OpenAI	59	$4.35	69	28.11	35.39	Model ↗ Providers ↗
Claude Opus 4.7 (max)	1M	Anthropic	57	$4.10	51	26.06	35.86	Model ↗ Providers ↗
Gemini 3.1 Pro Preview	1M	Google	57	$1.74	123	26.40	30.48	Model ↗ Providers ↗
GPT-5.5 (medium)	922k	OpenAI	57	$4.35	61	5.88	14.07	Model ↗ Providers ↗
Qwen3.7 Max	1M	Alibaba	57	--	--	--	--	Model ↗ Providers ↗
Gemini 3.5 Flash	1M	Google	55	$1.31	277	18.55	20.36	Model ↗ Providers ↗
Kimi K2.6	256k	Kimi	54	$0.70	97	2.85	53.92	Model ↗ Providers ↗
MiMo-V2.5-Pro	1M	Xiaomi	54	$0.71	57	2.68	46.77	Model ↗ Providers ↗











[LLM Leaderboard - Comparison of over 100 AI models]

[Full LLM Leaderboard — Advanced Explorer (2026)]

MADRID INNOVATION | MIL | AI Network

【AI Summer Camp】

QUIZ

Veamos cuánto habéis aprendido:

Sección 5

PRIVACIDAD Y COSTE EN EL MUNDO IA GEN

Dilema Open-Source vs Propietario

【AI Summer Camp】

Dilema Estratégico: Open Source vs. SaaS Privado

Balance entre Soberanía del Dato y Vanguardia Tecnológica

Modelos Open Source

LLaMA

Mistral AI

Ventajas:

- Control y soberanía total del dato.
- Coste fijo (CapEx) sin "Token Tax" variable.
- Ejecución 100% offline / On-Premise.

Riesgos:

- Complejidad técnica en el despliegue.
- Necesidad de inversión en hardware (GPUs).

Casos de Uso:

- Datos altamente sensibles (Salud, Defensa).
- Tareas de producción con volumen masivo y constante.

SaaS "Privado"

Google AI

ANTHROP\C

Ventajas:

- Rendimiento de frontera y actualizaciones continuas.
- Mantenimiento cero; despliegue inmediato.
- Escalabilidad infinita bajo demanda.

Riesgos:

- Privacidad: Los datos pueden entrenar al modelo.
- Vendor Lock-in y dependencia de la nube (Cloud Act).

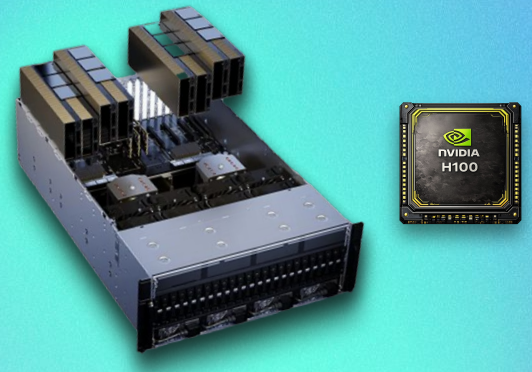
Casos de Uso:

- Prototipado rápido y I+D de alta complejidad.
- Aplicaciones que requieren el estado del arte (SOTA).

TEE

Trusted Execution Environment

Zona segura del procesador que protege la **confidencialidad de los datos** (nadie externo puede leerlos) e **integridad del código** (nadie puede modificarlos, ni siquiera el dueño del equipo en ciertos casos).



Privacidad 2.0: TEEs y Seguridad de Hardware en IA

Seguridad por Hardware

Los datos se procesan en una caja fuerte física dentro de la CPU/GPU.

SaaS suelen tener el riesgo de que los datos de entrada del usuario puedan ser utilizados para entrenar el modelo, perdiendo la confidencialidad

Ni siquiera el administrador del sistema puede ver qué está calculando la IA.

Permite ejecutar modelos en la nube con la soberanía total de un servidor local.

Clave para: Salud, Finanzas y Secreto Industrial.

Caso Real: La Fuga de Samsung

Cuando la IA compromete la Privacidad

En abril de 2023, ingenieros de Samsung subieron código sensible a ChatGPT para optimizarlo.

El problema: Los modelos SaaS entrenan con tus entradas. El código secreto de Samsung pasó a formar parte del cerebro de OpenAI.

Consecuencia: Prohibición total de IA externa en muchas Big Tech y aceleración de la demanda de IA Privada Local

【AI Summer Camp】

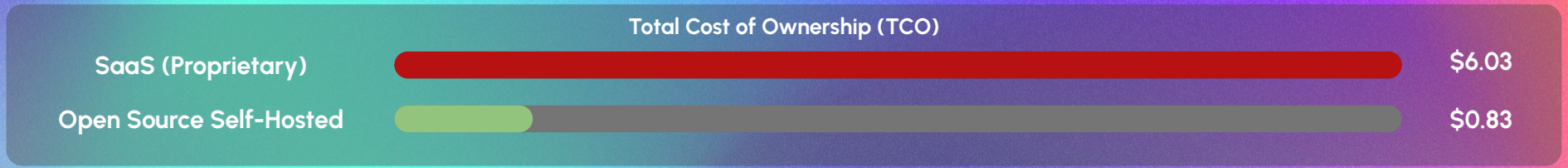
TCO: El Coste Real de Operar IA

SaaS (Modelos Propietarios)

- **Estructura:** Pago por uso (OpEx) mediante "Token Tax" variable.
- **Costes Ocultos:** Dependencia de red, latencia y escalado de precios sin previo aviso.
- **Ventaja:** Inversión inicial cero (CapEx) y mantenimiento nulo.
- **Estimación:** ~$6.03 por millón de tokens en modelos frontera.

Open Source (On-Prem / Self-Hosted)

- **Estructura:** Inversión en hardware (CapEx) con costes operativos fijos.
- **Eficiencia:** Optimización de infraestructura propia (H100/B200) para cargas constantes.
- **Ventaja:** Soberanía total y 100% offline tras la inversión inicial.
- **Estimación:** ~$0.83 por millón de tokens (hasta un 86% más barato).



Análisis de Ahorro: Un modelo Open Source en infraestructura optimizada puede ser un **86% más barato** a largo plazo para cargas de trabajo constantes.

*Basado en estudios de benchmarking de tokens/segundo en Infraestructura optimizada Open Source Self-Hosted (H100/B200) vs API pricing 2025.

MADRID INNOVATION | MIL | AI Network

Artificial Analysis: LLM Hosting API Prices and TCO Calculator

【AI Summer Camp】

Sección 6

TRANSFORMERS Y ALUCINACIONES

Probabilidad frente a comprensión

【AI Summer Camp】

¿Qué es un Transformer? El Motor de la IA Gen

Comprensión funcional: De palabras a pensamiento estadístico

Transformers son un tipo de arquitectura de red neuronal que transforma o cambia una secuencia de entrada en una secuencia de salida:

1.- Aprenden el contexto
2.- Calculan las relaciones entre los componentes de la secuencia de entrada.

Por ejemplo...

Secuencia de entrada: "¿De qué color es el cielo?"

Calcula relación matemática → entre las palabras color, cielo y azul.

Secuencia de salida: "El cielo es azul"

1. El Concepto: La IA no entiende, predice

Es una arquitectura que procesa datos en paralelo. No lee palabra por palabra como un humano, sino que analiza todas las piezas de información a la vez para detectar patrones globales.

2. El Mecanismo de Atención (Self-Attention)

Es el "foco" del modelo. Permite que el sistema integre el significado de palabras que están separadas en un texto para entender el contexto real.
Ejemplo: En "El perro no cruzó la calle porque estaba muy cansado", el modelo asigna más "atención" a la conexión entre estaba y perro, sabiendo exactamente a qué se refiere el pronombre.

3. Limitaciones Estructurales

- **Ventana de Contexto:** Memoria finita; olvida lo que está fuera de su rango.
- **Lost in the Middle:** Tiende a ignorar información en el centro de textos largos.
- **Loros Estocásticos:** Prioriza la probabilidad sobre la veracidad (causa de alucinaciones).

MADRID INNOVATION | MIL | AI Network

AWS: ¿Qué son los Transformers en IA GEN?

【AI Summer Camp】

Alucinaciones: Loros Estocásticos

La metáfora de los "loros estocásticos" sugiere que estas herramientas generan respuestas prediciendo y seleccionando palabras basándose en la probabilidad estadística de sus datos de entrenamiento, sin comprender el significado de su producción.

La IA no es una base de datos, es un **predictor de probabilidad**.

Por qué alucina: Prioriza la probabilidad de la siguiente palabra sobre la veracidad del hecho.

Solución RAG: Darle al "loro" el libro abierto (tu base de datos) para que lea en lugar de intentar recordar lo que aprendió hace 2 años.

Bloque 2

DINÁMICA DE EQUIPOS

Tomando decisiones como un CTO

【AI Summer Camp】

Dinámica de Grupo: El Desafío del CTO

Como Directores de Tecnología, vuestra misión es seleccionar la arquitectura óptima para tres retos de negocio reales.

Escenario 1: Customer Service

Implementación de un chatbot global de atención al cliente multilingüe con picos de tráfico masivos.

- Alta disponibilidad requerida.
- Tratamiento de datos personales básicos.

Escenario 2: Diagnóstico Médico

Sistema de apoyo a la decisión clínica en hospital local para análisis de informes de pacientes críticos.

- Datos de salud altamente sensibles.
- Necesidad de soberanía total.

Escenario 3: Análisis de Mercado

Herramienta interna de I+D para resumir informes financieros públicos y tendencias de competidores.

- Volumen de documentos extremo.
- Presupuesto operativo muy ajustado.

Vectores de Decisión Críticos

1. Seguridad (GDPR)
Protección del dato y cumplimiento normativo.

2. Presupuesto (TCO)
Coste total de propiedad (OpEx vs CapEx).

3. Latencia
Velocidad de respuesta e impacto en UX.

GRACIAS

Pablo Fernández Díaz

【AI Summer Camp】

MADRID INNOVATION 【MIL】 AI Network