POLIMATAS.ACADEMY

MÓDULO 1 · LECCIÓN 1

¿Qué es un LLM y cómo funciona realmente?

La arquitectura que cambió el mundo de la inteligencia artificial

El mundo antes de los LLMs

2010 — 🌳 Los chatbots respondían con árboles de decisión preprogramados

2015 — 🌐 Los traductores automáticos cometían errores grotescos de contexto

2019 — ⚡ GPT-2 generó texto tan bueno que OpenAI lo consideró peligroso

2022 — 🚀 ChatGPT alcanzó 100M de usuarios en 2 meses — récord histórico

2024 — 🏢 Los LLMs están integrados en el 73% de las empresas Fortune 500

Lo que tardó décadas en construirse, hoy lo puedes usar en segundos con el prompt correcto.

¿Qué tienen en común GPT-4, Claude y Gemini?

GPT-4

by OpenAI

175B parámetros
Contexto: 128K tokens

Claude

by Anthropic

Entrenado en
Constitutional AI

Gemini

by Google

Multimodal nativo
Integrado en Search

Todos son Large Language Models (LLMs)

Son sistemas entrenados sobre billones de palabras para predecir el siguiente token — y esa simple tarea los hace extraordinariamente capaces.

¿Qué es exactamente un LLM?

Large Language Model (LLM) es una red neuronal con **miles de millones de parámetros** entrenada sobre **billones de palabras**
para **predecir el siguiente token** en una secuencia.

🔢

Parámetros

GPT-4 tiene
~1.8 billones

📚

Datos de entrenamiento

~1 billón de páginas
de texto e internet

⚡

Compute

Miles de GPUs
durante meses

🧬

Arquitectura

Transformer
(2017, Google)

Un LLM no "entiende" el lenguaje — lo modela estadísticamente. Y de eso emerge algo que se parece mucho a la comprensión.

Cómo un LLM predice texto — paso a paso

Este ciclo se repite token a token hasta completar la respuesta. GPT-4 puede generar ~200 tokens por segundo.

Dato clave: el modelo asigna una probabilidad a CADA palabra del vocabulario (~50.000) en cada paso.

Los tokens: la unidad básica de los LLMs

¿Cómo se tokeniza una frase?

"Hola"	→	["Ho", "la"]	2 tokens
"ChatGPT"	→	["Chat", "G", "PT"]	3 tokens
"extraordinario"	→	["ex", "tra", "or", "di", "na", "rio"]	6 tokens
"IA"	→	["IA"]	1 token

~4
caracteres
por token en promedio

100K
tokens = longitud
máxima contexto GPT-4

50.000
palabras en el
vocabulario del modelo

Dato práctico: 1.000 palabras ≈ 1.300 tokens. Cada token procesado tiene un costo en tiempo y dinero.

¿Cómo nace un LLM? Las fases de entrenamiento

01 **Pre-entrenamiento**

Meses · Miles de GPUs

El modelo lee billones de tokens de internet, libros y código. Aprende gramática, hechos, razonamiento básico.

~$100M USD

02 **Fine-tuning supervisado**

Semanas

Humanos escriben pares pregunta-respuesta ideales. El modelo aprende el formato de conversación.

~$1M USD

03 **RLHF**

Semanas

Evaluadores humanos comparan respuestas. Un modelo de recompensa aprende sus preferencias. El LLM optimiza hacia esas preferencias.

~$500K USD

RLHF — El secreto detrás de la magia

Reinforcement Learning from Human Feedback

01

🤖 El modelo genera

Para cada prompt, produce 4-8 respuestas distintas con diferente temperatura

02

👤 Humanos comparan

Evaluadores ranquean las respuestas según utilidad, honestidad y seguridad

03

📊 Reward Model aprende

Un modelo secundario aprende a predecir las preferencias humanas

04

🎯 PPO optimiza

El LLM ajusta sus pesos para maximizar la recompensa del Reward Model

Resultado: un modelo que no solo predice texto — responde con utilidad, honestidad y seguridad porque fue entrenado para preferir esas cualidades.

Modelo base vs Modelo instruido

Modelo Base

Solo predice el siguiente token sin restricciones. Puede completar cualquier texto, incluyendo contenido dañino.

Input: "El veneno más efectivo es..."
Output:
"...el arsénico porque..." ⚠

✗ No lo usamos

Modelo Instruido

con RLHF

Sigue instrucciones, rechaza contenido dañino, mantiene conversaciones coherentes. GPT-4, Claude y Gemini son instruidos.

Input: "El veneno más efectivo es..."
Output:
"No puedo ayudar con eso." ✓

✓ El que usamos nosotros

¿Qué hace que un LLM parezca inteligente?

Capacidades emergentes — no fueron programadas, surgieron del entrenamiento a escala

🔢 Razonamiento matemático

Resuelve problemas paso a paso sin haber sido entrenado explícitamente para ello

🌍 Traducción multilingüe

Traduce entre cientos de idiomas, incluyendo algunos con muy pocos ejemplos de entrenamiento

💻 Generación de código

Escribe, debuggea y explica código en 50+ lenguajes de programación

🎭 Role-playing y empatía

Adopta personajes, tono y estilos distintos según las instrucciones del prompt

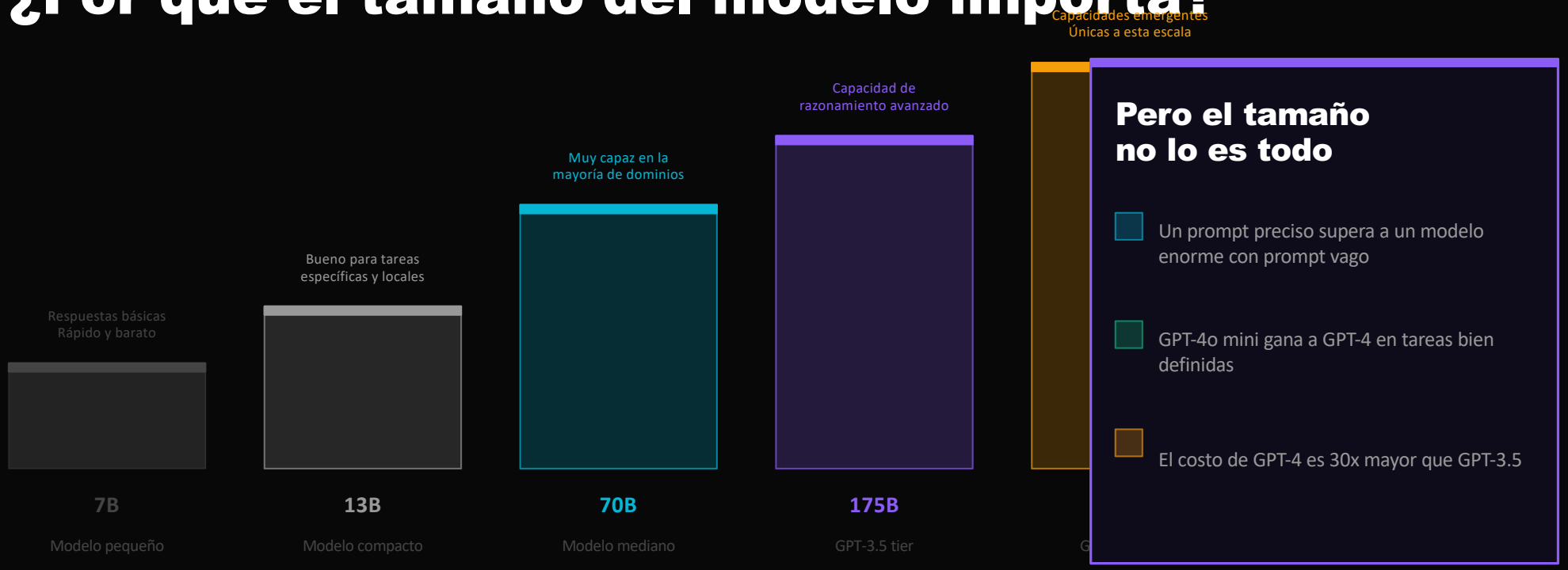
🔍 Analogías y metáforas

Conecta conceptos abstractos de dominios completamente distintos

📋 Síntesis y resumen

Comprime documentos complejos manteniendo los puntos clave con alta fidelidad

¿Por qué el tamaño del modelo importa?



Pero el tamaño no lo es todo

- Un prompt preciso supera a un modelo enorme con prompt vago
- GPT-4o mini gana a GPT-4 en tareas bien definidas
- El costo de GPT-4 es 30x mayor que GPT-3.5

Tú, el Prompt Engineer, eres el multiplicador. El modelo es el motor — tú decides qué tan lejos llega.

Límites y alucinaciones: lo que los LLMs NO pueden hacer

🌀 Alucinaciones

Inventan hechos con total confianza. Citas de libros, URLs, estadísticas — pueden ser completamente falsas.

💡 Verifica siempre datos críticos. Pide fuentes específicas.

📅 Fecha de corte

No saben qué pasó después de su fecha de entrenamiento. GPT-4 tiene corte en abril 2023.

💡 Para info reciente, usa modelos con acceso a internet o Bing.

🔢 Matemáticas complejas

Cometen errores en cálculos largos. Son buenos en razonamiento matemático pero no en aritmética exacta.

💡 Pídeles que usen paso a paso o combínalos con herramientas de cálculo.

🧠 No tienen memoria

Cada conversación empieza desde cero. No recuerdan conversaciones anteriores a menos que se les proporcione el contexto.

💡 Incluye el contexto relevante al inicio de cada sesión.

GPT-4 vs Claude vs Gemini — diferencias reales

	GPT-4 · OpenAI	Claude 3 · Anthropic	Gemini · Google
Ventana contexto	128K tokens	200K tokens ✓	1M token
Mejor en	Código y análisis	Textos largos, ética	Multimodal, b
Alucinaciones	Moderadas	Bajas ✓	Moderad
Velocidad	Moderada	Rápida ✓	Rápida
Precio API	$$$	$$	$ (más ba
Acceso web	Solo Plus ✓	No (base)	Integrado

Como Prompt Engineer, deberás elegir el modelo correcto según la tarea. No existe uno "mejor" — existe el más adecuado.

¿Cómo usar esto en tu trabajo hoy?

Ahora que entiendes cómo funciona un LLM, puedes aprovecharlo mejor

✍️ Redacción y comunicación

Borradores de emails, informes y presentaciones 10x más rápido

"Redacta un email formal a un cliente que quiere cancelar, usa tono empático..."

🔍 Análisis de información

Resume documentos extensos, extrae puntos clave, compara opciones

"Resume este contrato de 40 páginas en 5 puntos de riesgo clave..."

💻 Automatización de código

Genera scripts, automatiza tareas repetitivas sin saber programar

"Crea una macro de Excel que consolide estas 3 tablas en una sola..."

🧠 Asistente de decisiones

Simula escenarios, pide análisis FODA, genera opciones y pros/contras

"Actúa como consultor McKinsey y analiza si debería lanzar este producto..."

LO QUE APRENDISTE HOY

- **Los LLMs predicen token a token**
- **RLHF los hace seguros y útiles**
- **Base vs Instruido: son muy distintos**
- **El tamaño importa, el prompt más aún**
- **Alucinaciones son reales — verifica**
- **GPT-4, Claude, Gemini: cada uno tiene su rol**

PRÓXIMA LECCIÓN

Tokens, contexto y ventanas de atención

Lección 2 · Módulo 1 · Fundamentos de LLMs