

**\*\*2017: Transformer (Vaswani et al.)\*\* ← \*Base de todos los LLMs modernos\***

- \*\*1. Rama GPT (OpenAI)\*\***
  - 2018: GPT-1 (117M parámetros)
  - 2019: GPT-2 (1.5B)
  - 2020: GPT-3 (175B)
  - 2022: ChatGPT (GPT-3.5)
  - 2023: GPT-4 (multimodal)
  - 2024: GPT-4 Turbo (128k contexto)
  - 2025: GPT-5 (rumorado: >300B, multi-agente)
  - 2021: Codex → GitHub Copilot
  - 2023: ChatGPT Plugins → GPTs personalizados
  - 2022: InstructGPT (fine-tuning con RLHF)
- \*\*Derivados y Extensiones de GPT\*\***
  - 2023–2024: **\*\*Claude\*\*** (Anthropic)
    - Claude 1 → 2 → 3 (2024, contextos largos)
  - 2023: **\*\*Sage\*\*** (Poe, Quora)
  - 2024: **\*\*Grok\*\*** (xAI - Elon Musk, integrado en X)
  - 2024: **\*\*Groq\*\*** (motor LPU, ultra rápido inference)
  - 2025: **\*\*GPT Agents\*\*** → integración con acciones web, APIs y memoria
- \*\*2. Rama BERT (Google)\*\***
  - 2018: BERT (Bidirectional Encoder)
  - 2019: RoBERTa (Meta), ALBERT, DistilBERT
  - 2020: T5 (Text-to-Text Transfer Transformer)
  - 2023: FLAN-T5 (fine-tuning con instrucciones)
  - 2022: PaLM (540B)
  - 2023: PaLM-2
  - 2023: Gemini 1.0 (nace Gemini, multimodal)
  - 2024: Gemini 1.5 (1M tokens, más estable)
  - 2025: Gemini 2.0 (multimodal completo: texto, imagen, audio, video)
  - Otros modelos Google: UL2, RETRO (recall via memory)
- \*\*3. Rama Open-Source / Efficient LLMs\*\***
  - EleutherAI
    - 2020: GPT-Neo
    - 2021: GPT-J (6B)
    - 2022: GPT-NeoX (20B)
    - 2023: Pythia (accesibilidad + reproducibilidad)
  - Meta
    - 2022: OPT (175B)
    - 2023: LLaMA-1 (7B–65B)
      - Alpaca (Stanford), Vicuna (finetuned)
    - 2023: LLaMA-2 (70B, licencia comercial)
    - 2024: LLaMA-3 (con >400B parámetros previsto)
  - Mistral
    - 2023: Mistral 7B
    - 2024: Mixtral (MoE de expertos activados)
  - Falcon (TII, EAU, 40B–180B)
  - DeepSeek (China)
    - 2024: DeepSeek-V2 (236B)
    - DeepSeek Coder (optimizado para código)
    - 2025: DeepSeek-V3 (esperado)
  - Otros: Phi-2 (Microsoft, 2.7B), OpenChat, RWKV (RNN + Transformer)
- \*\*4. Otros Modelos Notables y Alternativos\*\***
  - 2021: MT-NLG (Microsoft/NVIDIA, 530B)
  - 2023: XGen (Salesforce, económico y eficiente)
  - 2024: Command R+ (Cohere, RAG optimized)
  - 2025: Mamba (State-space model, alternativa a Transformers)
  - 2025: Project Starline (Google, IA multimodal conversacional con presencia)
  - Apple (rumorado modelo multimodal privado para dispositivos 2025)
- \*\*Tendencias 2025\*\***
  - \*\*Multimodalidad total\*\***: texto, imagen, audio, video y sensores
  - \*\*Agentic AI\*\***: IA con autonomía operativa → AutoGPT, BabyAGI, Devin, ChatDev
  - \*\*Memoria y contexto extendido\*\***: modelos con “long-term memory” y 1M+ tokens
  - \*\*Integración con herramientas\*\***: navegación, ejecución de código, control de apps
  - \*\*Modelos especializados\*\***: en medicina, derecho, programación, ingeniería
  - \*\*Personalización a escala\*\***: LLMs personales con privacidad local
  - \*\*Infraestructura\*\***: Groq (LPU), Cerebras, Graphcore, NVIDIA H200, AI edge devices
  - \*\*Nuevas arquitecturas\*\***: MoE (Mixture of Experts), SSM (State Space Models)
  - \*\*Regulación e IA alineada\*\***: AI Act (UE), Biden Order (EE.UU.), NIST
  - \*\*Evaluaciones éticas\*\***: Red teams, interpretabilidad, IA responsable y sesgos

Es un nuevo paradigma para pensar, decidir y ejecutar. Desde la irrupción de los Transformers en 2017, hemos pasado de modelos con cientos de millones de parámetros a sistemas multimodales capaces de comprender texto, voz, imágenes y video en paralelo.

En 2023, más del 65% de las empresas líderes a nivel global ya habían incorporado IA en al menos un proceso clave. Para 2025, ese porcentaje superará el 85%.

Pero la verdadera pregunta no es **quién la usa**, sino **quién la entiende como ventaja competitiva central y actúa en consecuencia**.

- ¿Dónde estoy?
- ¿Qué modelo refleja mejor mi realidad tecnológica: Claude, LLaMA, Gemini, GPT-4, un modelo open-source... o ninguno?

Porque la IA avanza a velocidad exponencial. No espera aprobaciones jerárquicas, presupuestos congelados ni roadmaps de 18 meses.

Hoy, los modelos entrenan en una semana más información de la que cualquier comité humano podría procesar en una década.

Toman decisiones más rápido, con mayor contexto, y con una eficiencia que reconfigura las reglas de competencia.

Es cuánto tiempo más vas a sostener un modelo operativo que:

- no delega,
- no automatiza,
- no escala su capacidad de juicio.

Las empresas que prosperen no serán las que más gasten en IA, sino aquellas que rediseñen su mentalidad, su liderazgo y su forma de ejecutar **con la IA como núcleo operativo**.

Hoy, el verdadero riesgo no es tecnológico, es estratégico.

Y cada CEO, CIO, CTO o CPO debe hacerse una única pregunta urgente:

**¿Qué tan irrelevante será mi compañía si no actúo ahora?**