Módulo 2. LLMs

CÓRDOBA CREA FUTURO

1. El corazón del lenguaje: un viaje al interior de los LLM

1.1 Recapitulación de IA generativa y definición de LLM

Como recordatorio, la inteligencia artificial generativa es una categoría de algoritmos que, a diferencia de las IA analíticas que clasifican o predicen a partir de datos existentes, generan artefactos nuevos y coherentes. Dentro de esta familia, los *large language models* (LLM) son los motores más potentes y versátiles, especializados, como su nombre indica, en el lenguaje humano.

Entonces, ¿qué es exactamente un LLM? En su nivel más fundamental, un LLM es un tipo de red neuronal de tamaño colosal, entrenada con una cantidad de texto casi inimaginable (básicamente, una gran porción de todo lo que la humanidad ha escrito y digitalizado). Su objetivo principal, aunque suene simple, es extraordinariamente complejo: predecir la siguiente palabra más probable en una secuencia de texto.

Podemos imaginarlo como la función de «autocompletar» de tu teléfono, pero con superpoderes. Sin embargo, esta analogía se queda corta rápidamente. Al entrenarse para realizar esta tarea a una escala masiva, los LLM no solo aprenden gramática y vocabulario: desarrollan capacidades emergentes que van mucho más allá. Aprenden a captar el significado, el contexto, el estilo, el razonamiento e incluso ciertas formas de conocimiento sobre el mundo. Esta habilidad para predecir secuencias lógicas les permite realizar tareas asombrosas: resumir un libro, traducir idiomas en tiempo real, escribir código de programación, mantener una conversación coherente o explicar un concepto científico complejo.

Todos estos modelos modernos, desde ChatGPT hasta Gemini, se basan en una arquitectura de red neuronal revolucionaria conocida como *Transformer*, que fue el avance clave que desató su potencial actual.

1.2 El proceso de «pensamiento»: cómo los LLM entienden y generan lenguaje

Para que un *large language model* pueda procesar tu pregunta y generar una respuesta, debe convertir el lenguaje humano —caótico y ambiguo— en el lenguaje universal de las matemáticas: los números. Este proceso puede desglosarse en tres pasos fascinantes.

Paso 1: de palabras a números (*tokenización*)

Lo primero que hace un *large language model* es descomponer el texto de entrada en piezas más pequeñas llamadas *tokens*. Un *token* no es necesariamente una palabra completa: puede ser una palabra, una parte de una palabra (como «camin-» y «-ando»), un signo de puntuación o un espacio. Por ejemplo, la frase «Inteligencia Artificial» podría convertirse en la secuencia de *tokens*: [«Inteli», «gencia», « », «Artifi», «cial»]. Este método es increíblemente eficiente, ya que permite al modelo manejar un vocabulario casi infinito y comprender palabras nuevas o complejas descomponiéndolas en partes que ya conoce.

Figura 1. IA generativa - paso 1

Fuente: Kotwani, 2018, https://goo.su/xBRf

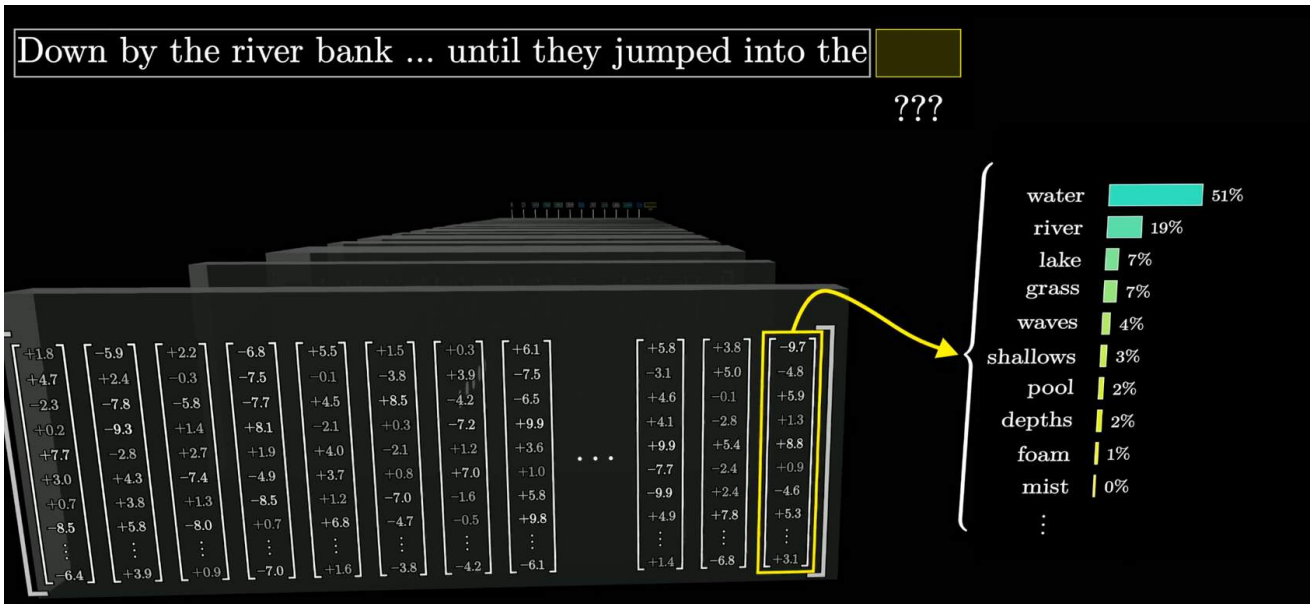
Un concepto importante en los *large language models* es la ventana de contexto, que es la cantidad de texto (o tokens) que el modelo puede considerar de una sola vez al generar una respuesta. Si la ventana de contexto es de, digamos, 4000 *tokens*, significa que el modelo solo «recuerda» o procesa hasta ese número de tokens de la conversación (incluyendo la pregunta del usuario y su propia respuesta parcial) antes de empezar a «olvidar» lo más antiguo. Los modelos más avanzados han ido ampliando estas ventanas.

Paso 2: el mapa del significado (word embeddings)

Una vez que el texto está tokenizado, el modelo debe entender qué significa cada *token*. Aquí es donde ocurre la magia de los *word embeddings* (incrustaciones de palabras). Cada *token* se convierte en una larga lista de números, conocida como

un vector. Este vector no es aleatorio: representa las «coordenadas» del *token* en un vasto «mapa de significado» multidimensional.

Figura 2. IA generativa - paso 2

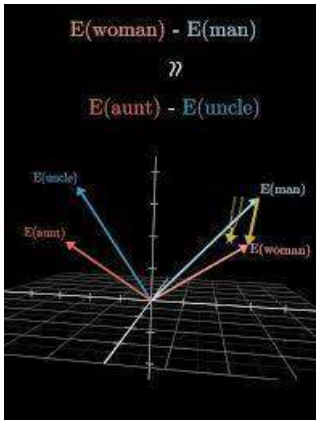


Fuente: Kotwani, 2018, https://goo.su/xBRf

Imagina un mapa gigante donde las palabras se colocan según su contexto y significado. Palabras como «perro», «gato» y «hámster» estarían agrupadas en una región del mapa, la zona de «mascotas». Cerca de allí, pero en otra agrupación, podrían estar «león», «tigre» y «pantera». Lo más asombroso es que el mapa no solo captura similitudes, sino también relaciones. La distancia y dirección entre «rey» y «reina» sería muy similar a la distancia y dirección entre «hombre» y «mujer». Esto permite al modelo realizar una especie de álgebra semántica, como en la famosa ecuación: vector(«Rey») − vector(«Hombre») + vector(«Mujer») da como resultado un vector muy cercano al de vector(«Reina»).

Este proceso es fundamental porque revela una verdad profunda sobre cómo funcionan estos modelos. Su «comprensión» no es como la nuestra, basada en la experiencia vivida y la conciencia. Es una comprensión estadística y relacional, construida a partir de las posiciones y relaciones matemáticas entre miles de millones de palabras en sus datos de entrenamiento. El modelo no «sabe» lo que es un rey, pero «sabe» que la palabra «rey» aparece en contextos similares a «reina», «corona» y «castillo». Esta naturaleza puramente matemática es también la razón por la que los LLM a veces pueden «alucinar» o inventar información: no tienen un anclaje en la realidad verificable, solo en los patrones estadísticos del texto con el que fueron alimentados. Su inteligencia es de correlación, no de causación.

Figura 3. Relaciones semánticas en el espacio vectorial de *word embeddings*



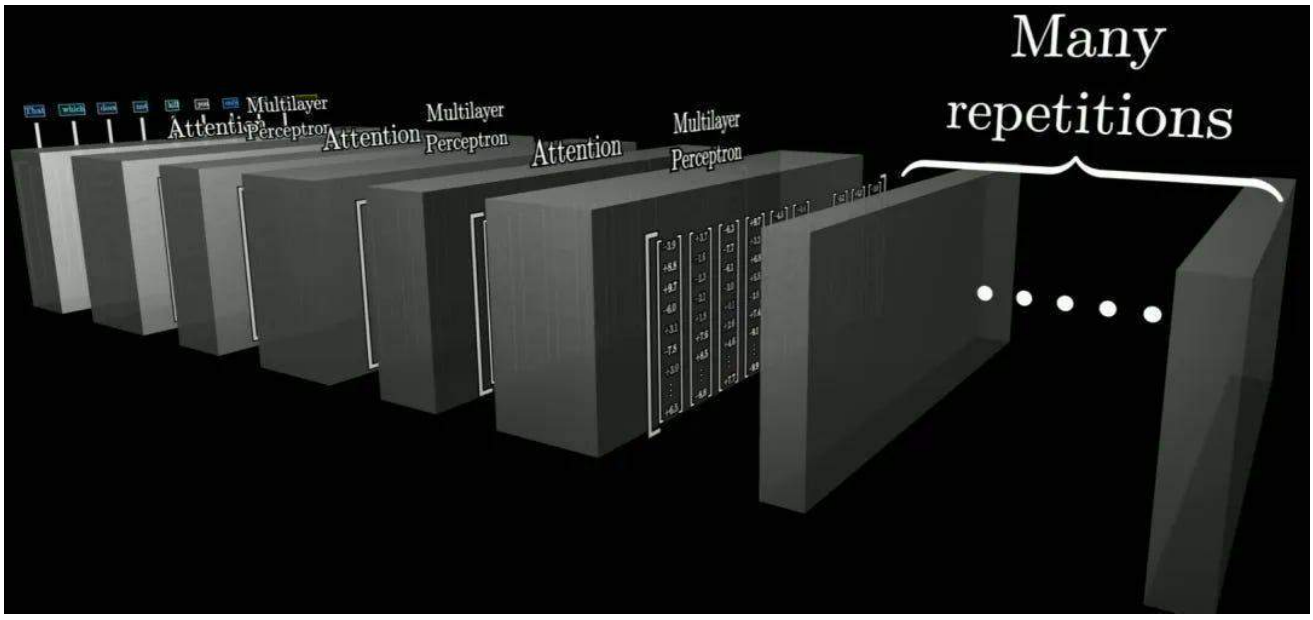
Fuente: Kotwani, 2018, https://goo.su/xBRf

Paso 3: el secreto está en el contexto

Saber el significado de las palabras individuales no es suficiente; el contexto lo es todo. Considera la frase: «El banco estaba cerca del río». Sin contexto, «banco» podría ser una institución financiera o un asiento. Para resolver esta ambigüedad, necesitamos mirar las otras palabras de la frase. Aquí es donde entra en juego la arquitectura *Transformer*, presentada por investigadores de Google en 2017 en un artículo titulado *Attention Is All You Need* (Vaswani et al., 2017).

El componente clave del *Transformer* es el mecanismo de atención. Si pensamos en cómo leemos nosotros: cuando vemos la frase «el robot dejó caer la pelota porque estaba oxidado», nuestro cerebro instantáneamente presta más «atención» a la palabra «robot» para entender a qué se refiere «estaba». El mecanismo de atención hace exactamente eso de forma matemática. A medida que procesa la frase, asigna una puntuación de importancia a cada palabra en relación con todas las demás. Esto permite al modelo sopesar la influencia de cada palabra y entender las relaciones a larga distancia dentro de un texto, capturando el contexto con una eficacia sin precedentes.

Figura 4. IA generativa - paso 3



Fuente: Kotwani, 2018, https://goo.su/xBRf

Esta idea de la «atención» fue tan poderosa que se convirtió en el «*big bang*» de la IA moderna. No solo revolucionó el procesamiento del lenguaje, sino que su principio fundamental —permitir que cada elemento de una secuencia pondere la importancia de todos los demás— se ha generalizado a otros dominios. Los Vision Transformers (ViTs), por ejemplo, aplican el mismo mecanismo a parches de píxeles en una imagen, logrando resultados de vanguardia en visión por computadora. Esta unificación arquitectónica es la razón por la que hemos visto una explosión de capacidades multimodales, donde un solo modelo puede entender texto, imágenes y sonido, ya que ahora existe un lenguaje común para procesar diferentes tipos de datos.

1.3 El poder en la práctica: utilidad y ejemplos de los LLM en el mundo real

La teoría es fascinante, pero el verdadero impacto de los LLM se ve en su aplicación práctica, que ya está transformando innumerables aspectos de nuestra vida y trabajo.

- **Asistentes de escritura y creatividad.** Son una herramienta increíble para superar el bloqueo del escritor. Pueden ayudarte a redactar un correo electrónico profesional, generar una lluvia de ideas para un ensayo o componer un poema al estilo de tu autor favorito.

- **Herramientas de programación.** Para los desarrolladores de software, los LLM son como un compañero de programación experto. Herramientas como GitHub Copilot,

impulsadas por modelos de OpenAI, pueden sugerir líneas de código, completar funciones enteras, explicar errores complejos y traducir código de un lenguaje a otro, acelerando drásticamente el proceso de desarrollo.

- **Educación personalizada.** Imagínate un tutor personal disponible 24 hs/día, 7 días a la semana. Un LLM puede explicar la fotosíntesis de diez maneras distintas hasta que la entiendas, crear cuestionarios de práctica sobre cualquier tema o simular conversaciones para ayudarte a aprender un nuevo idioma.

- **Análisis y resumen de información.** En un mundo saturado de información, los LLM son una herramienta invaluable para la productividad. Pueden leer un informe de 100 páginas y darte los puntos clave en un solo párrafo, resumir una cadena de correos electrónicos o transcribir y analizar una reunión de una hora en segundos.

- **Interacción multimodal.** Los modelos más avanzados ya no se limitan al texto. Pueden analizar una imagen y describirla en detalle para una persona con discapacidad visual, interpretar los datos de una gráfica financiera, escuchar una conversación y transcribirla o incluso ver un vídeo y explicar lo que está sucediendo. Esta capacidad de procesar e integrar información de diferentes modalidades abre un universo de nuevas aplicaciones.

04:15

CONTINUAR

2. La búsqueda de la inteligencia artificial general (AGI)

Cuando hablas de IA generativa y LLM, es inevitable mencionar la «meta final» que muchos imaginan: la inteligencia artificial general (AGI, por sus siglas en inglés). ¿Qué es la AGI? A diferencia de las IA actuales, que son «expertas» en tareas específicas —por ejemplo, un modelo que juega ajedrez a nivel superhumano, pero que no puede hacer nada más, o un chatbot que sabe conversar sobre textos, pero no manejar un robot físico—, la AGI se refiere a una inteligencia artificial con capacidades generales al nivel de un ser humano en prácticamente cualquier tarea cognitiva. Es decir, una IA que pueda aprender y entender cualquier problema como lo harías tú, y aplicarlo en distintos dominios: desde matemáticas hasta literatura, pasando por entender emociones, tener sentido común, resolver problemas nuevos y adaptarse a diferentes contextos (IBM, s. f.).

2.1 Contexto e historia del concepto

La idea de una «máquina pensante» que iguale la inteligencia humana existe desde el nacimiento mismo del campo de la IA en los años 1950. En 1956, durante la conferencia de Dartmouth donde nació el término «inteligencia artificial», ya se proponía como hipótesis que «cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede describirse tan precisamente que una máquina pueda simularlo» (IBM, s. f.).

Sin embargo, durante décadas, la investigación se enfocó en lograr avances en IA estrechas o específicas (lo que en inglés se denomina ANI, *artificial narrow intelligence*): programas expertos en dominios concretos como jugar ajedrez, diagnosticar enfermedades, reconocer voz, conducir vehículos, etc. Y de hecho,

viste grandes éxitos en esas áreas (como el caso de *Deep Blue* ganándole al campeón de ajedrez). Pero esas no eran inteligencias «generales», ya que no podían salir de su tarea.

El término *artificial general intelligence* en sí se popularizó más tarde. En 2007, el investigador Ben Goertzel, con la sugerencia de Shane Legg, cofundador de DeepMind, utilizó AGI para distinguir los esfuerzos dirigidos a crear IA con capacidad amplia y flexible de las IA estrechas tradicionales. También se le suele llamar «IA fuerte» (en contraste con la «IA débil» o limitada): la IA fuerte implicaría incluso que la máquina tiene una verdadera conciencia o entendimiento, no solo manipula símbolos. No todos usan los términos de la misma forma; algunos reservan «IA fuerte» específicamente para indicar conciencia, mientras que «AGI» sería simplemente inteligencia al nivel humano, sin pronunciarse sobre si la máquina es consciente o no. En cualquier caso, desde fines de la década de 2000, empezó a haber conferencias y grupos de investigación dedicados a la AGI como objetivo a largo plazo, aunque seguía siendo algo más propio de la ciencia ficción que de laboratorios prácticos (IBM, s. f.).

2.2 ¿Cómo medir o saber si se logra la AGI?

Esta pregunta es compleja y discutida. Tradicionalmente, se mencionaba el test de Turing: si una máquina podía conversar de forma indistinguible de un humano, se la consideraba inteligente. De hecho, superar por completo este test —engañar a evaluadores humanos durante largas conversaciones— sería una señal clara de inteligencia humana en lenguaje. Sin embargo, eso no implica que cubra otras habilidades, como el sentido común físico, el razonamiento matemático avanzado o las capacidades motoras.

Los modelos actuales de lenguaje han demostrado pasar con facilidad esta prueba, por lo que se considera obsoleta para identificar una AGI. Por eso, investigadores han propuesto otros criterios y *benchmarks*. Por ejemplo, se podría decir que se alcanza la AGI cuando un sistema puede aprender cualquier tarea nueva con la

misma rapidez y eficiencia que lo harías tú, o cuando es capaz de transferir conocimientos libremente de un área a otra.

En 2023, Google DeepMind propuso un marco con cinco niveles de desempeño —emergente, competente, experto, virtuoso y superhumano— junto con el grado de autonomía, para clasificar sistemas en el camino hacia la AGI. También se han diseñado pruebas como el Abstraction and Reasoning Corpus (ARC), que plantea desafíos de lógica visual novedosos para evaluar el razonamiento general y el aprendizaje flexible de una IA. Se espera que una AGI sea capaz de resolverlo con relativa facilidad, mientras que los modelos actuales apenas alcanzan resultados muy bajos. Por ejemplo, en 2023, los modelos más avanzados no superaban el 20 % de aciertos en ARC. Un dato reciente muestra que Grok 4, el modelo desarrollado por xAI, obtuvo un 15,9 % en ARC-AGI, lo que evidencia cuán difícil sigue siendo para la IA actual resolver este tipo de pruebas de «inteligencia general».

En resumen, no existe aún un consenso único sobre cómo medir la AGI, pero el objetivo es que la IA logre un rendimiento comparable al de una persona promedio en una amplia gama de tareas cognitivas. Una AGI debería comprender el lenguaje, interpretar contextos culturales, resolver problemas nuevos, tener sentido común e, incluso, según algunas posturas, demostrar creatividad o conciencia de sí misma. Alcanzar este nivel también implicará definir con mayor precisión qué se entiende por «inteligencia» y «comprensión» en una máquina, lo cual sigue siendo parte filosófica del desafío.

2.3 La carrera hacia el futuro: actores y enfoques

Por lo que dijimos antes, existe también una competencia de filosofías sobre cómo construir la AGI y, lo que es más importante, cómo garantizar que sea segura y beneficiosa para la humanidad.

OpenAI: —

su misión declarada al fundarse es «asegurar que la inteligencia artificial general beneficie a toda la humanidad». Su enfoque principal ha sido escalar agresivamente el tamaño y la capacidad de sus modelos, operando bajo la hipótesis de que la cantidad (de datos y cómputo) también aporta calidad, lo cual conduce a una mayor inteligencia.

Google DeepMind: —

formado por la unión de los dos principales laboratorios de IA de Google, su lema es «resolver la inteligencia para avanzar en la ciencia y la humanidad». Tiene una larga historia de investigación fundamental y aborda el problema desde múltiples ángulos, incluyendo aprendizaje por refuerzo y neurociencia computacional.

Anthropic: —

fundada por exmiembros de alto nivel de OpenAI, nació de una profunda preocupación por la seguridad. Su enfoque principal es la investigación en alineamiento de la IA: cómo garantizar que los objetivos de una IA superinteligente estén alineados con los valores humanos. Han sido pioneros en técnicas como la «IA constitucional», donde el modelo se entrena para seguir un conjunto de principios (una «constitución») en lugar de depender únicamente de la retroalimentación humana.

La existencia misma de estas diferentes filosofías revela una tensión fundamental que define el campo de la IA hoy en día: el conflicto entre la velocidad del progreso (*capability*) y la necesidad de control (*safety*). Cada nuevo lanzamiento de un

modelo más potente se celebra como un hito de capacidad, pero también reaviva el debate sobre los riesgos y el alineamiento. El desarrollo de la IA no sigue una línea recta de progreso, sino que es un constante tira y afloja entre el impulso de avanzar lo más rápido posible y el imperativo de garantizar que lo que construimos sea seguro y beneficioso para todos.

Para concluir, vale la pena entender la AGI como un horizonte hipotético: es una idea que nos invita a reflexionar sobre qué significa la inteligencia y qué responsabilidades tendríamos si llegáramos a crear máquinas tan capaces como nosotros. Si bien todavía usamos IA «débiles» en la vida cotidiana, la noción de AGI nos impulsa a considerar tanto las posibilidades extraordinarias —como curar enfermedades o resolver problemas globales— como los desafíos éticos —control, alineación con valores humanos, impacto en los empleos — que vendrían con una inteligencia artificial de nivel humano.

CONTINUAR

3. Principales LLM de las empresas líderes

Aunque la investigación en IA es un esfuerzo global con raíces profundas en la academia, el desarrollo de los *LLM* de vanguardia se ha concentrado en un puñado de grandes empresas tecnológicas. Esto se debe a los requisitos astronómicos de recursos: acceso a cantidades masivas de datos, posesión de supercomputadoras que cuestan miles de millones de dólares y capacidad para atraer al escaso talento de élite en el campo.

Curiosamente, aunque estas empresas están convergiendo tecnológicamente —casi todas construyen *LLM* multimodales basados en la arquitectura *Transformer*—, están divergiendo en sus estrategias de negocio y distribución. La verdadera batalla no gira solo en torno a quién tiene el mejor modelo hoy, sino sobre qué ecosistema —cerrado o abierto— dominará el futuro.

3.1 OpenAI – GPT (Generative Pre-trained Transformer)

Figura 5. ChatGPT

Fuente: captura de pantalla de ChatGPT (https://chatgpt.com/)

OpenAI es una organización fundada en 2015 con la misión explícita de desarrollar una AGI segura y benéfica. Fue cofundada por, entre otros, Elon Musk (quien luego se apartó) y Sam Altman, y desde sus inicios recibió apoyo de empresas como Microsoft. OpenAI comenzó como una entidad sin fines de lucro, pero luego adoptó una estructura de «capped-profit» para atraer inversión, lo que le permitió financiar la investigación masiva necesaria para sus modelos de lenguaje.

OpenAI presentó la serie GPT (Generative Pre-trained Transformer), que ha marcado hitos en la historia de la IA.

GPT-2 (2019):

fue uno de sus primeros *LLM* públicos. Contenía aproximadamente 1.500 millones de parámetros y mostró, por primera vez, la capacidad de generar textos sorprendentemente coherentes en inglés. OpenAI fue cautelosa inicialmente y no liberó de inmediato el modelo más grande de GPT-2, citando preocupaciones sobre su posible mal uso —por ejemplo, en la generación de noticias falsas—.

GPT-3 (2020):

representó un salto enorme en escala, con 175.000 millones de parámetros. GPT-3 demostró una versatilidad notable: con una sola arquitectura podía responder preguntas, redactar ensayos breves, traducir, generar fragmentos simples de código, entre otras tareas, con mínima instrucción. A esta capacidad se la llamó «*few-shot learning*», ya que con solo ver unos ejemplos en el *prompt* podía realizar tareas nuevas.

Se lanzó como una API —es decir, la posibilidad de usar GPT en otras aplicaciones, sin liberar el modelo en sí— y fue la base de numerosos prototipos de asistentes inteligentes entre 2020 y 2021. GPT-3 atrajo la atención fuera del ámbito técnico por primera vez, ya que personas ajenas al desarrollo pudieron probarlo en demostraciones y comprobar la coherencia de sus respuestas. También se comenzó a explorar su uso con fines educativos y creativos.

GPT-3.5 (2022):

no fue un modelo nombrado así públicamente en su momento, pero así se conoce hoy a las versiones mejoradas de GPT-3 que dieron lugar a ChatGPT. A finales de 2022, OpenAI lanzó ChatGPT al público, inicialmente de forma gratuita, y este chatbot, basado en GPT-3.5, se volvió mundialmente conocido. ChatGPT lograba conversaciones fluidas y útiles gracias a un entrenamiento adicional con retroalimentación humana (*reinforcement learning from human feedback*, RLHF), lo que le permitió sonar más seguro, evitar ciertos sesgos y seguir instrucciones conversacionales. En pocos días alcanzó más de un millón de usuarios y, en apenas dos meses, superó los 100 millones de usuarios activos. Este crecimiento sin precedentes marcó el inicio de una expansión masiva del interés por los chatbots de IA en múltiples industrias.

GPT-4 (marzo 2023):

sus especificaciones exactas (como número de parámetros) no fueron reveladas públicamente, pero los expertos estiman que podría tener del orden del trillón (millón de millones) de parámetros, entrenado con una mezcla de datos de texto e imagen. GPT-4 introdujo mejoras notables en razonamiento, entendimiento de instrucciones complejas y capacidad para manejar entradas visuales. De hecho, GPT-4 es multimodal

en la entrada, lo que significa que, además de texto, puede procesar imágenes: puedes darle, por ejemplo, la foto de un diagrama y pedir que lo explique, o cargar una imagen y hacer preguntas sobre su contenido. Esto amplió las posibilidades de uso (aunque en la práctica la funcionalidad de visión se habilitó más ampliamente meses después). En términos de rendimiento, GPT-4 obtuvo resultados que sorprendieron al alcanzar niveles humanos en muchos exámenes: por ejemplo, en pruebas estandarizadas como el examen de abogado o el GRE, sus puntajes estaban en el rango de los alumnos destacados. En el popular *benchmark* MMLU (Massive Multitask Language Understanding), que abarca preguntas de múltiples disciplinas académicas, GPT-4 logró aproximadamente un 86 % de precisión, superando a modelos anteriores por un margen amplio. En cuanto a la ventana de contexto, la versión estándar de GPT-4 manejaba hasta 8000 *tokens*, y OpenAI ofreció una versión ampliada a 32 000 *tokens* para ciertos usuarios, lo que ya suponía un salto respecto de GPT-3, que tenía una capacidad de 2000. GPT-4 se puso a disposición mediante API —bajo pago— y a través de ChatGPT, inicialmente para usuarios de ChatGPT Plus con prioridad. Posteriormente, algunas funciones se extendieron a la versión gratuita con ciertos límites. Su llegada consolidó el liderazgo de OpenAI en la carrera de LLM durante 2023.

GPT-4o («omni») (2024):

este modelo «insignia» multimodal entiende y genera texto, voz e imagen (e incluso puede razonar a partir de video o imagen). Está optimizado para conversaciones rápidas y naturales; por ejemplo, en modo voz responde en aproximadamente 320 ms.

GPT-4.1 (abril 2025):

esta versión se lanzó sin representar un avance enorme respecto de GPT-4o, especializándose principalmente en desarrollo de código, mejor capacidad de seguir instrucciones y contextos más largos.

GPT-4.5:

sucesor iterativo de GPT-4 orientado al entendimiento sutil, el «EQ» conversacional y la creatividad (mejor lectura de matices e intenciones implícitas). OpenAI lo presentó como *research preview* a comienzos de 2025 y publicó su *system card* con metodología y evaluaciones. En la documentación de la plataforma se lo listó con ventana de 128K tokens (*preview*). Fue pensado para escritura y diseño con mayor sensibilidad de tono (por ejemplo, *feedback* empático en redacciones, guías para docentes, redacción creativa de materiales). Sin embargo, este modelo no tuvo mucho éxito; el *preview* evolucionó y OpenAI recomienda hoy usar GPT-4.1, o3 o GPT-5, según el caso. El *preview* de 4.5 fue depreciado en la API más adelante.

o1 (modelos razonadores):

familia nueva (2024) entrenada para «pensar antes de responder». Están diseñados para razonamiento complejo, resolver problemas de varios pasos (ciencia, matemáticas, código) y planificar tareas para flujos tipo «agentes». OpenAI describe el cambio como dedicar más cómputo al *thinking* interno antes de producir la respuesta.

o3:

la evolución 2025 de los razonadores; modelo más potente en razonamiento de OpenAI hasta esa fecha. Mejora código, matemáticas, ciencia y percepción visual (gráficos, diagramas), y marcó el estado del arte en varios *benchmarks* (SWE-bench, MMMU, etc.), con foco en consultas donde la respuesta no es obvia y hace falta análisis multifacético. También existe o3-mini como opción más rápida y económica.

GPT-5:

presentado en agosto de 2025 como el modelo más avanzado y un «salto significativo» en inteligencia respecto de versiones anteriores, con desempeño destacado en código, matemáticas, escritura, salud y percepción visual. OpenAI lo describe como un sistema unificado que «sabe cuándo responder rápido y cuándo pensar más», integrando *thinking* de manera nativa. En la web y en la documentación de la plataforma se posiciona, además, como referente para código y tareas *agénticas* (orquestar cadenas largas de llamadas a herramientas). Es el modelo que actualmente se encuentra disponible por defecto en ChatGPT.

OpenAI, gracias a ChatGPT, posee el LLM más utilizado del planeta. A octubre de 2025, ChatGPT cuenta con aproximadamente 700 millones de usuarios activos semanales, habiendo superado los 800 millones en ciertos momentos pico de ese año. Esto equivale a cientos de millones de personas que utilizan la

herramienta cada día para estudiar, trabajar o por simple curiosidad. ChatGPT se ha convertido en una especie de «asistente universal» al alcance de cualquiera con conexión a internet. Su sitio web llegó a ser el quinto más visitado del mundo, recibiendo on the order of ~5 mil millones de visitas al mes (Exploding Topics, s.f.).

3.2 Google – Gemini

Figura 6. Gemini

Fuente: captura de pantalla de Gemini (https://gemini.google.com/)

Google ha sido durante años líder en investigación de IA (recordemos que fue Google Brain quien inventó la arquitectura *Transformer* en 2017). Sin embargo, en el terreno de los *chatbots*, Google inicialmente se movió con más cautela que OpenAI. Tras el éxito de ChatGPT, aceleró sus proyectos de LLM conversacionales.

Originalmente, presentó LaMDA (*Language Model for Dialogue Applications*) en 2021, un modelo enfocado en diálogos. LaMDA saltó a la fama pública cuando, en

2022, un ingeniero de Google afirmó (erróneamente) que el modelo se había vuelto «sensible» o consciente, lo que generó titulares, pero fue refutado por la comunidad. A inicios de 2023, Google lanzó Bard, su primer *chatbot* público, basándolo inicialmente en LaMDA. Bard se presentó como un producto experimental, con acceso limitado mediante lista de espera en EE. UU. y el Reino Unido. En esos primeros meses, Bard quedó algo rezagado en capacidades frente a ChatGPT (que ya utilizaba GPT-4 para los usuarios Plus). No ayudó que, en su demostración inicial, Bard cometiera un error factual astronómico, lo que infamemente le costó 120 000 millones de dólares en valor bursátil a Google en un solo día por la reacción negativa.

Para mayo de 2023, Google reaccionó integrando en Bard los avances de otro gran modelo llamado PaLM 2. Este era un LLM más potente que LaMDA, entrenado con alrededor de 340 mil millones de parámetros y con sólidas habilidades en idiomas y programación. Con PaLM 2, Bard mejoró y, además, se expandió globalmente a más de 180 países y decenas de idiomas. Google también comenzó a combinar Bard con sus propias herramientas, permitiendo que se conecte con Google Search, Gmail, Docs, entre otras, para recuperar información; una ventaja potencial, dada la amplia base de datos de Google.

A finales de 2023, Google llevó a cabo un *rebranding* importante: anunció Gemini, un nuevo nombre que englobaría su nueva generación de modelos de lenguaje y que sustituiría la marca Bard. En diciembre de 2023 se presentó Gemini 1.0 en varios tamaños: Nano, Pro y Ultra. Gemini es fruto de la colaboración entre Google Brain y DeepMind (las dos divisiones se fusionaron en 2023 bajo el nombre «Google DeepMind»). La expectativa era muy alta: se afirmaba que Gemini buscaría «superar a GPT-4» mediante una arquitectura novedosa y multimodal desde el inicio.

Modelos y versiones

Gemini 1.0 (diciembre de 2023):

se lanzaron las versiones Nano, Pro y Ultra. Nano era un modelo pequeño, optimizado para funcionar localmente (por ejemplo, se integró en los celulares Pixel 8 para funciones básicas sin conexión); Pro, el modelo general para múltiples tareas; y Ultra, el más potente y con mayor capacidad de procesamiento. Bard comenzó a utilizar Gemini 1.0 Pro en su versión pública a finales de 2023, mientras se preparaba un Bard «Advanced» con Gemini Ultra para casos más complejos.

Gemini 1.5 (febrero de 2024):

muy pronto, Google actualizó a la versión 1.5, introduciendo Gemini 1.5 Pro con mejoras y, notablemente, una ampliación drástica del contexto: hasta 1 millón de tokens en ciertos entornos. Si bien, al momento del lanzamiento, el público en general solo podía acceder al modelo con un contexto de 128 000 tokens (una cifra ya muy elevada en comparación con la competencia), con el tiempo esta característica se fue habilitando para todos los usuarios. Este salto en la capacidad de contexto formó parte de una estrategia de Google para diferenciarse. En mayo de 2024 también se presentó Gemini 1.5 Flash, optimizado para ofrecer respuestas más rápidas, aunque con menor profundidad.

Gemini 2.0 (enero de 2025):

a comienzos de 2025 se lanzó Gemini 2.0, con foco en multimodalidad total y capacidades de agente (es decir, poder no solo conversar, sino también tomar acciones

como navegar la web, ejecutar código, entre otras). Gemini 2.0 Flash y Flash-Lite pusieron el énfasis en la velocidad y la eficiencia en el costo.

Gemini 2.5 (marzo de 2025):

en el evento Google I/O 2025, la empresa presentó Gemini 2.5. Allí se introdujo Gemini 2.5 Pro Experimental (marzo de 2025), calificado como «su modelo más inteligente hasta la fecha», con mejoras en razonamiento, codificación y un modo de «pensamiento en cadena» (*chain-of-thought*) que le permitía desglosar problemas en pasos lógicos antes de responder. Este modelo mantenía de forma nativa la multimodalidad (texto, imágenes, audio y video) y contaba con una ventana de 1 millón de *tokens* desde el lanzamiento. Poco después, en abril de 2025, Gemini 2.5 Flash se convirtió en el modelo por defecto, priorizando la velocidad, mientras que Gemini 2.5 Pro quedó como la opción para tareas complejas, con un modo denominado «Deep Think» para razonamientos prolongados. En junio de 2025, tanto la versión 2.5 Pro como Flash estaban disponibles de forma generalizada en la API, incluyendo un submodelo Flash-Lite para clientes que deseaban reducir costos a cambio de una leve disminución en la calidad.

Aunque ChatGPT acaparó titulares, Google Gemini también ha acumulado una base enorme de usuarios gracias a la amplia presencia de Google. Para mediados de 2025, Gemini contaba con alrededor de 450 millones de usuarios activos mensuales en todo el mundo. Google ha promocionado que Gemini tenía 350

millones de usuarios activos a comienzos de 2025, con unos 35 millones diarios y más de 80 millones de descargas de su aplicación móvil.

La ventaja de Google es su ecosistema: integró Gemini en una gran cantidad de productos (Búsqueda, YouTube, Android, Workspace), por lo que el número de personas que interactúan, aunque sea de manera indirecta, con sus LLM es altísimo. En el buscador de Google, muchos usuarios comenzaron a recibir respuestas generativas (la «Search Generative Experience» o SGE) en la parte superior de sus resultados, impulsadas por estos modelos (Meetanshi, s. f.).

3.3 Anthropic – Claude

Figura 7. Claude

Fuente: captura de pantalla de Claude (https://chat.ai-pro.org/)

Anthropic es una *startup* de IA fundada en 2021 por exinvestigadores de OpenAI, entre ellos Dario Amodei. Surgió tras divergencias en cuanto al enfoque: Anthropic se centra especialmente en el desarrollo de IA ética y alineada con valores

humanos, investigando técnicas para que los LLM sean útiles y eviten resultados perjudiciales. Su proyecto principal es Claude, un asistente de lenguaje similar en espíritu a ChatGPT. El nombre «Claude» es un homenaje al científico Claude Shannon, pionero de la teoría de la información.

Desde sus inicios, Anthropic adoptó un enfoque denominado «*Constitutional AI*»: en lugar de depender únicamente de la retroalimentación humana para alinear el modelo (como en el caso del RLHF), definieron una «constitución» de principios que la IA sigue para autocorregir sus respuestas. Por ejemplo, principios como «no aconsejar acciones ilegales», «ser honesto y no engañar al usuario», entre otros, fueron utilizados para afinar a Claude. El objetivo era reducir sesgos y resultados tóxicos de manera más autónoma.

Evolución de los modelos Claude

Claude 1: —

lanzado inicialmente en versión limitada en marzo de 2023, ofrecía un rendimiento comparable al de GPT-3.5, aunque con un sesgo más prudente, ya que tendía a rechazar con mayor firmeza las consultas inapropiadas. Anthropic también lanzó en paralelo Claude Instant, una versión más pequeña y rápida, pensada para aplicaciones que requerían menor latencia o costo. Durante 2023, Claude estuvo disponible a través de API y por medio de socios: por ejemplo, la aplicación de debate Quora Poe integró a Claude, y las empresas podían pagar para utilizarlo en sus servicios.

Claude 2: —

también podía generar respuestas más extensas y se posicionó como un modelo muy útil para la redacción de textos largos, el análisis de documentos voluminosos o las conversaciones prolongadas. En *benchmarks* como MMLU, Claude 2 alcanzó alrededor del 76% de precisión, un poco por debajo de GPT-4, pero mostrando un desempeño competitivo. Su punto fuerte era el contexto extendido y una personalidad amable y

menos propensa a desviarse: según los usuarios, Claude seguía bien el formato y ofrecía respuestas seguras, aunque a veces pecaba de excesivamente conservador, omitiendo respuestas donde GPT-4 sí se arriesgaba.

Claude 3 (marzo de 2024): —

Anthropic introdujo la familia Claude 3 con tres variantes: Claude Haiku, Claude Sonnet y Claude Opus. La idea era similar a la de otras compañías: Haiku representaba la versión ligera y rápida; Sonnet, la opción intermedia con equilibrio entre costo y capacidad; y Opus, la más potente para tareas complejas. Claude 3 amplió sus capacidades a la multimodalidad en la entrada (texto e imagen), mantuvo ventanas de contexto extensas y mejoró el rendimiento general. Durante 2024, Anthropic continuó con actualizaciones incrementales y se lanzó Claude 3.5, en la que la versión Haiku 3.5 ya superaba a la Opus 3 original en algunas pruebas, lo que evidenciaba avances rápidos. También comenzaron a experimentar con funciones de «*computer use*»: a comienzos de 2025, Claude podía controlar un entorno simulado de computadora (abrir un navegador, ejecutar código) en pruebas, anticipando las futuras capacidades de agente.

Claude 4 (mayo de 2025): —

este fue un avance importante, anunciado el 22 de mayo de 2025. Claude 4 se presentó con dos modelos principales: Claude 4 Opus y Claude 4 Sonnet. Opus 4 es el buque insignia *frontier* de Anthropic (un modelo de frontera, probablemente con el mayor tamaño y costo de entrenamiento de su historia), mientras que Sonnet 4 representa una versión algo más pequeña y eficiente para uso general. La versión Haiku no se actualizó inicialmente a 4.0, permaneciendo en 3.5 como la opción rápida. Claude 4 incorporó mejoras notables en codificación, razonamiento avanzado y capacidad para ejecutar tareas de forma autónoma. La compañía destacó que Claude 4 fue concebido como un «colaborador virtual» capaz de trabajar junto a humanos en proyectos complejos a lo largo del tiempo. Esto sugiere que el modelo fue optimizado para mantener el contexto en conversaciones o proyectos muy prolongados (posiblemente, mediante nuevas técnicas de memoria dinámica) y para alternar entre modos rápidos y modos de pensamiento profundo, según la dificultad de la consulta. De hecho, Claude 4 introdujo un modo dual de razonamiento: puede responder casi de forma instantánea en consultas simples, pero ante problemas complejos activa un modo de «pensamiento extendido» (*extended thinking*), en el que realiza *chain-of-thought* y uso de herramientas antes de responder. Este modo le permite, por ejemplo, si se le pide «búscame información sobre X y genera un informe», pausar su respuesta, hacer búsquedas en la web, leer los resultados e

incorporar esos datos a su respuesta final. Así, Claude 4 se promocionó en su lanzamiento como uno de los modelos más avanzados en cuanto a versatilidad.

Poco después, Anthropic continuó refinando sus modelos: en septiembre de 2025 lanzó Claude 4.5 (Claude 4.5 Sonnet). Esta actualización se enfocó especialmente en mejorar aún más las capacidades de programación y el uso de herramientas. De hecho, Anthropic proclamó que Claude Sonnet 4.5 era el mejor modelo de codificación del mundo, superando a cualquier otro en *benchmarks* como SWE-bench (*Software Engineering Benchmark*).

Obtuvo un 77,2 % en una prueba de OSWorld sobre uso de computadora, frente al 42,2 % que había alcanzado Claude 4 solo unos meses antes, lo que representa un salto notable en apenas cuatro meses y evidencia la rapidez de mejora. Claude 4.5 también introdujo nuevas funciones en sus productos, como la posibilidad de guardar estados de conversación (*checkpoints*), un SDK para agentes y mejoras en la memoria de contexto que le permiten trabajar hasta por 30 horas seguidas en tareas de múltiples pasos.

A diferencia de OpenAI o Google, Anthropic no lanzó inicialmente a Claude como un servicio masivo para consumidores, sino a través de empresas y socios. Sin embargo, tras la apertura de su propia interfaz web en 2023 y las mejoras posteriores, Claude ha ido ganando usuarios directos. A comienzos de 2025, el LLM de Anthropic contaba con unos 18 millones de usuarios mensuales, una cifra que lo deja bastante rezagado en términos de uso si se lo compara con los modelos de otras compañías.

Claude puede no ser el más conocido entre el público general, pero en ámbitos técnicos es respetado y se ha posicionado como uno de los modelos más avanzados

desde el punto de vista técnico. Su enfoque en la alineación ética también ha influido en otras empresas: OpenAI y Google han incorporado principios similares de IA constitucional en sus actualizaciones recientes (Intuition Labs, s. f.; Anthropic, s. f.).

3.4 Meta – LLaMA

Figura 8. LLaMA

Fuente: captura de pantalla de LLaMA (https://www.llama.com/)

Meta (antes Facebook) ha seguido un camino distinto: liberar modelos de IA «abiertos» para la comunidad. En lugar de ofrecer un *chatbot* al público general, Meta ha compartido sus LLM con investigadores y desarrolladores, permitiendo que estos los utilicen, modifiquen y desplieguen por su cuenta. El más notable es la familia LLaMA (*Large Language Model Meta AI*).

Meta lanzó LLaMA 1 en febrero de 2023 como un modelo de investigación. No se ofreció acceso abierto total, pero el modelo se filtró en internet poco después, lo que paradójicamente impulsó mucha experimentación en la comunidad. Incluso surgieron afinaciones creadas por terceros, como Alpaca o Vicuna, basadas en LLaMA.

Al observar el interés generado, en julio de 2023 Meta decidió lanzar LLaMA 2 abiertamente, esta vez con una licencia permisiva (gratuita para uso comercial bajo ciertas condiciones). LLaMA 2 se presentó con versiones preentrenadas y versiones *fine-tuned* para *chat*. Su importancia radica en que cualquier persona podía descargarlo y ejecutarlo en sus propios servidores (incluso en computadoras personales potentes, en el caso de los modelos más pequeños), lo que democratizó el acceso a LLM avanzados sin depender de una API de pago. Microsoft se alió con Meta para ofrecer LLaMA 2 en Azure Cloud y para integrarlo con herramientas como Windows.

Meta no se detuvo en LLaMA 2 y continuó investigando modelos más grandes y potentes, manteniendo su filosofía *open-source*. A mediados de 2024, surgieron informes de que Meta buscaba desarrollar una nueva generación de modelos tan capaces como GPT-4. En septiembre de 2024, Meta anunció LLaMA 3, cuya variante más grande alcanzó los 405 000 millones de parámetros (LLaMA 3.1). Este modelo 3.1 se convirtió en el modelo de IA de código abierto más grande y poderoso disponible hasta la fecha. Meta afirmó que LLaMA 3.1 superaba a GPT-3.5 en muchas tareas y se acercaba a GPT-4 en algunas métricas, todo ello siendo accesible de forma abierta para la comunidad. También se presentaron variantes más pequeñas para usos que requieren menos recursos. LLaMA 3 incorporó capacidades multimodales limitadas, mejoras en el seguimiento de instrucciones y entrenamiento en múltiples idiomas (Meta, s.f.a.).

En 2025, Meta avanzó hacia LLaMA 4. Un indicio temprano fue un artículo que mencionaba «LLaMA 4 Maverick», un experimento de Meta con un modelo de arquitectura *Mixture-of-Experts* de 17 B parámetros activos y 128 expertos, que

supuestamente superó a GPT-4 y a Gemini 2.0 Flash en ciertas tareas multimodales. Esto sugiere que Meta está explorando diseños más eficientes (utilizar expertos especializados en lugar de un único modelo gigante) para alcanzar o superar a los modelos de OpenAI y Google con menor costo. A finales de 2025, según algunas fuentes, Meta ya tenía LLaMA 4 en pruebas internas y posiblemente publicaría versiones para la comunidad en 2026 (Meta, s.f.b.).

Además de los modelos LLaMA base, Meta ha lanzado derivados especializados: Code LLaMA (agosto de 2023) para generación de código, y modelos de audio y video para comprensión multimodal, entre otros. Sus productos de consumo también han integrado IA: en septiembre de 2023, Meta introdujo en Instagram, WhatsApp y Messenger los «AI stickers» y *chatbots* de celebridades impulsados por sus modelos. Sin duda, Meta considera la IA como un elemento clave para sus redes sociales y el metaverso.

Aunque Meta no tiene un «ChatGPT» propio abierto a todo el público, el impacto de LLaMA en la comunidad es enorme. El número de descargas pasó de 650 millones a principios de diciembre de 2024 a 1000 millones en marzo de 2025, mostrando un crecimiento explosivo. En la práctica, esto significa que LLaMA es el motor detrás de innumerables aplicaciones y proyectos: desde asistentes personales en teléfonos Android rooteados, pasando por mods en videojuegos para crear NPC conversacionales, hasta servicios empresariales privados que prefieren una solución de código abierto por motivos de privacidad.

3.5 xAI – Grok

Figura 9. xAI Grok

Fuente: captura de pantalla de xAI Grok (https://askaichat.app/)

xAI es una empresa de inteligencia artificial fundada por Elon Musk en julio de 2023. Musk, quien fue cofundador de OpenAI, pero se apartó en 2018, creó xAI con la premisa de «entender la verdadera naturaleza del universo» mediante IA y con un enfoque de «búsqueda de la verdad». En la práctica, xAI busca desarrollar un modelo de lenguaje de próxima generación que compita con OpenAI y otros, pero enfatizando, según Musk, menos corrección política y más libertad de respuesta dentro del marco legal.

El primer producto de xAI es Grok, un *chatbot* conversacional cuyo nombre proviene de la jerga de ciencia ficción (significa «entender profundamente», del libro *Stranger in a Strange Land*). Aunque xAI es nueva, ha iterado rápidamente. En noviembre de 2023, xAI lanzó una vista previa de Grok para un número limitado de usuarios, inicialmente disponible para suscriptores premium de la plataforma X/Twitter. Esta versión temprana probablemente se basaba en un modelo existente modificado (se especuló que una variante de LLaMA o un GPT-3.5 licenciado, dada la rapidez de desarrollo). Grok ya mostraba integración con búsqueda en tiempo real, lo que le permitía realizar consultas en internet para obtener información actualizada al responder, de forma similar a lo que hace Bing Chat. Esto se alinea con la idea de Musk de un *chatbot* que siempre esté al día y pueda comentar noticias.

En diciembre de 2024, Grok se habilitó para usuarios gratuitos, con ciertos límites. Esto amplió la base de personas que podían probarlo más allá de los suscriptores de pago. Para diferenciar opciones, xAI introdujo distintos planes: uno gratuito con limitaciones, y opciones de pago como Premium+ y SuperGrok, que ofrecían acceso prioritario a los modelos más avanzados y un mayor número de consultas.

- **Grok 3 (mediados de 2024):** aunque no se publicitó ampliamente cada versión intermedia, xAI lanzó Grok 2 y 3 conforme mejoraban sus modelos. Probablemente, Grok 2 consistió en una mejora incremental durante la primavera de 2024, y Grok 3 a fines de 2024, aumentando el número de parámetros y afinando el modelo con más datos. Musk mencionó que Grok estaba «entrenado en el *fuegohose* de X» (el torrente de datos públicos de Twitter), aprovechando el acceso a esa enorme fuente de texto conversacional y noticias.

- **Grok 4 (julio de 2025)**: xAI anunció Grok 4 como su modelo insignia y lo puso a disposición de manera más amplia durante el verano de 2025. Grok 4 se promociona como «el modelo más inteligente del mundo» y presenta integración nativa de herramientas y búsqueda en tiempo real.

xAI también lanzó variantes como Grok 4 Fast, con menor costo y mayor rapidez, aunque algo menos preciso, para que los usuarios puedan elegir según sus necesidades.

Grok, fiel a la personalidad de Musk, intenta diferenciarse en tono. Se promociona como más divertido, dispuesto a hacer bromas o referencias a memes de internet, y menos filtrado en ciertos temas, aunque xAI aseguró que cumpliría con las leyes y no permitiría contenidos ilícitos. Esto atrajo a algunos usuarios curiosos que percibían a ChatGPT como demasiado «censurado» en ciertas respuestas. No obstante, un riesgo de esta libertad es que Grok podría generar contenido controvertido; hubo reportes de exempleados preocupados por la poca moderación en algunos casos, por lo que xAI tendría que equilibrar esto para el público general.

Grok comenzó de forma cerrada, pero en 2025 se liberó globalmente de manera gratuita, lo que impulsó su adopción. Para septiembre de 2025, se estimaba que Grok alcanzaba alrededor de 30 millones de usuarios activos mensuales. Para aumentar la base, xAI integró Grok dentro de X (Twitter) como un *bot* al que los usuarios pueden consultar directamente en la plataforma. También lanzaron Grok Companions, personajes con avatar 3D que interactúan con el usuario (similar a los *bots* de celebridades de Meta), para atraer al público más joven.

CONTINUAR

4. Comparativa de los modelos de vanguardia

Como vimos anteriormente, la mayoría de las empresas promocionan su propio modelo como el más avanzado del mercado, sin embargo, para probar esto existen una serie de pruebas estandarizadas, conocidas como *benchmarks*, para medir y comparar el rendimiento de los modelos en diversas tareas. De esta forma podemos tener una apreciación cuantitativa del rendimiento de los distintos modelos en diferentes situaciones de uso.

Tabla 1. Comparación de modelos

Métrica/*Benchmark*	**Gemini 2.5 Pro**	**GPT-5**	**Claude 4.5**	**Grok 4**
MMLU (conocimiento general)	~86-91 %	**91.4 %**	~87-89 %	No disponible
GPQA (razonamiento lógico)	86.4 %	**~88-88.4 %**	~83-84 %	87.5 % (Grok 4) / 85.7 % (Grok 4 Fast)
AIME 2025 (matemáticas)	~88 %	**94.6 %**	87 %	92.0 % (Grok 4 Fast)
SWE-Bench (codificación)	~67-69 %	74.9 %	**77.2 %**	~75 %

MMMU (multimodalidad)	~82 %	**84.2 %**	~78 %	No disponible
Ventana de contexto	**1 millón de tokens**	~400K+ tokens	200K tokens	256K *tokens*

Fuente: elaboración propia con base en Kavukcuoglu, 2025.

Como muestra la tabla, las capacidades de los distintos modelos varían según la tarea y el uso que se les dé. Por esta razón, se recomienda que, al momento de elegir un modelo, se realicen pruebas basadas en los requerimientos personales de cada usuario, con el fin de determinar cuál se adapta mejor a sus necesidades.

Nota aclaratoria sobre uso de IA

Este material fue asistido con herramientas de IA generativa para tareas de borrador, síntesis, reescritura y apoyo en la organización de contenidos. Cada sección fue revisada, editada y validada por el equipo humano, que verificó la precisión conceptual, la coherencia pedagógica y las fuentes citadas. Se invita a contrastar con las referencias bibliográficas incluidas y la documentación oficial. Dado que los modelos de IA evolucionan con rapidez, ciertas especificaciones técnicas podrían actualizarse; este texto refleja el estado del conocimiento al momento de su elaboración.

CONTINUAR

Referencias

Anthropic. (s. f.). *Claude Sonnet 4.5 (anuncio y especificaciones)*. https://www.anthropic.com/news/claude-sonnet-4-5

Exploding Topics. (s. f.). *ChatGPT users: estadísticas y uso*. https://explodingtopics.com/blog/chatgpt-users#:~:text=,5%20billion%20prompts%20each%20day

IBM. (s.f.). *What is artificial general intelligence (AGI)?* IBM Think. https://www.ibm.com/think/topics/artificial-general-intelligence ibm.com+1

Intuition Labs. (s. f.). *Anthropic Claude 4: LLM evolution*. https://intuitionlabs.ai/articles/anthropic-claude-4-llm-evolution

Kavukcuoglu, K. (2025). Gemini 2.5: Our most intelligent AI model. Google DeepMind Blog blog.google.com

Kotwani, R. (2018). But what is a GPT? Visual intro to Transformers. *Medium*. https://medium.com/lazy-by-design/but-what-is-a-gpt-visual-intro-to-transformers-3blue1brown-d078447b8ef4

Meetanshi. (s. f.). *Google Gemini statistics*. https://meetanshi.com/blog/google-gemini-statistics/#:~:text=3,They

Meta AI. (s.f.a.). *Introducing Llama 3.1*. https://ai.meta.com/blog/meta-llama-3-1/#:~:text=Introducing%20Llama%203,capable%20openly%20available%20foundation%20model

Meta AI. (s.f.b.). *Llama 4: Multimodal intelligence.* https://ai.meta.com/blog/llama-4-multimodal-intelligence/#:~:text=AI%20ai,4o%20and%20Gemini%202.0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need.* https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

CONTINUAR