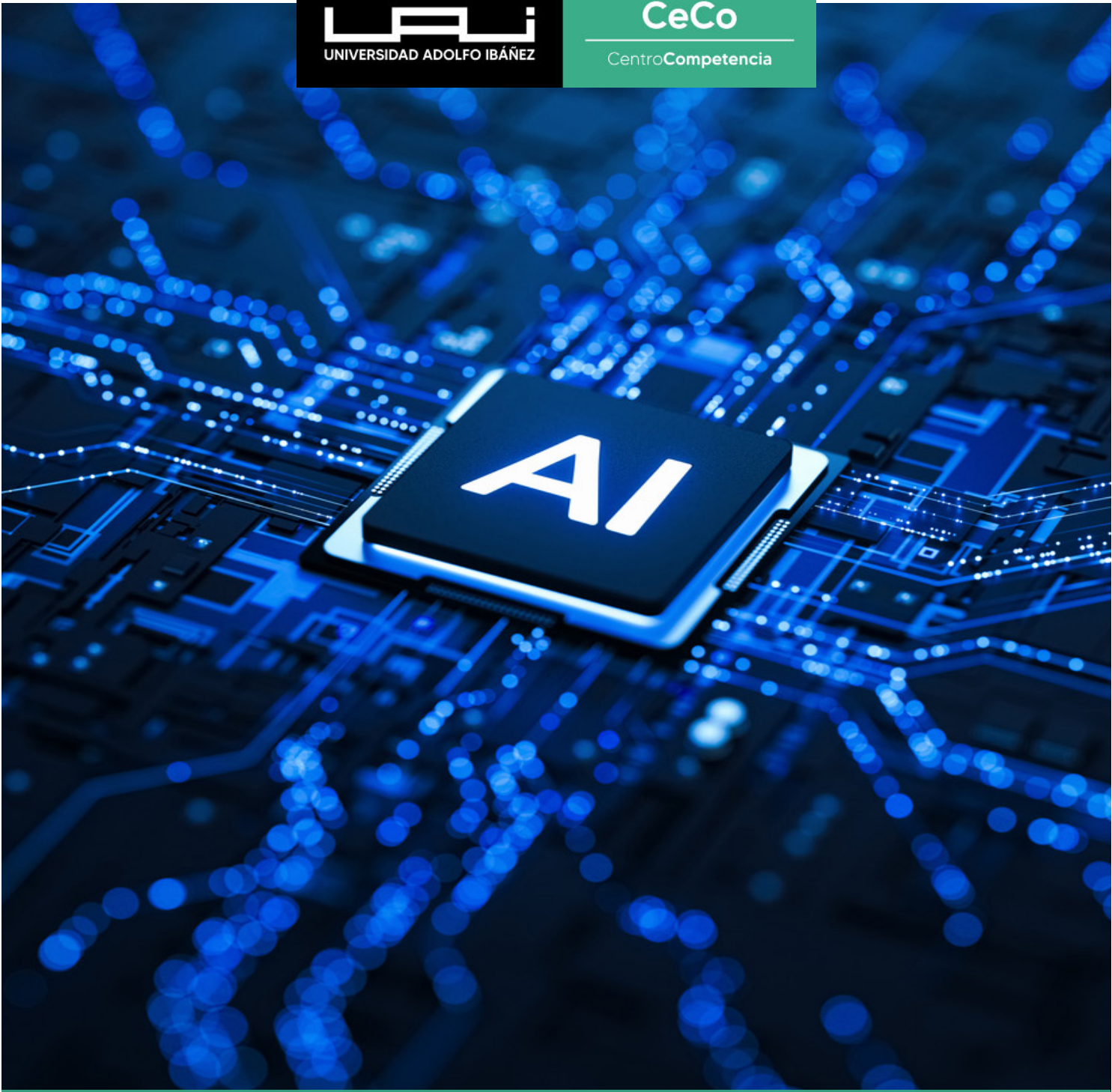


A close-up, high-angle shot of a square, dark-colored microchip with the letters 'AI' in white. The chip is resting on a complex circuit board. The background is a dense network of glowing blue lines and dots, resembling a digital or neural network, creating a bokeh effect. The overall color palette is dominated by deep blues and greens.

AI

MERCADOS EN FORMACIÓN: COMPETENCIA EN LA INDUSTRIA DE LA IA GENERATIVA

Iván Paredes

MERCADOS EN FORMACIÓN: COMPETENCIA EN LA INDUSTRIA DE LA IA GENERATIVA

Agosto 2026



Iván Paredes

Economista (UNAM, CIDE), científico de datos (ITAM), con 12 años de experiencia en competencia, regulación, y análisis de datos. He trabajado en la autoridad de competencia de México, y actualmente en el organismo regulador de Telecom.

Abstract: Este trabajo analiza la evolución de la AI generativa, tipologías y cadena de valor, destacando el papel central de los modelos de lenguaje (LLMs) como capa de orquestación de sistemas multimodales capaces de generar texto, imágenes, audio y video.

A partir de una revisión estructural del sector, se identifican los principales eslabones de la industria, infraestructura de cómputo, datos, modelos, middleware y aplicaciones. El estudio muestra que el mercado presenta una alta concentración inicial aparente, impulsada por barreras de entrada como los elevados costos de entrenamiento. Sin embargo, los modelos abiertos generan una oportunidad de dar contestabilidad al mercado.

I. PANORAMA DE LA AI GENERATIVA

1. Introducción

La inteligencia artificial generativa es una de las tecnologías de mayor impacto económico de la última década. Los modelos avanzados de lenguaje y sistemas multimodales impactan sectores tan diversos como servicios financieros, educación, salud, medios de comunicación, desarrollo de software y administración pública. Su aplicación ha generado expectativas comparables a las asociadas con la difusión de internet o la computación en la nube.

La creación de estos modelos requiere grandes inversiones en infraestructura computacional, datos, talento especializado y capacidades de investigación, lo que ha favorecido la aparición de un reducido grupo de empresas con una posición estratégica en distintos eslabones de la cadena de valor. Al mismo tiempo, hay una lucha geopolítica que ha permitido el desarrollo de modelos de código abierto, no restrictivos en China.

En este contexto, el presente trabajo analiza la evolución de la inteligencia artificial generativa, sus principales tipologías y la estructura de su cadena de valor. A partir de una revisión de los distintos eslabones de la industria se examinan las barreras de entrada, los factores que favorecen la concentración y los mecanismos que pueden impulsar la competencia.

Este estudio argumenta que, aunque el mercado presenta signos de alta concentración inicial derivados de economías de escala, efectos de red e integración vertical, la rápida innovación tecnológica y la expansión de modelos abiertos continúan generando oportunidades para la entrada de nuevos participantes y la contestabilidad del mercado.

Así pues, si bien la competencia en el mercado de los modelos de inteligencia generativa requiere vigilancia y análisis constantes por parte de las autoridades de competencia y supervisión, el verdadero problema, sin embargo, seguramente es lo que va a generar en el mercado laboral, con profesiones que desaparecerán y el hecho de que cada vez se va a requerir menos mano de obra. Tema, sobre el cual no profundizaremos en este documento.

2. Clasificación por output

La inteligencia artificial generativa (IA generativa) es la creación de contenido “nuevo” a partir de patrones aprendidos en grandes volúmenes de datos. Los modelos generativos pueden producir “nuevo” texto, imágenes, audio, video o código (Oluwagbenro, 2024).

Los modelos de lenguaje masivo (LLMs por sus siglas en inglés) son los más conocidos dentro de la IA generativa, pero no son el único tipo (Wheeler, 2025). En paralelo, se han desarrollado otros tipos de modelos especializados en distintos tipos de datos y tareas. Por ejemplo:

1. Los modelos de difusión¹, empleados para la generación de imágenes, video y audio. Estos modelos funcionan aprendiendo a transformar datos aleatorios en contenido coherente y realista.²

¹ También llamados denoising diffusion probabilistic models — DDPMs — o score-based generative models. Stable Diffusion 3.5". Stability AI. Archived from the original on October 23, 2024. Retrieved October 23, 2024.

² Los modelos de difusión funcionan destruyendo datos de entrenamiento mediante la adición sucesiva de ruido gaussiano y luego aprendiendo a recuperar los datos revirtiendo este proceso de ruido.

Entre los ejemplos más influyentes se encuentran Stable Diffusion (O'Connor, 2022), NanoBanana de Google³, Runway Gen-2⁴ y Sora de OpenAI⁵.

2. Las Redes Generativas Antagónicas (GANs)⁶ han sido ampliamente utilizadas en la creación de rostros realistas conocidos como deepfakes. StyleGAN es un ejemplo de este tipo de modelos (Fu, J. et al., 2022).

3. Los modelos de audio y música han avanzado en la generación de voces sintéticas y composiciones musicales "originales". Herramientas como Suno⁷, Jukebox de OpenAI⁸ y MusicLM⁹ de Google permiten producir canciones completas. The Velvet Sundown es un grupo de IA que generó millones de reproducciones usando este tipo de aplicaciones (Bakare, 2025).

4. Finalmente, emergen los modelos multimodales, capaces de integrar distintas entradas y salidas como texto, imagen, voz y video en un mismo sistema. Ejemplos de ello son GPT-5.2¹⁰, Gemini 1.5¹¹ y Kosmos-1¹², que representan la convergencia de la inteligencia artificial generativa hacia asistentes más generales y versátiles, capaces de interactuar con los usuarios en múltiples formatos y contextos. En la mayoría de los casos, siguen siendo diferentes modelos accesibles desde un solo punto de acceso (Bubeck et al., 2023).

No obstante, de todos los modelos, los LLMs son el centro de la IA generativa, pues al ser el lenguaje natural son el medio más flexible de interacción: se usan para escribir, traducir, programar, responder preguntas, resumir documentos y coordinar otras herramientas (Wang et al., 2025). Así, los LLMs se han convertido en la "capa de orquestación" que coordina tareas multimodales.

3. Clasificación por acceso y propiedad

Los modelos también se pueden clasificar según su propiedad y acceso. No obstante, la idea de lo que es *open-source* en AI Generativa es un debate en construcción, a diferencia de lo que pasa en la industria del software (Nobel et al., 2025). El debate gira en torno a la apertura de distintos componentes técnicos y documentales: código fuente, parámetros del modelo (*weights*, o pesos), datos de entrenamiento, documentación y licencias. Algunos modelos son totalmente privados, otros tienen acceso libre al código, pero no a los parámetros ni a los datos de entrenamiento, y solo los más abiertos tienen acceso a todo, o casi todo.

La disponibilidad del código fuente y de los parámetros entrenados permite comprender cómo funciona el modelo y modificarlo. En contraste, la falta de acceso a los parámetros entrenados, incluso con acceso al código, dificulta el uso del modelo. Por ello, la apertura de los pesos es indispensable. Los pesos o parámetros de un modelo de frontera pueden pesar entre 100 GB y más de 1 TB.¹³

3 Google DeepMind. (2025). Gemini 2.5 Flash Image. Google AI Studio. <https://aistudio.google.com/models/gemini-2-5-flash-image>.

4 Runway. (s. f.). Runway. <https://app.runwayml.com/>.

5 OpenAI. (2024, septiembre 12). Video generation models as world simulators. OpenAI. <https://openai.com/index/video-generation-models-as-world-simulators>.

6 Una red generativa antagónica (GAN) es una arquitectura de aprendizaje profundo. Entrena dos redes neuronales de modo que compiten entre sí para generar nuevos datos más auténticos a partir de un conjunto de datos de entrenamiento determinado. (Oluwagbenro, 2024).

7 Suno. (s. f.). Suno AI. Recuperado 17 de septiembre de 2025, de <https://www.suno.ai>.

8 OpenAI. (2020). Jukebox. OpenAI. <https://openai.com/index/jukebox/>.

9 Google AI. (2023). MusicLM: Generating music from text. Google Research. <https://google-research.github.io/seanet/musiclm/examples/>.

10 OpenAI. (2024). *GPT-4o* (Version 4o) [Large language model]. <https://openai.com/index/hello-gpt-4o/>.

11 Google DeepMind. (2024). Gemini 1.5 [Large language model]. <https://blog.google/technology/ai/gemini-1-5/>.

12 Microsoft. (2023). *Kosmos-1* [Multimodal large language model]. <https://www.microsoft.com/en-us/research/blog/kosmos-1/>.

13 Mistral AI. (s. f.). Model weights. Mistral AI Documentation. Recuperado el 26 de septiembre de 2025, de <https://docs.mistral.ai/getting-started/models/weights/Mistral-AI-Documentation>.

Finalmente, respecto a los datos de entrenamiento, Nobel et al. (2025) sostienen que es necesario publicar los datos completos para garantizar auditabilidad y reproducibilidad, mientras que otros argumentan que basta con ofrecer metadatos sobre su origen y características. De todas maneras, existen obstáculos a la hora de compartir los datos, específicamente relacionados con la protección de datos personales, las licencias de contenido y los costos técnicos de almacenamiento y documentación (Nobel et al., 2025).

El aspecto jurídico también es central. Existen modelos que liberan código y pesos, pero bajo licencias restrictivas que limitan el uso comercial, la redistribución o la modificación. Entre ellos se encuentran, por ejemplo, LLaMA 2, LLaMA 3.1, y LLaMA 3.2 (todos de Meta). Este fenómeno ha sido denominado *open-washing*, ya que aparenta apertura sin garantizar las libertades asociadas al software libre (TechTurismo Media, 2024).

En un extremo se encuentran los modelos cerrados (*closed-source*), como GPT-4 de OpenAI o Claude de Anthropic, cuyo código¹⁴, pesos¹⁵ y datos de entrenamiento¹⁶ no son públicos. Estos modelos se ofrecen bajo esquemas comerciales de acceso restringido mediante *APIs* o integraciones en productos, lo que permite a las empresas mantener ventajas competitivas y un mayor control sobre su uso. En el otro extremo están los modelos abiertos (*open-source*), como Deepseek R.1, Qwen3 o Falcon¹⁷, cuyos pesos y, en algunos casos, el código de entrenamiento se libera al público.

4. Clasificación por número de parámetros

Otra forma de clasificar los modelos es de acuerdo con el número de parámetros y el uso de recursos. El número de parámetros está relacionado directamente con las capacidades de procesamiento e “inteligencia” de los modelos. Algunos modelos están diseñados para operar con cientos de miles de millones de parámetros, requieren infraestructura de cómputo de primer nivel y son capaces de ejecutar tareas altamente complejas, incluyendo razonamiento multimodal (Bowdon, 2025).

En contraste, modelos más pequeños, de entre 3 y 20 mil millones de parámetros, están optimizados para la eficiencia, pueden ejecutarse en servidores estándar e incluso en hardware local, y responden a la necesidad de desplegar IA en contextos con limitaciones de recursos. Esta distinción entre “modelos gigantes o de frontera” y “modelos ligeros” refleja un trade-off entre poder predictivo y viabilidad de uso en entornos cotidianos (Johnson, 2025).

La elección entre ejecutar un modelo avanzado como DeepSeek R.1 con 236 mil millones de parámetros y uno más ligero como Llama 3.2 3B de 3 mil millones de parámetros implica diferencias significativas en los requisitos de hardware. Llama 3.2, al ser un modelo pequeño y optimizado, puede ejecutarse en hardware de uso personal, como una GPU con 8 GB de VRAM o una CPU con suficiente RAM. Por su parte, DeepSeek R.1, al ser un modelo más grande y complejo, requiere recursos considerablemente superiores, como múltiples GPUs de alto rendimiento (80 GB de VRAM cada una) o acceso a clústeres de servidores en la nube con optimizaciones específicas.

14 El conjunto de instrucciones y algoritmos escritos en lenguajes de programación (como Python) que definen la arquitectura y la lógica del modelo.

15 Los parámetros numéricos internos del modelo que se ajustan durante el entrenamiento. Representan la “fuerza” de las conexiones entre las neuronas artificiales de la red.

16 El inmenso conjunto de texto (y a veces código) con el que se alimenta al modelo durante su fase de entrenamiento.

17 Technology Innovation Institute. (2023). *Falcon-40B* [Large language model]. Hugging Face. <https://huggingface.co/tiiuae/falcon-40b>.

Tabla 1. Diferencias entre modelos de frontera y modelos ligeros

Característica	LLM Avanzado (ej. GPT-4, Llama-3 70B, Mistral Large, >100B parámetros)	LLM Ligero (ej. Llama-3.2 3B, DS distilled <10B parámetros)
Parámetros	100B – 1,500B	7B – 13B
Memoria GPU necesaria (FP16)	~800 GB – 1.6 TB	~8 – 48 GB
GPUs para entrenamiento	2,000 – 25,000 GPUs A100/H100	64 – 256 GPUs, A100/H100
Tiempo de entrenamiento	3–6 meses (full training)	2–4 semanas
Costo de entrenamiento (USD)	\$50M – \$1000M	\$2M – \$5M
Inferencia en <i>batch</i> (GPU RAM mínima)	8×H100 (80GB) o 16×A100 (40GB)	1×A100 (40GB) o incluso 1×L4/T4 con cuantización ¹⁸
Costo de inferencia (por millón de tokens, nube)	\$2 – \$5 USD (OpenAI, Anthropic)	\$0.10 – \$0.50 USD
Velocidad de inferencia	15–40 tokens/s por GPU	50–150 tokens/s en una GPU
Casos de uso típicos	Razonamiento complejo, multi-agente, multimodalidad	Chatbots empresariales, RAG, prototipos.
Infraestructura necesaria	Clúster HPC con interconexión NVLink/InfiniBand, consumo >5 MW	Servidor único con 1–4 GPUs, factible on-prem o en cloud

Fuente: Elaboración propia con base en Meta (2024), NVIDIA (2024), Epoch AI (2025), Stanford AI Index (2025), OpenAI (2026), Anthropic (2026) y SemiAnalysis (2024). Los valores corresponden a órdenes de magnitud representativos.

Una última clasificación relevante proviene de su especialización funcional. Algunos modelos son generalistas, entrenados con grandes cantidades de datos heterogéneos para desempeñar múltiples tareas de lenguaje, mientras que otros son especializados o afinados en dominios concretos. Ejemplos de estos últimos incluyen modelos entrenados para programación (Codex, Code LLaMA), para razonamiento científico y matemático (Galactica, DeepSeek Math), o para aplicaciones legales y médicas (Anisuzzaman et al., 2025). No obstante, esta clasificación es menos relevante, toda vez que el reentrenamiento de un modelo puede realizarse relativamente rápido y con recursos disponibles para empresas pequeñas y medianas.¹⁹

5. Evolución de la AI generativa

De acuerdo con Vipra y Korinek (2023), el desarrollo de la inteligencia artificial generativa deriva de avances acumulativos en tres grandes áreas: modelos probabilísticos, disponibilidad de datos y capacidad de procesamiento computacional. En 2017, la publicación de *Attention Is All You Need* introdujo la arquitectura *Transformer*, un diseño de redes neuronales basado en mecanismos de atención que permite procesar grandes volúmenes de texto de manera más eficiente y capturar relaciones de largo alcance entre las palabras. Esta innovación sentó las bases para el desarrollo de los modelos de lenguaje de gran escala (LLMs).

18 IBM. (s. f.). ¿Qué es la cuantificación? IBM Think. <https://www.ibm.com/mx-es/think/topics/quantization>.
19 Un interesante ejemplo es la AI Matria desarrollada en México. Celestial Dynamics. (2025). *Matria AI: Liderando la nueva era de modelos de inteligencia artificial en México*. Matria AI. <https://matria.ai/>.

Posteriormente, modelos como BERT (Google, 2018) y GPT-2 (OpenAI, 2019) demostraron el potencial de esta arquitectura para comprender y generar lenguaje natural, al tiempo que evidenciaron que incrementar el número de parámetros de los modelos, junto con mayores volúmenes de datos y capacidad de cómputo, podía traducirse en mejoras sustanciales de su desempeño.²⁰

El lanzamiento de GPT-3 (OpenAI, 2020) con 175 mil millones de parámetros marcó un hito en la generación de texto fluido. Poco después, se popularizaron modelos multimodales como CLIP y DALL·E (OpenAI, 2021), así como *Stable Diffusion* (Stability AI, 2022), que democratizaron la generación de imágenes. Estos modelos volvieron a la IA generativa un fenómeno social y económico, integrándose en buscadores, redes sociales y plataformas de comercio.

6. Evaluación de los modelos de AI Generativa

La evaluación de los modelos de inteligencia artificial generativa es compleja, pues requiere analizar simultáneamente diversas aristas (Ramamoorthy, 2025). Por ello, han surgido múltiples enfoques que combinan métricas automáticas, pruebas de desempeño estandarizadas y evaluaciones humanas.

En el caso de los LLMs, la evaluación suele realizarse con *benchmarks* estandarizados como *Massive Multitask Language Understanding* (MMLU), *HellaSwag*, *TruthfulQA* o *HumanEval*, que miden la capacidad del modelo para responder preguntas de razonamiento, mantener la coherencia, evitar sesgos y producir código funcional.²¹

Para los modelos de generación de imágenes y video, se emplean métricas como el *Inception Score* (IS)²² o el *Fréchet Inception Distance* (FID)²³, que comparan la distribución estadística de las imágenes generadas frente a conjuntos de referencia. En el ámbito de la generación de música y audio, la evaluación combina métricas automáticas (como similitud espectral o diversidad de patrones sonoros) con estudios de percepción realizados a través de *tests* de escucha, en los que los participantes valoran naturalidad, expresividad o adecuación estilística.

No obstante, en muchos casos estas pruebas deben complementarse con evaluaciones humanas. La percepción estética, la originalidad y la adecuación al *prompt* no siempre se capturan de manera objetiva y requieren validación humana.

Plataformas como *LMarena.ai*, que recogen y sistematizan evaluaciones directas de los usuarios a través de comparaciones de respuestas de distintos modelos, representan una aproximación relevante y complementaria.²⁴ Este tipo de evaluación colectiva permite incorporar la percepción humana en gran escala,

aportando información sobre calidad percibida, adecuación al contexto y utilidad práctica, dimensiones que difícilmente pueden medirse con métricas automáticas tradicionales.

20 Our World in Data. (2023). *Exponential growth of parameters in notable AI systems* [Chart]. Retrieved from <https://ourworldindata.org/grapher/exponential-growth-of-parameters-in-notable-ai-systems>.

21 Ibid.

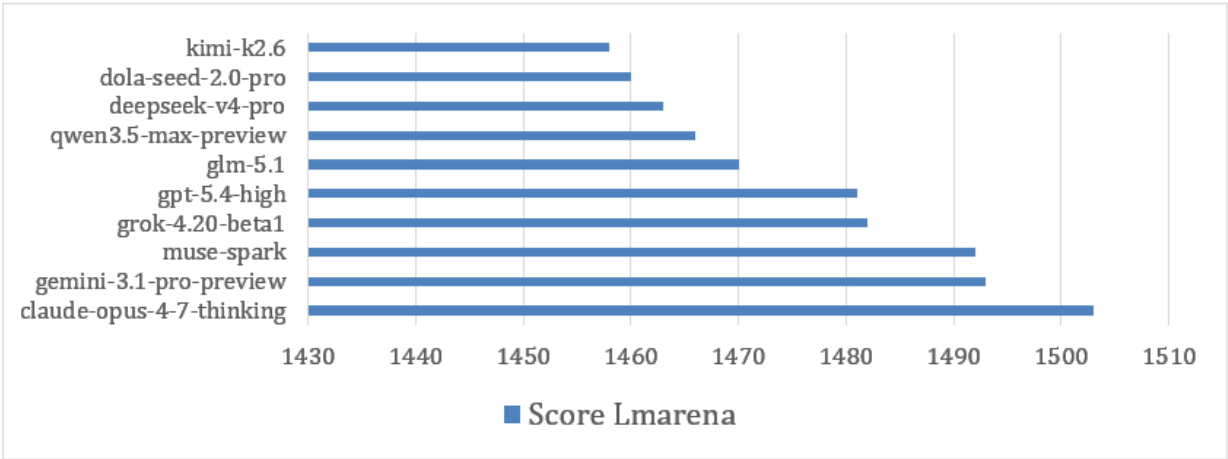
22 El Inception Score (IS) fue uno de los primeros indicadores diseñados para evaluar la calidad de imágenes generadas. Su lógica es doble: 1. Si la imagen es clara y realista, la red debería asignarle una distribución de probabilidad muy concentrada en una sola clase (es decir, alta entropía negativa) y 2. Si el modelo solo genera imágenes de una o dos clases, la distribución estará concentrada y entonces hay baja diversidad.

23 El Fréchet Inception Distance (FID) es más moderno y hoy es mucho más usado. Se basa en comparar distribuciones estadísticas entre imágenes reales y generadas. Un FID bajo significa que las imágenes generadas tienen distribución de características cercana a la de las imágenes reales.

24 LMArena. (s. f.). LMArena. <https://lmarena.ai/>.

De acuerdo a Lmarena, (a diciembre de 2025) Gemini-3-Pro es el modelo más eficiente y con mejor desempeño ante usuarios. Este *ranking* cambia constantemente. Hace apenas un par de meses Gpt-5.1-High estaba en el primer lugar. Además, las diferencias entre los modelos son marginales.

Gráfica 1. Score humano de calidad del modelo para los 10 modelos más avanzados de abril 2025



Fuente: Elaboración propia con datos en Lmarena (<https://arena.ai/>).

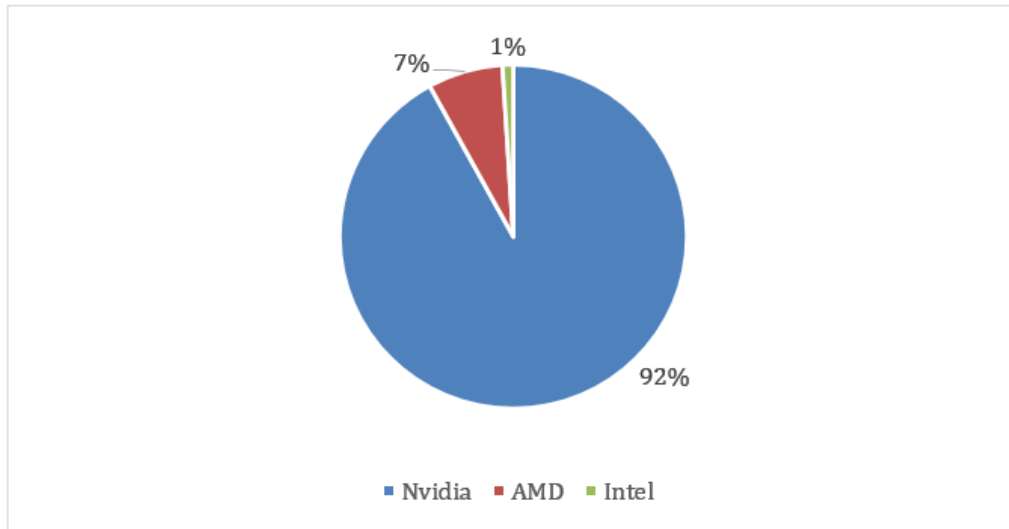
II. LA CADENA DE VALOR DE LA IA GENERATIVA

La cadena de valor de la inteligencia artificial generativa se estructura en múltiples eslabones interdependientes, que van desde la infraestructura física hasta las aplicaciones finales orientadas al usuario.

1. Unidad de Procesamiento Gráfico (GPUs)

El primer eslabón de la cadena productiva de la IA generativa es la infraestructura física: chips especializados (GPUs, TPUs, ASICs) y centros de datos. En el caso de las empresas de chips, el mercado está dominado por una empresa: NVIDIA. En el mercado también operan AMD e Intel, aunque de manera marginal.

Gráfica 2. Participación de mercado en 2024 en millones de chips comercializados por las empresas de GPUs

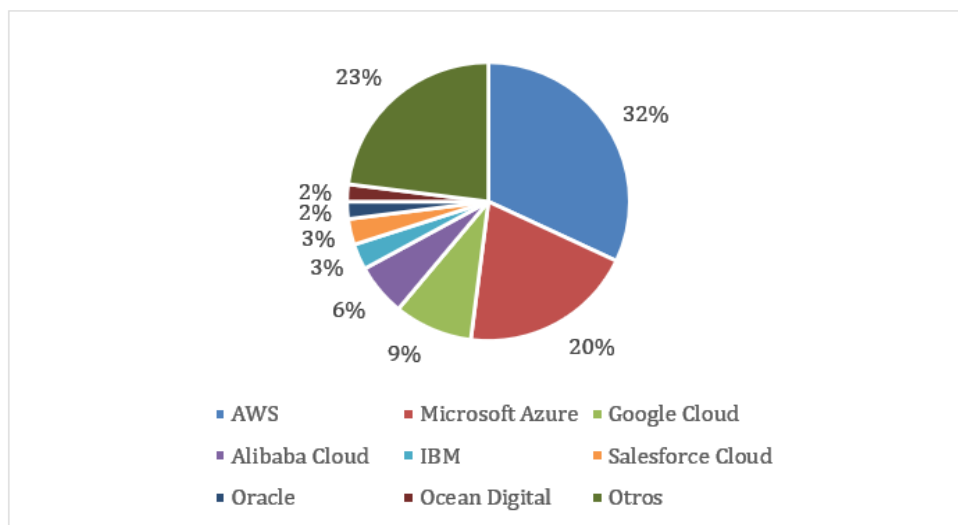


Fuente: Elaboración propia con datos de Yahoo Finance. (Agosto, 2024).²⁵

2. Centro de datos/Procesamiento

En lo referente al segmento de centros de datos, Google, Amazon, Azure y Alibaba dominan el segmento de centros de procesamiento, aunque aún existe una serie de pequeños jugadores locales que siguen operando en segmentos de nicho.²⁶

Gráfica 3. Participación de mercado en 2024 por ingresos las empresas de cómputo en la nube



Fuente: Elaboración propia con datos de Emma. (2025, enero 14).²⁷

²⁵ Yahoo Finance. (2024, August 14). Nvidia secures 92% of the GPU market. Yahoo Finance. <https://finance.yahoo.com/news/nvidia-secures-92-gpu-market-150444612.html>.

²⁶ Emma. (2025, enero 14). Cloud Market Share Trends to Watch in 2025. emma Blog. <https://www.emma.ms/blog/cloud-market-share-trends>.

²⁷ Nguyen, K. (2025, mayo 26). AWS vs Azure vs GCP vs Alibaba Cloud: Which cloud platform is best for your business? Niteco. <https://niteco.com/articles/choosing-cloud-platform-aws-azure-gcp-alibaba/>.

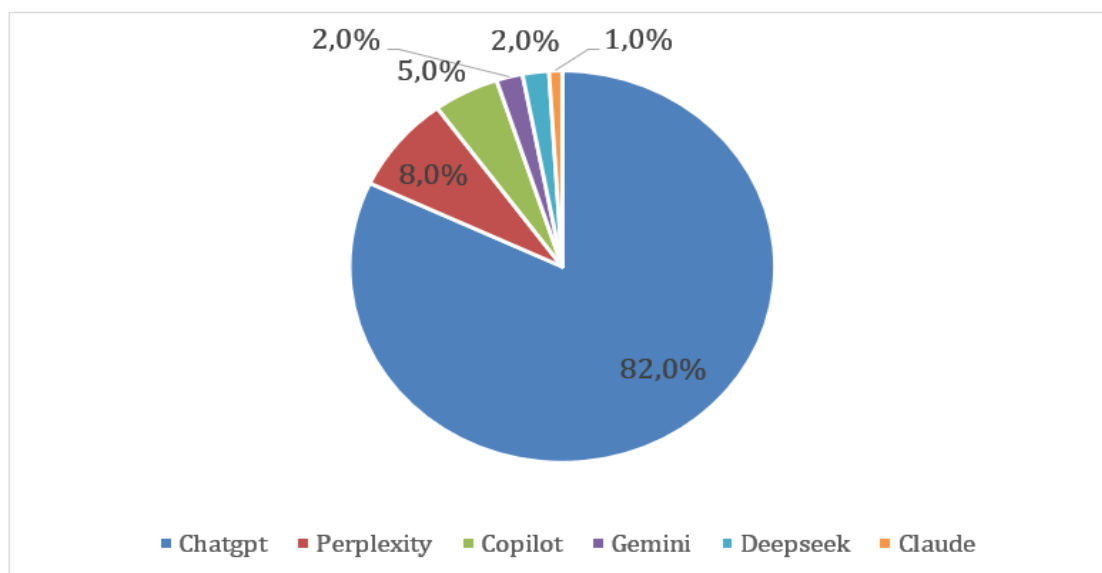
3. Etiquetado de datos

Un eslabón fundamental y con frecuencia subestimado en la cadena de valor de la IA generativa es el de las empresas especializadas en la curación, etiquetado y venta de datos de alta calidad. Estas compañías proveen conjuntos de datos estructurados y anotados que son indispensables para el entrenamiento, ajuste fino y evaluación de modelos de lenguaje, visión por computadora y multimodalidad. Empresas como Scale AI; Appen; Labelbox; y Hugging Face, son actores clave en este nivel: además de su rol en middleware (software que funciona como una capa intermedia entre dos sistemas, aplicaciones, servicios o bases de datos para que puedan comunicarse), operan como repositorios y distribuidores de *datasets* abiertos y especializados. En este nivel no se cuenta con información sobre participaciones de mercado. No obstante, hay numerosas empresas que operan en el mercado (MarketsandMarkets, 2023). Incluso algunas empresas como OpenAI y Meta tienen sus propios departamentos de etiquetado de datos (Kumar & Sharma, 2022). Algunas empresas de LLM incluso usan datos generados por sus propios modelos.

4. Modelos fundacionales

Posteriormente, en el eslabón de los modelos (LLMs, VLMs), Microsoft, OpenAI, Anthropic, Deepseek, Google, Qwen y Meta dominan el mercado. Estos modelos, entrenados con enormes recursos de datos y cómputo, corresponden a la última etapa de la cadena productiva y son los modelos contratados por los usuarios finales.

Gráfica 4. Participación de mercado en 2024 por ingresos de modelos de LLM²⁸

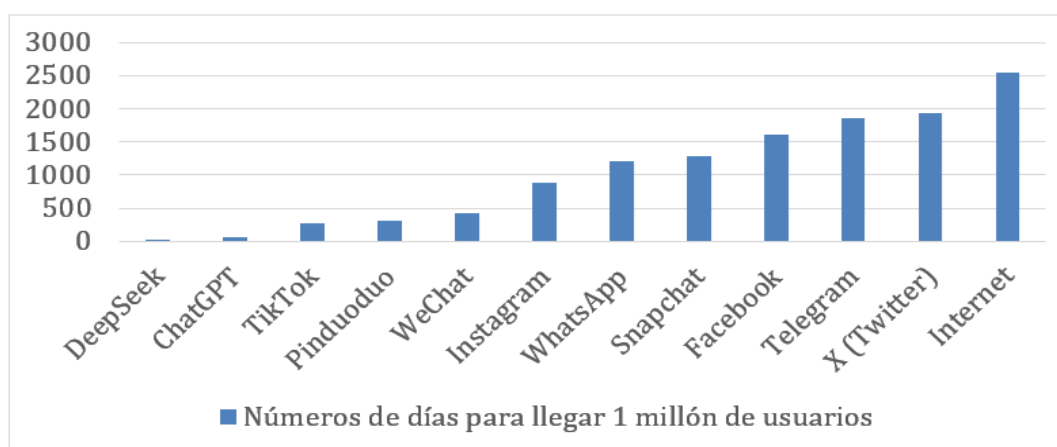


Fuente: Elaboración propia con datos de IoT Analytics. (2024). *Leading generative AI companies and platforms in 2024*. <https://iot-analytics.com/leading-generative-ai-companies/>.

²⁸ IoT Analytics. (2024). *Leading generative AI companies and platforms in 2024*. <https://iot-analytics.com/leading-generative-ai-companies/>.

La determinación de la participación de mercado de estas empresas puede resultar engañosa. Una alta participación de Chatgpt no equivale a un dominio del mercado, quizá solo temporalmente. La aparición de un nuevo modelo, como ocurrió con Deepseek en 2025 o con Claude en 2026, supone cambios fuertes en la estructura del mercado. Por ejemplo, mientras que Chatgpt tardó 2 meses en alcanzar el millón de usuarios, Deepseek lo hizo en 7 días. Los LLM no generan una fuerte retención del cliente al mismo tiempo que tienen la capacidad de ser multihoming.

Gráfica 5. Número de días que le tardo a cada modelo alcanzar 1 millón de usuarios²⁹



Fuente: Elaboración propia con datos de Chen Long, Zheng Kaiwen & Shi Sen. (2025, 6 marzo). DeepSeek's Chinese-style Innovation: A New Form of AI Economics (Part 1). CKGSB Knowledge. https://english.ckgsb.edu.cn/knowledge/professor_analysis/deepseek-redefining-global-ai-trends/.

5. Middleware y herramientas de integración - Automatización y Orquestación

Entre los modelos fundacionales y las aplicaciones finales existe un conjunto de plataformas y herramientas conocido como middleware. Su función es simplificar el desarrollo de aplicaciones basadas en inteligencia artificial, proporcionando componentes reutilizables para conectar modelos de lenguaje con bases de datos, sistemas empresariales, APIs y otras aplicaciones. En lugar de que cada empresa tenga que desarrollar desde cero la infraestructura necesaria para utilizar un modelo de IA, el middleware ofrece bibliotecas, servicios y entornos que reducen significativamente el tiempo, costo y complejidad del desarrollo. Aquí, plataformas como Hugging Face y frameworks como LangChain juegan un rol fundamental:

- Hugging Face Hub funciona como un repositorio central para modelos, datasets y demos, actuando como el "GitHub de la IA" (Metz, 2023). Esta plataforma facilita el acceso a modelos preentrenados para desarrolladores e investigadores. Además, con HUGS (Hugging Face Generative AI Services), es posible desplegar modelos abiertos de forma rápida y optimizada en hardware como GPUs, reduciendo costos y complejidad técnica.
- LangChain es un framework de código abierto que proporciona componentes para construir aplicaciones basadas en modelos de lenguaje. Entre otras funciones, permite gestionar prompts, conectar modelos con bases de datos y APIs, desarrollar agentes inteligentes y orquestar flujos de trabajo complejos.

²⁹ Chen Long, Zheng Kaiwen & Shi Sen. (2025, 6 marzo). DeepSeek's Chinese-style Innovation: A New Form of AI Economics (Part 1). CKGSB Knowledge. https://english.ckgsb.edu.cn/knowledge/professor_analysis/deepseek-redefining-global-ai-trends/.

Las opciones son variadas y hay indicios de que la participación de mercado en este segmento está repartida entre numerosas empresas (McKinsey, 2023).

Asimismo, en la parte de automatización y orquestación, existen herramientas como Zappier³⁰, n8n³¹ y Microsoft Power Automate³², una plataforma de automatización que permite integrar modelos generativos en procesos empresariales sin escribir código complejo.

Finalmente, en el último eslabón nos encontramos con las aplicaciones finales y sectoriales. Aplicaciones como asistentes virtuales, herramientas de productividad, legaltech, edtech, fintech, marketing, y generación de imágenes o video materializan el valor final.³³

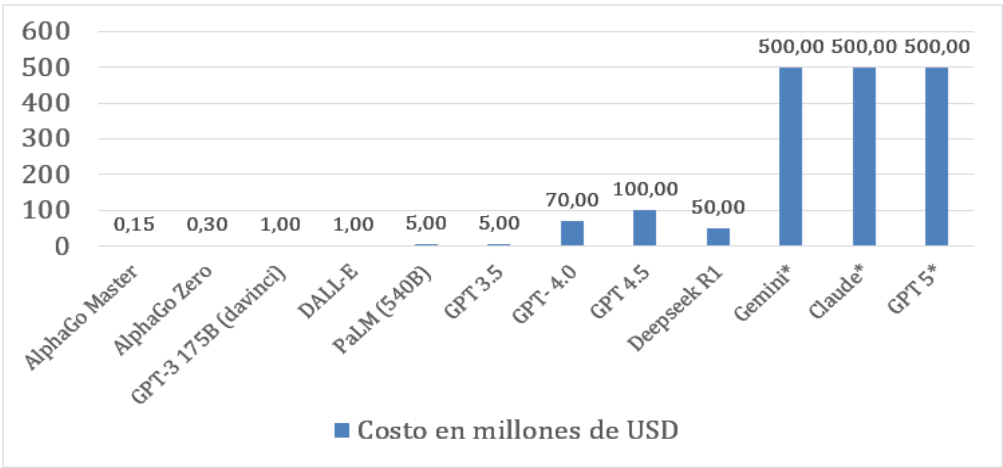
III. PANORAMA COMPETITIVO DE LA AI

1. Barreras competitivas

El mercado de inteligencia artificial actual está en una fase de alta concentración inicial, derivada de las fuertes barreras de entrada. Estas barreras responden a una serie de factores estructurales que refuerzan las ventajas de los incumbentes. Por ejemplo:

- **Costos de entrenamiento y economías de escala:** El entrenamiento de modelos de frontera implica gastos hundidos en cómputo (GPUs/TPUs, energía, redes de alta velocidad), ingeniería (paralelización, *distributed training*), y *evaluation/safety*.³⁴

Gráfica 6. Costos de entrenamiento por modelo



Datos estimados para Gemini, Claude y GPT 5. Fuente: <https://epoch.ai/> (2026, abril 28)³⁵

30 Microsoft. (2023). *Microsoft Power Automate*. Microsoft Corporation. <https://powerautomate.microsoft.com>.

31 n8n. (2023). *Workflow automation platform*. n8n GmbH. <https://n8n.io>.

32 Zapier. (2023). *Automate your work today*. Zapier Inc. <https://zapier.com>.

33 En el sector fintech, destacan aplicaciones como Kasisto KAI (asistente virtual para banca digital), Upstart (evaluación crediticia mediante IA) y Stripe con AI (automatización de atención al cliente y detección de fraude). En marketing, ejemplos relevantes incluyen Jasper AI (generación de copys y campañas publicitarias), Copy.ai (redacción de contenido para marketing digital) y Synthesia (creación de videos con avatares para publicidad y capacitación).

34 Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. arXiv pre-print arXiv:2405.21015.

35 <https://epoch.ai/> (2026, Abril 28).

Estos costos generan economías de escala y de alcance: cada dólar adicional rinde más cuando la base de usuarios crece (amortización de costos fijos) y cuando la empresa reutiliza la misma tubería MLOps para múltiples productos (texto, imagen, voz).³⁶⁻³⁷ En principio, los incumbentes con acceso prioritario a hardware, datos y *experiencia técnica* tienen una ventaja sobre los pequeños participantes. No obstante, si bien esto es cierto, algunos importantes participantes, como Deepseek, han mostrado que se puede ser más costo-eficiente y alcanzar resultados similares.³⁸

- **Acceso a datos de alta calidad:** La calidad y diversidad de datos (texto, código, imágenes, voz, multimodal logs) condicionan la frontera de desempeño. Los datos curados, con licencias claras y feedback humano experto, son escasos y costosos. Las firmas con productos masivos (buscadores, suites de productividad, redes sociales, IDEs) acumulan datos de interacción de gran valor para entrenamiento y aprendizaje por refuerzo a partir de retroalimentación humana o de IA (*training y reinforcement learning from human/AI feedback*), lo que, en principio, crea una ventaja acumulativa difícil de replicar. Así Google, que maneja Gemini, tiene una ventaja sobre competidores como OpenAI o Deepseek.

Los entrantes dependen de datos abiertos (a menudo ruidosos), acuerdos de intercambio de datos, datos sintéticos y aprendizaje activo (*data sharing, synthetic data y active learning*). La incertidumbre legal (derechos de autor, minería de texto y de datos [*text-and-data mining*]) añade fricción y favorece a quienes pueden costear litigios y licencias.

Esta barrera se ve atenuada por la existencia de varias empresas en el eslabón de etiquetamiento y por el abaratamiento de lo que se denomina datos sintéticos (*synthetic data*, datos artificiales generados por algoritmos (especialmente IA) que imitan las características estadísticas y patrones de los datos reales.³⁹

- **Talento especializado y know-how:** Los equipos con experiencia en sistemas distribuidos, entrenamiento de escala (*training at scale*), y alineación de incentivos (*alignment*) son escasos. El know-how organizacional se acumula con cada ciclo de entrenamiento, generando activos intangibles y costos de cambio de talento (retención, golden handcuffs). La productividad de estos equipos también depende de contar con flujos de trabajo reproducibles, almacenes de características, gobernanza de datos (*feature stores, data governance*) y acuerdos de cómputo que disminuyen las fricciones operativas. Esta combinación refuerza ventajas incumbentes.

De hecho, hace poco se filtró que en septiembre de 2025 Meta, de Mark Zuckerberg, realizó ofertas compensatorias por hasta 300 millones de dólares para reclutar a ingenieros e investigadores de IA de élite. Estas ofertas tenían como objetivo específico atraer a expertos provenientes de empresas rivales como Google DeepMind y OpenAI.⁴⁰

36 McKinsey & Company. (2024). The cost of compute: A \$7 trillion race to scale data centers. McKinsey & Company. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers>.

37 Sutton, R. (2023, November 28). The economics of foundation models. Medium. <https://medium.com/economics-of-ai/the-economics-of-foundation-models-cb697cf3cd85>.

38 Gibney, E. Secrets of DeepSeek AI model revealed in landmark paper. Nature.

39 Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. Electronics, 13(17), 3509.

40 Verma, P. (2025, September 6). Mark Zuckerberg is offering top AI talent up to \$300 million. Wired. <https://www.wired.com/story/mark-zuckerberg-meta-offer-top-ai-talent-300-million/>.

- **Efectos de red y lock-in de plataformas:** En la capa de distribución (APIs, SDKs, *marketplaces*, integraciones), surgen efectos de red: más desarrolladores generan más integraciones lo que provoca más usuarios finales resultando en más datos para mejorar modelos.⁴¹

El lock-in ocurre cuando el ecosistema requiere herramientas propietarias (CUDA), fine-tuning y guardrails específicos de un proveedor, o plugins exclusivos. Migrar de un LLM a otro por parte de los desarrolladores implica costos de cambio (reentrenar/adaptar modelos, reescribir prompts/chains, rehacer compliance), lo que consolida el poder de mercado. Los frameworks abiertos (Hugging Face, LangChain, LlamaIndex) atenúan el lock-in, pero no eliminan dependencias de cómputo o distribución.

- **Regulación y estándares emergentes:** La regulación de la IA es complicada y sus efectos, particularmente en los mercados, se verán tiempo después de su aplicación. La regulación que se emita debe enfocarse solo en proteger derechos: propiedad intelectual, transparencia y rendición de cuentas. Lo importante en todo caso, es que existan requisitos de interoperabilidad para aumentar la competencia y la comparación entre modelos.
- **Integración vertical:** La integración vertical entre un proveedor de IA y otro productor confiere control sobre el diseño y la fabricación de chips, el cómputo en la nube y la distribución mediante aplicaciones. Quien coordina estos eslabones internaliza márgenes y logra ventajas en costo y tiempo de salida al mercado. Además, puede apalancar ventas atadas (créditos de nube + acceso preferente a modelos + integración nativa) y datos de uso para retroalimentar el *fine-tuning*. El caso más claro es Google con Anthropic, pero también está el caso de Microsoft con OpenAI.⁴²

Para rivales especializados en un solo eslabón (solo modelo, o solo app), competir exige alianzas estratégicas o diferenciación radical (abierto, *on-prem*, *privacy-first*, latencia local, nichos sectoriales).

Las alianzas estratégicas entre laboratorios de IA y proveedores de servicios en la nube se han convertido en un rasgo distintivo de la industria de modelos fundacionales. Estas asociaciones permiten a los laboratorios acceder a recursos de cómputo a gran escala, mientras que los proveedores de nube consolidan su posición como plataformas dominantes de distribución. El caso más emblemático es el de OpenAI y Microsoft, donde la nube de Microsoft ofrece la infraestructura y, a cambio, obtiene acceso preferente para integrar los modelos en productos como *Copilot*.⁴³

Un patrón similar se observa en la inversión de Google en Anthropic⁴⁴, que asegura que los modelos Claude se ejecuten de forma optimizada en Google Cloud, y en la colaboración de AWS con Anthropic y Hugging Face, donde Amazon Web Services no solo facilita cómputo sino también acceso a clientes corporativos a través de Bedrock. Estas alianzas generan sinergias técnicas y comerciales, pero también refuerzan dependencias, vinculan el desarrollo de modelos a los intereses de las grandes empresas.

Por su parte, los conglomerados tecnológicos han optado por la adquisición de startups de IA especializada. Por ejemplo:

41 Agranat, R., & Gal, M. S. (2025). Fueling Concentration: AI Agents and Network Effects. *Network Law Review*, Spring, 2016-2023.

42 Competition and Markets Authority. (2025). AI Foundation Models Update Paper. CMA.

43 Microsoft. (2025, January 21). *Microsoft and OpenAI evolve partnership to drive the next phase of AI*. Microsoft Corporate Blog.

44 Rooney, K. (2026, April 24). Google to invest up to \$40 billion in Anthropic as search giant spreads its AI bets. CNBC. <https://www.cnbc.com/2026/04/24/google-to-invest-up-to-40-billion-in-anthropic-as-search-giant-spreads-its-ai-bets.html>.

- La compra de Runway ML y MosaicML por parte de Databricks, que fortaleció su posición en el mercado de *machine learning as a service*.
- Apple, que ha adquirido pequeñas empresas de IA enfocadas en procesamiento de lenguaje y visión por computadora para reforzar su ecosistema de productos móviles.
- Salesforce, ha incorporado startups de IA conversacional y analítica predictiva para enriquecer su CRM con funciones avanzadas.

Estas adquisiciones ilustran cómo los conglomerados internalizan innovaciones, controlan aplicaciones finales y, al mismo tiempo, cierran el paso a posibles competidores independientes.

2. Tipos de clientes y su capacidad para enfrentar presiones competitivas

La IA generativa ha configurado distintos tipos de clientes que responden a necesidades diferenciadas según el contexto de uso y el grado de intermediación tecnológica. Estos tipos se agrupan en tres grandes categorías: consumidores individuales (B2C), empresas (B2B) y desarrolladores/startups (B2B2C). Analizar su posición relativa permite evaluar qué tan capaces son de resistir las presiones competitivas ejercidas por los proveedores de modelos fundacionales.

Consumidores individuales (B2C): Este tipo de cliente se centra en la oferta de productos de IA generativa diseñados para uso personal, en los que el usuario final interactúa directamente con el sistema sin mediación institucional. Los principales casos de uso incluyen: i) chatbots y asistentes virtuales como ChatGPT, Claude o DeepSeek, y ii) aplicaciones de generación de imágenes, video, música y texto como MidJourney, DALL·E, Runway o Suno.

El modelo de negocio predominante se basa en suscripciones, planes freemium y micropagos. Este grupo de clientes depende de la fijación de precios y de las políticas de acceso definidas por las grandes tecnológicas. No obstante, los usuarios de este grupo pueden cambiar entre plataformas fácilmente, lo que obliga a los proveedores a competir y frena el poder que podrían tener.

Empresas (B2B): Las empresas utilizan soluciones de IA generativa integradas en sus flujos de trabajo corporativos, con aplicaciones clave en: i) productividad y ofimática (Microsoft Copilot, Google Workspace), ii) CRM y atención al cliente (Salesforce Einstein GPT, Zendesk), y iii) recursos humanos (análisis de CVs, formación, evaluación de desempeño).

En este caso predominan modelos de negocio basados en licencias corporativas, tarifas por usuario o consumo de API. Las empresas pueden hacer frente a la presión competitiva pues poseen cierto poder de negociación gracias al volumen. No obstante, pueden enfrentar elevados costos de cambio. Esto genera una relación más dependiente con los proveedores de modelos fundacionales.

Desarrolladores y startups (B2B2C): Un tercer tipo de cliente está conformado por desarrolladores independientes y startups, que requieren acceso flexible a infraestructura y modelos de IA para crear nuevas aplicaciones. Sus principales vías de entrada son: i) APIs de modelos de lenguaje y multimodales, y ii) servicios en la nube como AWS Bedrock, Azure OpenAI Service o Google Vertex AI, que facilitan entrenamiento, ajuste fino y despliegue con menores costos iniciales.

Su capacidad para resistir presiones competitivas es baja, dado que dependen en gran medida de APIs propietarias y enfrentan elevados costos de cómputo y tienen poca capacidad de realizar grandes inversiones (*deep pocket*).

IV. DISCUSIÓN Y CONCLUSIONES

Los altos costos hundidos, acceso privilegiado a datos, talento especializado, efectos de red y dependencia a un proveedor (*lock-in*) de plataformas apuntan a un escenario donde pocos actores concentran una proporción significativa de los recursos estratégicos. Esta concentración podría traducirse en poder de mercado, donde las empresas incumbentes no solo fijan estándares tecnológicos y regulatorios, sino que también condicionan la dirección de la innovación. La integración vertical refuerza este panorama, pues el control sobre chips, nubes, modelos y aplicaciones reduce los márgenes de maniobra de proveedores intermediarios y usuarios empresariales.

Por un lado, las empresas líderes destinan inversiones masivas en entrenamiento de modelos de frontera, seguridad y escalabilidad. Estas mismas inversiones generan incentivos para recuperar costos a través de estrategias de exclusividad (contratos cerrados, *lock-in* de plataformas, ventas atadas), que pueden ralentizar la difusión de beneficios a la sociedad. La innovación puede quedar subordinada a la apropiación de rentas mediante modelos de suscripción, tarifas de uso de APIs o integración forzada con servicios complementarios.

Sin embargo, hay elementos que cuestionan seriamente la capacidad de los incumbentes de lograrlo. O es que acaso debemos ignorar que OpenAI está al borde de la quiebra o que Anthropic está en serios problemas financieros luego de perder su contrato con el Pentágono.⁴⁵

El mercado de inteligencia artificial generativa se encuentra en una fase de crecimiento acelerado y, aún conserva un margen significativo para la entrada de nuevos participantes, como lo demuestra la irrupción de modelos como DeepSeek R1, Kimi K2 y Hailuo⁴⁶ o Gemini 3.0 Pro. Estas entradas al mercado han demostrado que nuevos modelos pueden convertirse en auténticos *game changers*, redefiniendo dinámicas de competencia y elevando las expectativas sobre innovación y acceso.⁴⁷

Si bien los modelos de frontera (*frontier models*) marcan la pauta en términos de capacidad y alcance, la brecha entre estos y los modelos de gama media se ha ido reduciendo progresivamente. Incluso arquitecturas más sencillas logran hoy desempeños suficientemente altos para una amplia gama de aplicaciones prácticas. Es decir, a una persona promedio le resulta indiferente que el modelo tenga la capacidad de resolver un examen de matemáticas a nivel de doctorado o solo de maestría. Esta persona usa la AI para ordenar el correo y correr procesos repetitivos que no requieren un nivel abstracción de genios.

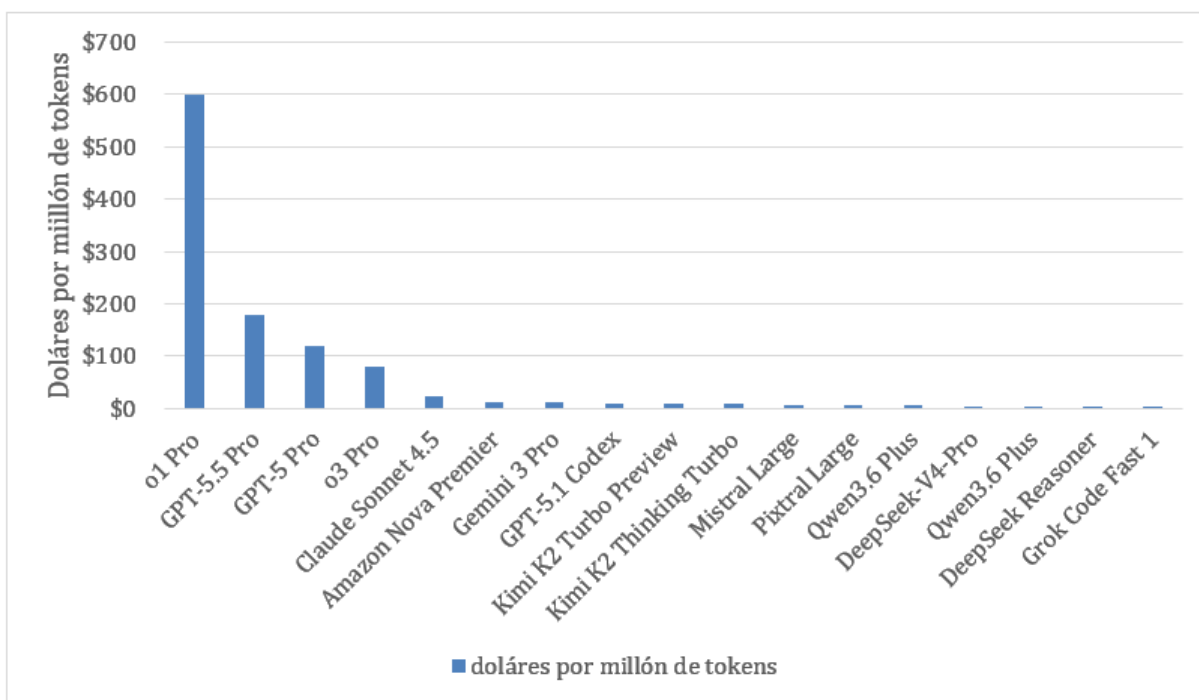
Este fenómeno disminuye el poder exclusivo de los *frontier models* y redistribuye la competencia hacia factores relacionados con la usabilidad, la adaptación sectorial y el costo de implementación, más que únicamente con el rendimiento técnico. Especialmente si el costo entre un modelo de frontera como GPT 5 Pro es 10 veces mayor que otros similares como Qwen o DeepSeek V4.

45 BBC News. (2025, 4 de julio). AI claims and a hoax spokesman: Viral band confuses world of music. BBC News. <https://www.bbc.com/news/articles/cvg3vlyzkkqeo> y Bosa, D. (2026, June 26). OpenAI, Anthropic face a new AI spending reality as users shift to efficiency. CNBC. <https://www.cnbc.com/2026/06/26/openai-anthropic-new-ai-spending-reality-as-users-shift-to-efficiency.html>.

46 <https://hailuoai.video/>.

47 El Financiero, *Gemini 3 pone a temblar a OpenAI: Altman declara 'código rojo' para mejorar su IA*.

Gráfica 7. Costos de los modelos por millón de tokens de salida



Fuente: Elaboración propia con datos de <https://www.llm-prices.com/> (2026, April 28)⁴⁸

Así, la concentración de mercado no se dará por las capacidades técnicas de los modelos, sino cada vez más por su costo y su integración con herramientas de productividad, especialmente si el costo de un modelo de frontera es 10 veces mayor que el de uno de cuasifrontera.⁴⁹ En este contexto, los modelos *open source* han demostrado ser un contrapeso estratégico frente al dominio de los actores cerrados.

En este documento, sin embargo, hay varios temas que no se abordan: 1) el análisis de prácticas anticompetitivas en la industria de la AI Generativa; 2) los efectos geopolíticos de la IA generativa; 3) los impactos regulatorios no previstos de normas y la interoperabilidad entre jurisdicciones para evitar asimetrías; y 4) el tema más importante de todos: el efecto, a mediano y largo plazo, de la AI en el mercado laboral; ahí ciertamente está el problema.

⁴⁸ <https://www.llm-prices.com/> (2026, April 28).

⁴⁹ Plataformas como Microsoft Office, Google Workspace o Salesforce incorporan IA generativa en flujos de trabajo cotidianos, lo que genera efectos de lock-in y barreras indirectas a la entrada, reforzando la posición de las empresas con una base instalada de usuarios y con control de infraestructuras críticas.

REFERENCIAS

- Anisuzzaman, D. M., Malins, J. G., Friedman, P. A., & Attia, Z. I. (2025). Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1), 100184.
- Agranat, R., & Gal, M. S. (2025). Fueling concentration: AI agents and network effects. *Network Law Review*.
- Bakare, L. (2025, julio 14). An AI-generated band got 1m plays on Spotify. *The Guardian*. <https://www.theguardian.com/technology/2025/jul/14/an-ai-generated-band-got-1m-plays-on-spotify-now-music-insiders-say-listeners-should-be-warned>.
- BBC News. (2025, 4 de julio). AI claims and a hoax spokesman: Viral band confuses world of music. *BBC News*. <https://www.bbc.com/news/articles/cvg3vlzzkqeo> y Bosa, D. (2026, June 26). OpenAI, Anthropic face a new AI spending reality as users shift to efficiency. *CNBC*. <https://www.cnbc.com/2026/06/26/openai-anthropic-new-ai-spending-reality-as-users-shift-to-efficiency.html>.
- Bowdon, C. (2025, agosto 25). How many parameters does GPT-5 have? *R-bloggers*. <https://www.r-bloggers.com/2025/08/how-many-parameters-does-gpt-5-have/>.
- Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. <https://arxiv.org/abs/2303.12712>.
- Chen, L., Zheng, K., & Shi, S. (2025). DeepSeek's Chinese-style innovation. *CKGSB Knowledge*. https://english.ckgsb.edu.cn/knowledge/professor_analysis/deepseek-redefining-global-ai-trends/.
- Competition and Markets Authority. (2025). AI Foundation Models Update Paper. *CMA*.
- Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv*.
- Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). How much does it cost to train frontier AI models? *Epoch AI*.
- Fu, J., Li, S., Jiang, Y., et al. (2022). StyleGAN-Human. <https://stylegan-human.github.io/>.
- Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>.
- Goyal, M., & Mahmoud, Q. H. (2024). Synthetic data generation techniques. *Electronics*, 13(17), 3509.
- Google AI. (2023). MusicLM. <https://google-research.github.io/seanet/musiclm/examples/>.
- Google DeepMind. (2024). Gemini 1.5. <https://blog.google/technology/ai/gemini-1-5/>.
- Google DeepMind. (2025). Gemini 2.5 Flash Image. <https://aistudio.google.com/models/gemini-2-5-flash-image>.

IBM. (s. f.). ¿Qué es la cuantificación? <https://www.ibm.com/mx-es/think/topics/quantization>.

IoT Analytics. (2024). Leading generative AI companies. <https://iot-analytics.com/leading-generative-ai-companies/>.

Johnson, J. (2025). Small language models. Hugging Face Blog. <https://huggingface.co/blog/jjokah/small-language-model>.

Kumar, D., & Sharma, P. (2022). Data acquisition strategies. Journal of Technology Management, 15(3), 45-62.

MarketsandMarkets. (2023). Data annotation market report. <https://www.marketsandmarkets.com>.

McKinsey. (2023). The economic potential of generative AI. <https://www.mckinsey.com>.

McKinsey & Company. (2024). The cost of compute. <https://www.mckinsey.com>.

Metz, C. (2023). Hugging Face wants to become the GitHub of AI. The New York Times.

Microsoft. (2023). Kosmos-1. <https://www.microsoft.com/en-us/research/blog/kosmos-1/>.

Microsoft. (2025, January 21). Microsoft and OpenAI evolve partnership to drive the next phase of AI. Microsoft Corporate Blog.

Microsoft. (2023). Power Automate. <https://powerautomate.microsoft.com>.

Mistral AI. (s. f.). Model weights documentation. <https://docs.mistral.ai>.

Nguyen, K. (2025). Cloud platform comparison. <https://niteco.com>.

Nobel, P., Rozenshtein, A. Z., & Sharma, C. (2025). Unbundling AI openness.

n8n. (2023). Workflow automation platform. <https://n8n.io>.

O'Connor, R. (2022). Diffusion models introduction. <https://www.assemblyai.com>.

Oluwagbenro, M. B. (2024). Generative AI: Concepts and applications. Authorea.

OpenAI. (2020). Jukebox. <https://openai.com/index/jukebox/>.

OpenAI. (2024). GPT-4o. <https://openai.com/index/hello-gpt-4o/>.

OpenAI. (2024). Video generation models. <https://openai.com>.

Our World in Data. (2023). Exponential growth of parameters. <https://ourworldindata.org>.

Ramamoorthy, L. (2025). Evaluating generative AI. International Journal of Future Multidisciplinary Research.

Rooney, K. (2026, April 24). Google to invest up to \$40 billion in Anthropic as search giant spreads its AI bets.

CNBC. <https://www.cnbc.com/2026/04/24/google-to-invest-up-to-40-billion-in-anthropic-as-search-giant-spreads-its-ai-bets.html>.

Runway. (s. f.). Runway. <https://app.runwayml.com/>.

Suno. (s. f.). Suno AI. <https://www.suno.ai>.

Sutton, R. (2023). Economics of foundation models. Medium.

TechTurismo Media. (2024). Open-washing en IA. <https://techturismo.media>.

Technology Innovation Institute. (2023). Falcon-40B. <https://huggingface.co/tiiuae/falcon-40b>.

Vaswani, A., et al. (2017). Attention is all you need. NeurIPS.

Vipra, J., & Korinek, A. (2023). Market concentration implications. CSET.

Wang, Z., et al. (2025). History of LLMs. AI and Ethics, 5(3), 1955–1971.

Wheeler, K. (2025). Top 10 generative AI tools. AI Magazine.

Yahoo Finance. (2024). Nvidia secures 92% GPU market. <https://finance.yahoo.com>.

Zapier. (2023). Automation platform. <https://zapier.com>.



Este documento se encuentra sujeto a los términos y condiciones de uso disponibles en nuestro sitio web:
<http://www.centrocompetencia.com/terminos-y-condiciones/>

Cómo citar este artículo:

Iván Paredes, "Mercados en Formación: Competencia en la Industria de la IA Generativa", *Investigaciones CeCo* (agosto, 2026),
<http://www.centrocompetencia.com/category/investigaciones>

Envíanos tus comentarios y sugerencias a centrocompetencia@uai.cl
CentroCompetencia UAI – Av. Presidente Errázuriz 3485, Las Condes, Santiago de Chile