



El estado de la economía de la IA

25 de junio de 2026

Azeem Azhar, William Gildea, Hannah Petrovic, PhD, Nathan Warren y Marija Gavrilov

Visión exponencial
www.exponentialview.co

Producido de forma independiente por Exponential View. Elaborado a partir de información pública y modelos propios de Exponential View; todas las conclusiones son responsabilidad nuestra. © Epiplus1 Ltd 2026

Por qué hemos hecho esto

Existe un problema de visibilidad en la economía de la IA. Hasta ahora, ha sido imposible desglosar la demanda real de los clientes.

La oferta en la economía de la IA está bien documentada. La mayoría de las empresas de semiconductores y los proveedores de servicios en la nube cotizan en bolsa y divulgan sus actividades con cierto detalle. Los analistas del lado vendedor han hecho un excelente trabajo desglosando su desempeño.

La parte de la demanda, es decir, qué es lo que los clientes pagan realmente y si los ingresos son reales, ha sido poco clara.

Los laboratorios más grandes son privados, e incluso las empresas públicas ocultan los ingresos por IA dentro de los totales de sus segmentos.

Sin comprender la demanda real, es imposible juzgar la salud de la economía de la IA, que sustenta una valoración bursátil de 22,7 billones de dólares y ha impulsado el crecimiento del PIB estadounidense en los últimos seis trimestres.

Esperamos que este informe sirva como fuente de referencia sobre la situación actual, libre de exageraciones y temores, y que nos ayude a todos a tener una conversación más informada sobre la influencia que la IA ejerce sobre la economía y el mundo en general.

Un agradecimiento especial a quienes revisaron amablemente un borrador inicial de esta presentación y nos brindaron sus comentarios:
Alex Imas, Shanu Mathew, Patrick Rutherford, Jaime Sevilla y Amy Sutter.

– Azeem y el equipo de Exponential View

Un modelo de ingresos patentado a nivel de línea: obtenido, puntuado, triangulado y deduplicado.

1. Fuente

De abajo hacia arriba, más de 1000 empresas

Cada línea de ingresos se rastrea hasta las declaraciones primarias, las cuentas auditadas, las transcripciones y los informes fidedignos; además de la atribución en la nube donde los ingresos de una empresa privada aparecen en las cuentas de una empresa pública (OpenAI a través de Azure, Anthropic a través de Bedrock).

Entre las señales blandas adicionales que utilizamos se incluyen fuentes no oficiales como:

- Comentarios públicos de ejecutivos y partes relacionadas.
- Métricas de muestra y de proxy.
- Comentarios, estimaciones no verificadas y filtraciones en medios tradicionales, nuevos medios y redes sociales.

Identificamos, investigamos y mantenemos nuestros conjuntos de datos mediante la investigación de analistas, complementada con un sistema propio que escanea, rastrea y sintetiza información relevante.

2. Puntuación de confianza

La confianza se evalúa antes de que cuente.

Cada línea lleva una puntuación de confianza **rigurosa** antes de entrar en cualquier modelo, por lo que los datos de entrada débiles no pueden inflar el número.

Damos la máxima calificación a las cifras presentadas, por encima de otras fuentes primarias, estimaciones corroboradas por terceros y afirmaciones de una sola fuente.

Todos los números derivados heredan la calificación más baja de las fuentes de entrada.

Registro de auditoría: Tabla de origen de ejemplo

source_name	source_reference	grade	source_datapoint	value
source_1	source_1_ref	A	source_1_dp1	1000
source_2	source_2_ref	B	source_2_dp1	1000
source_3	source_3_ref	C	source_3_dp1	1000
source_4	source_4_ref	D	source_4_dp1	1000
source_5	source_5_ref	E	source_5_dp1	1000
source_6	source_6_ref	F	source_6_dp1	1000
source_7	source_7_ref	G	source_7_dp1	1000
source_8	source_8_ref	H	source_8_dp1	1000
source_9	source_9_ref	I	source_9_dp1	1000
source_10	source_10_ref	J	source_10_dp1	1000

3. Modelar y triangular

Los modelos financieros de la empresa se contrastan con los de arriba hacia abajo.

Creamos modelos completos para cada empresa. En concreto, para las finanzas de GenAI (separadas de los informes de ingresos brutos), que abarcan los principales factores que influyen en los ingresos, la rentabilidad y los costes en la cuenta de resultados, los flujos de caja y los balances.

Estos modelos se comparan con indicadores indirectos independientes: silicio (ingresos del fabricante de chips), coste de fabricación, combinación de segmentos, investigación de la industria, tráfico y capacidad.

Registro de auditoría: Ejemplo de modelo de ingresos de la empresa

	2016	2017	2018	2019	2020	2021	2022	2023	2024
Revenue (\$M)	100	150	200	250	300	350	400	450	500
Operating Profit (\$M)	20	30	40	50	60	70	80	90	100
Net Income (\$M)	15	22	30	38	46	54	62	70	78
EPS (\$)	1.5	2.2	3.0	3.8	4.6	5.4	6.2	7.0	7.8
Operating Margin (%)	20	20	20	20	20	20	20	20	20
Net Income Margin (%)	15	15	15	15	15	15	15	15	15
EPS Margin (%)	15	15	15	15	15	15	15	15	15
Operating Cash Flow (\$M)	10	15	20	25	30	35	40	45	50
Free Cash Flow (\$M)	8	12	16	20	24	28	32	36	40
Capex (\$M)	2	3	4	5	6	7	8	9	10
Debt (\$M)	50	45	40	35	30	25	20	15	10
Equity (\$M)	50	55	60	65	70	75	80	85	90

4. Eliminar duplicados

El gasto solo se contabiliza una vez.

Los ingresos se contabilizan en cada nivel, pero nunca se suman entre ellos: se atribuyen al valor añadido, de modo que el mismo dólar no se contabiliza dos o tres veces.

Por ejemplo, un gasto de \$100 en una aplicación que envía \$60 a un proveedor de modelos, que gasta \$30 en alojamiento de inferencia, se cuenta como \$100, no como \$190:

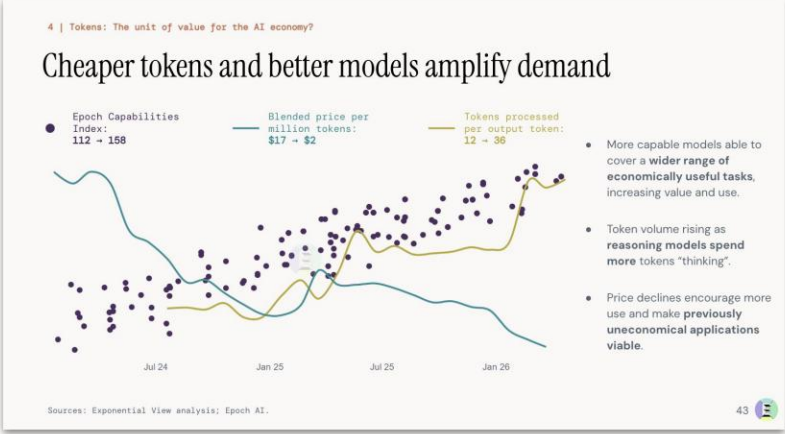
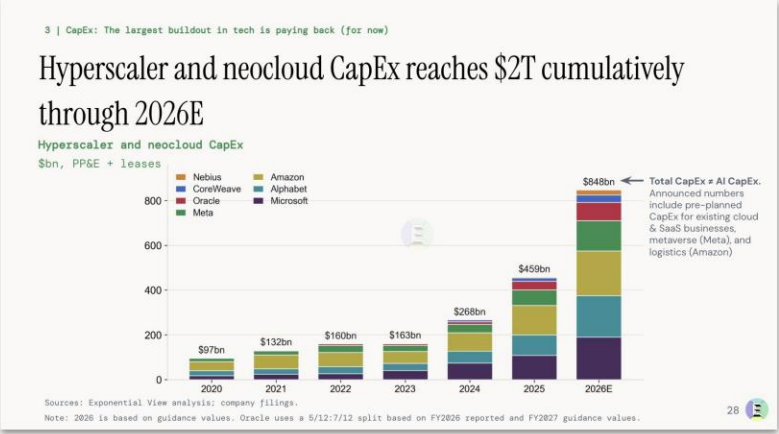


La línea superior

La demanda de IA se ve respaldada con mayor claridad por los ingresos reales que los cambios de plataforma anteriores. Los ingresos del ecosistema de IA generativa ya han superado los 175 mil millones de dólares anuales (tras eliminar la doble contabilización de los ingresos de los proveedores).

La intensidad de la inversión de capital (CapEx) está creciendo muy por encima de los estándares históricos de las grandes empresas tecnológicas para impulsar el desarrollo de la IA. Además, la financiación por parte de terceros está entrando cada vez más en el panorama financiero.

La incógnita reside en si la reducción de costes de la inteligencia artificial podrá generar el volumen y el margen de beneficio suficientes para cubrir el desarrollo de la infraestructura.



Contenido

1.	Demanda real	Clientes externos, ingresos reales, crecimiento sin precedentes.	06-18
2.	La economía	a gran escala aún es pequeña y está en sus inicios. Existen avances, pero son desiguales y no se pueden cuantificar.	19-26
3.	Costos de capital	La mayor inversión en infraestructura tecnológica de la historia está dando sus frutos (por ahora).	27-38
4.	Etiquetas	¿La unidad de valor para la economía de la IA, o no?	39-51
	Pila de 5	Donde se captura el valor. La pila convierte el capital y la energía en cognición.	52-62





1 | Demanda:

Es real, grande y rápido.

Los ingresos provienen de clientes externos reales.

El sector está creciendo tres veces más rápido que cualquier otra ola tecnológica anterior.

Esta demanda ha creado un superciclo informático: diez veces más capacidad de procesamiento, nueva generación de energía, centros de datos más grandes y una acumulación creciente de pedidos pendientes que la oferta no puede satisfacer.

1 | Demanda: Es real, grande y rápida

Ingresos de 110.000 millones de dólares en los últimos 12 meses, ahora a un ritmo de 175.000 millones de dólares.

Ingresos de la economía de IA generativa, sin duplicados, miles de millones
 de dólares al año, enero de 2023 - junio de 2026



Fuente: Análisis de Exponential View.
 Nota: Datos globales excluyendo China. Se han eliminado los ingresos duplicados por aplicaciones, modelos de negocio y alojamiento de infraestructura. No incluye la fabricación de chips.

1 | Demanda: Es real, grande y rápida

La demanda real y externa impulsa los ingresos de la IA.

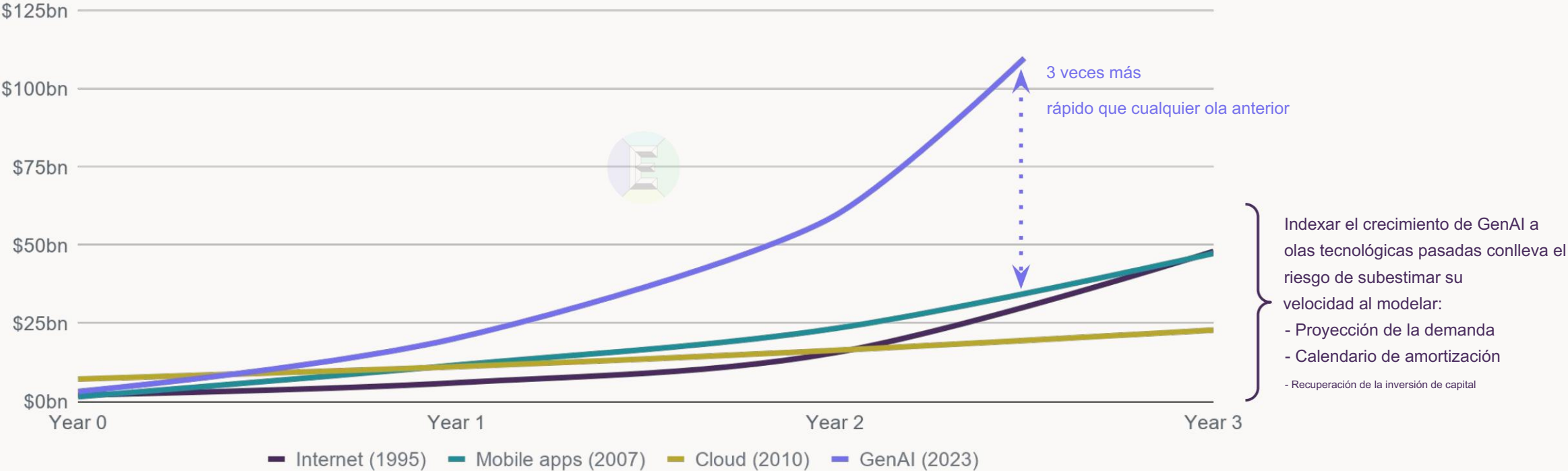
Ilustración de cómo modelamos y eliminamos los ingresos duplicados entre proveedores.



1 | Demanda: Es real, grande y rápida

La IA está creciendo tres veces más rápido que cualquier ola de TI.

Trayectoria de ingresos reales alineada con el año cero (miles de millones de dólares al año), ajustada por inflación.



Fuentes: Análisis de Exponential View; Departamento de Comercio de EE. UU.; informes de empresas; UBS; Oficina de Estadísticas Laborales de EE. UU.
Nota: Utilizamos los ingresos del primer año completo, por lo que hemos comenzado a medir GenAI a partir de enero de 2023.

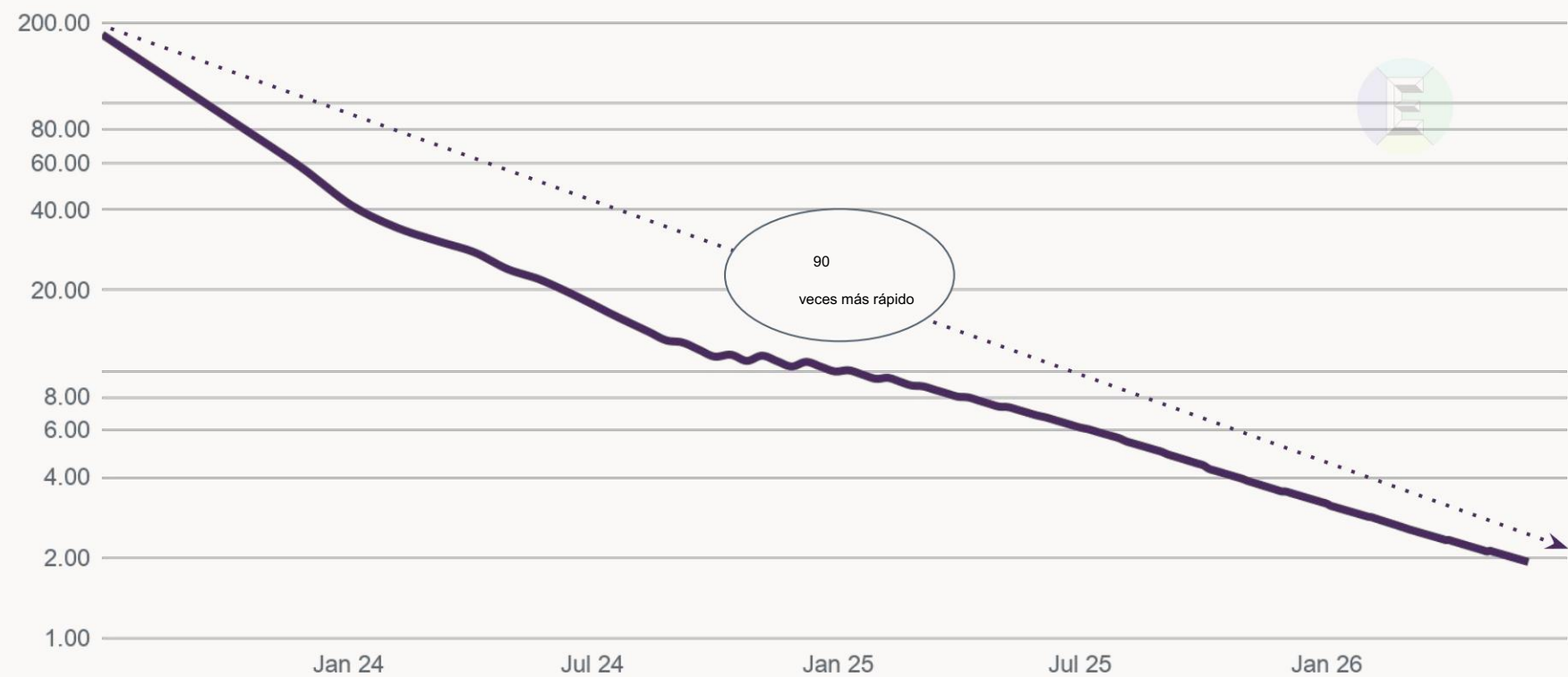


1 | Demanda: Es real, grande y rápida

Cada nuevo ingreso de mil millones de dólares llega más rápido que el anterior.

Es hora de añadir 1.000 millones de dólares adicionales a los ingresos acumulados.

días, escala logarítmica



En 2023, la industria de la IA necesitó 180 días para agregar \$1 mil millones en ingresos acumulados. ganancia.

Ahora necesita menos que dos días.

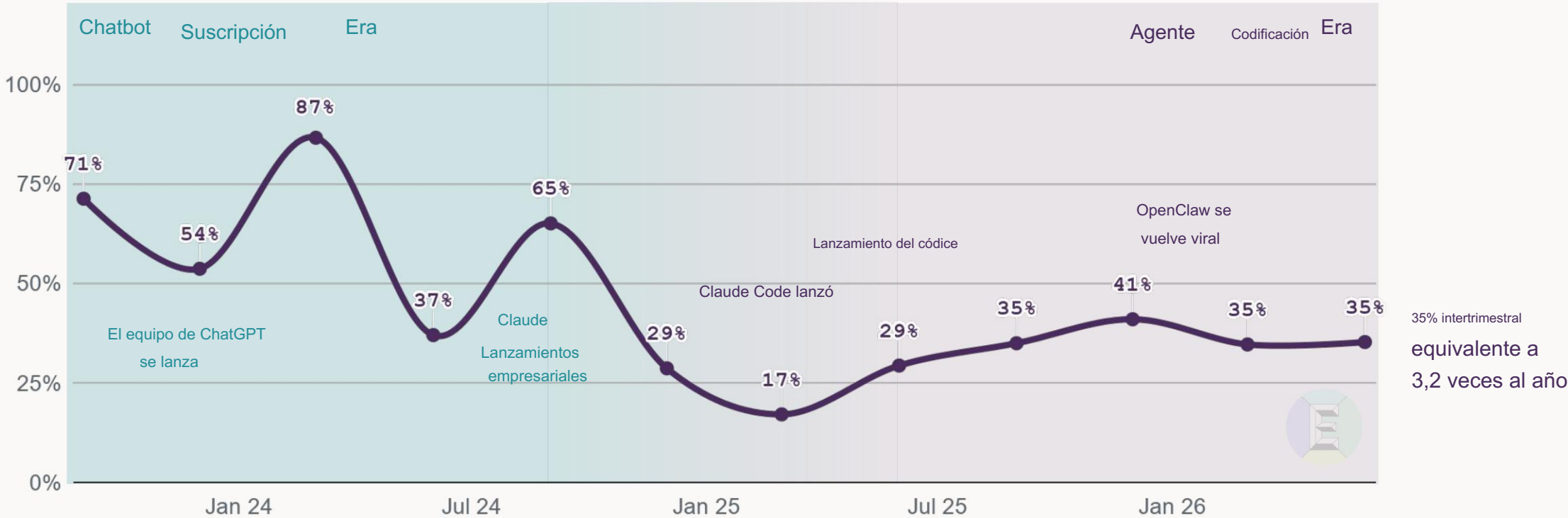
Fuente: Análisis de Exponential View.
Nota: Datos globales excluyendo China. Se han eliminado los ingresos duplicados por aplicaciones, modelos de negocio y alojamiento de infraestructura. No incluye la fabricación de chips.



1 | Demanda: Es real, grande y rápida

El crecimiento se ha mantenido en cada fase de adopción.

Crecimiento de los ingresos trimestre tras trimestre
% de cambio con respecto al trimestre anterior

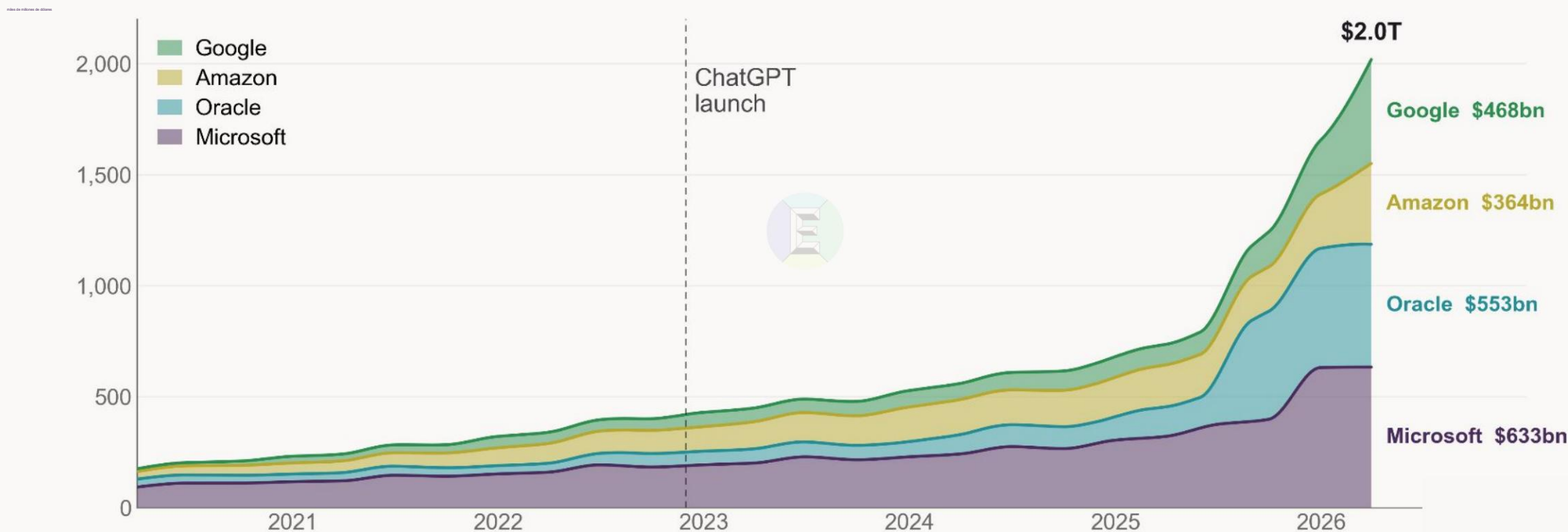


Fuente: Análisis de Exponential View.

1 | Demanda: Es real, grande y rápida

Este rápido crecimiento de la demanda se está reflejando en una acumulación de contratos para los proveedores de servicios en la nube a gran escala.

Cartera de pedidos combinada de hiperescaladores (obligaciones de rendimiento restantes)



Fuentes: Análisis de Exponential View; documentos presentados por la empresa.

Nota: Microsoft = RPO total (incl. M365/Dynamics); Amazon = RPO total de la empresa (principalmente AWS); Google = cartera de ingresos (principalmente Cloud).

El trimestre de Oracle finaliza un mes antes.

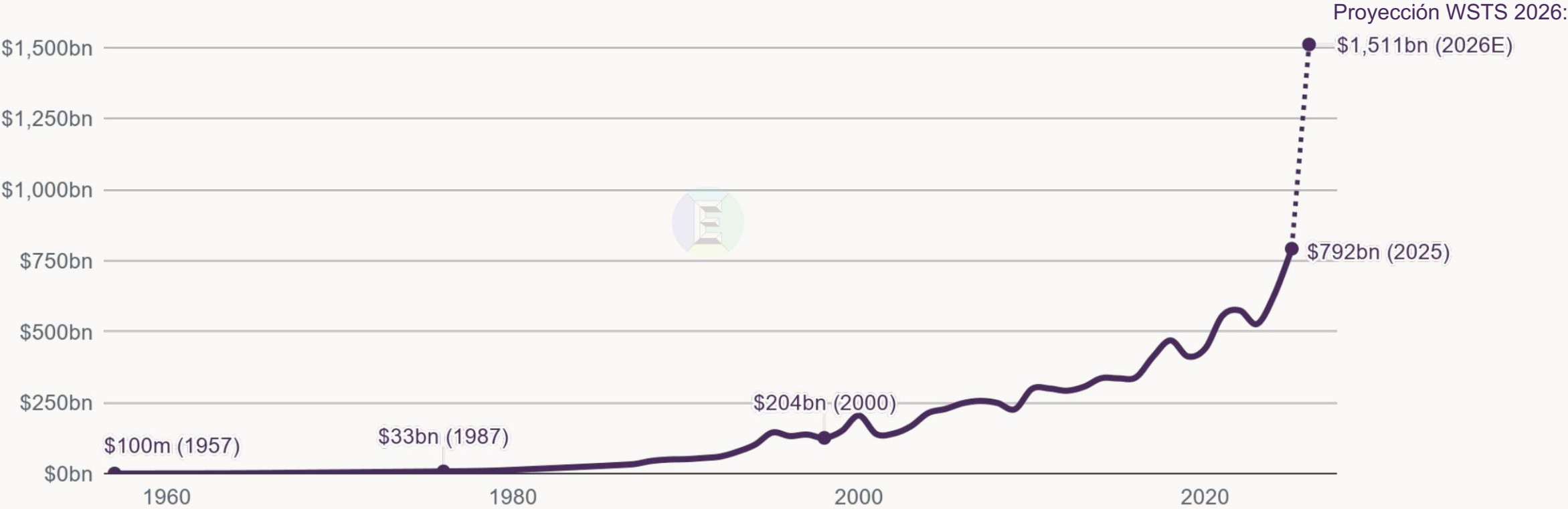


1 | Demanda: Es real, grande y rápida

La demanda ha puesto en marcha un superciclo computacional.

Ingresos del mercado mundial de semiconductores

miles de millones de dólares al año



Fuentes: Análisis de Exponential View; Estadísticas del Comercio Mundial de Semiconductores; Museo de Historia de los Semiconductores de Japón.

Nota: Valor nominal en dólares estadounidenses.

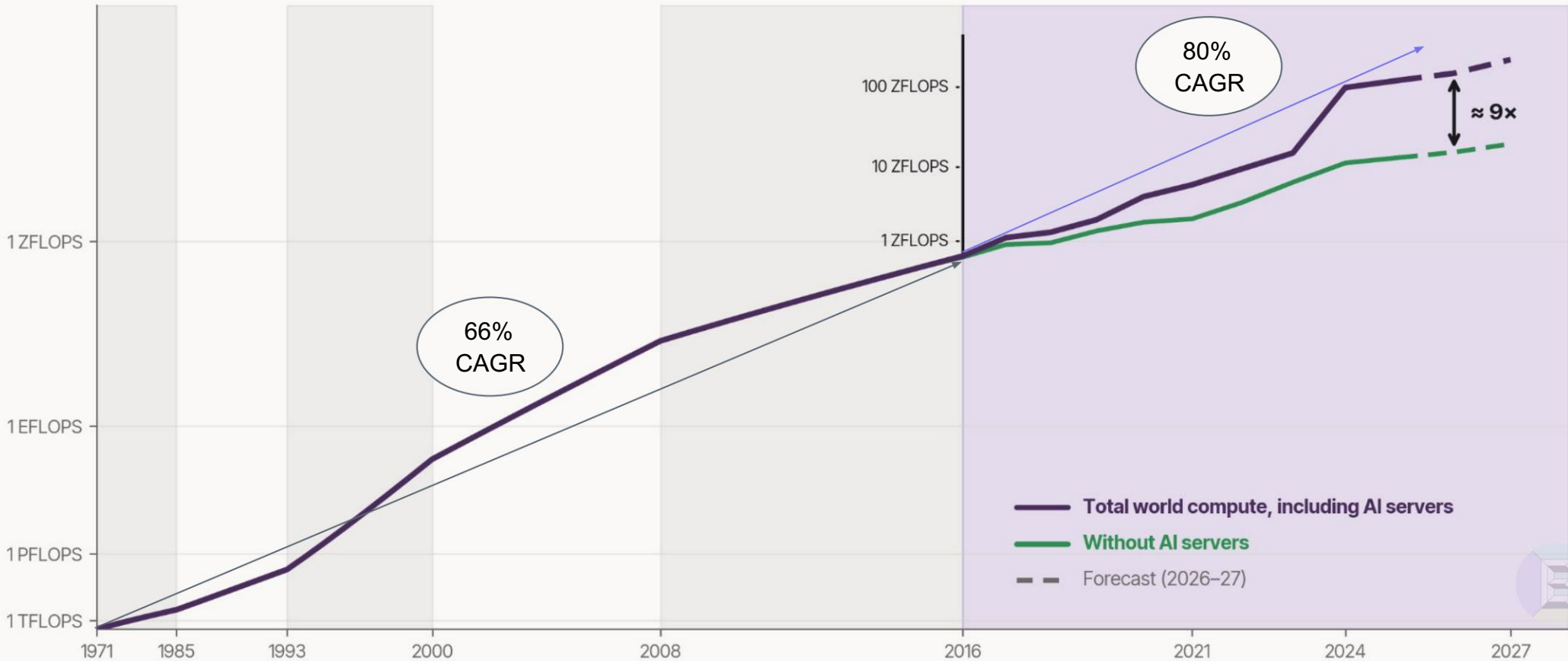


1 | Demanda: Es real, grande y rápida

La IA ha impulsado un repunte en una tendencia de crecimiento de la computación que se remonta a 50 años atrás.

Computación global desde 1971

FLOPS



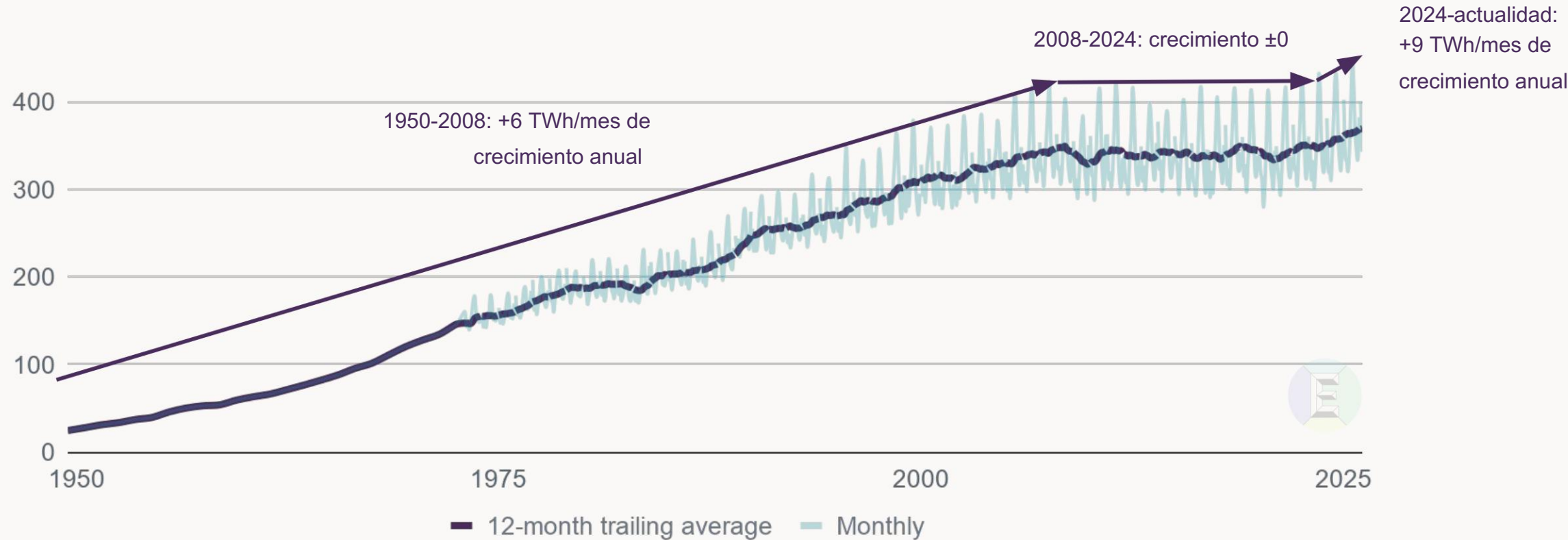
Fuente: Análisis de Exponential View.

Nota: Incluye mainframes, minicomputadoras, PC, servidores, teléfonos inteligentes, IoT y computación de IA. Los FLOPS del servidor de IA se derivan de la base instalada de GPU de Nvidia por generación, utilizando FP8 desde Hopper en adelante.

1 | Demanda: Es real, grande y rápida

La demanda de IA está reactivando un sector energético estadounidense moribundo.

generación neta de electricidad en EE. UU.
TWh/mes



Fuentes: Análisis de Exponential View; Administración de Información Energética de EE. UU.

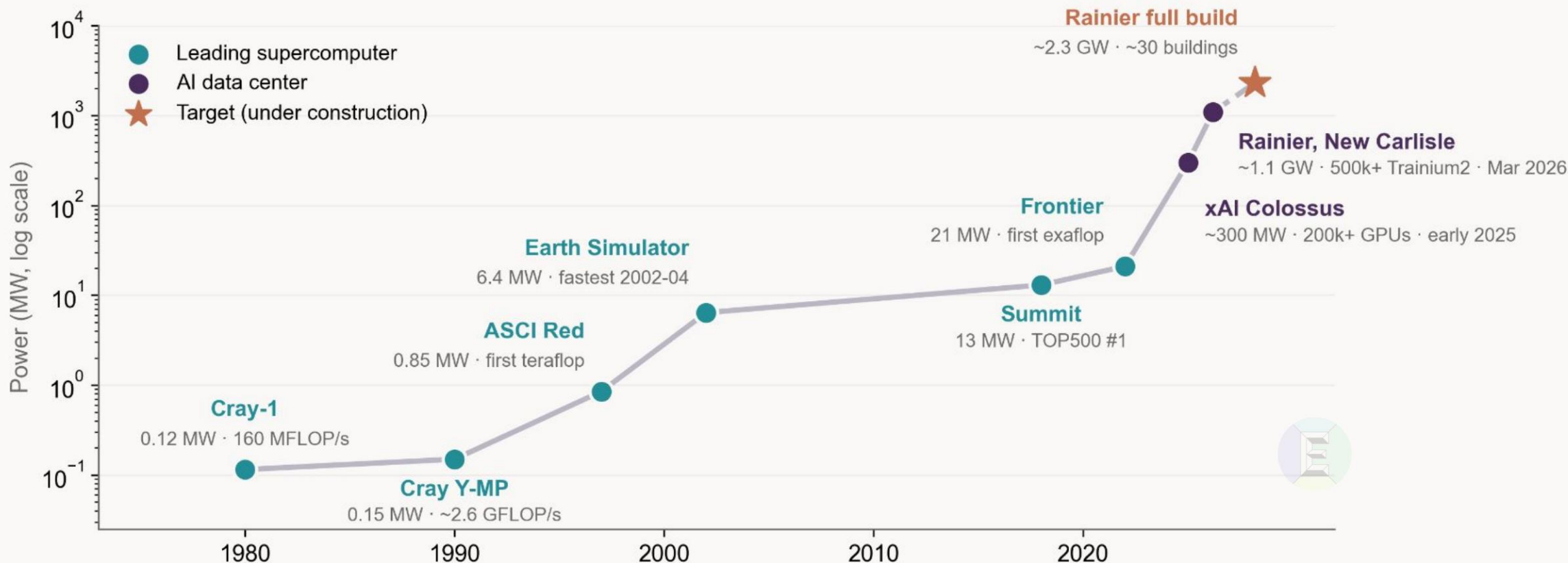


1 | Demanda: Es real, grande y rápida

El tamaño de los centros de datos más grandes se ha multiplicado por 50 en cuatro años.

El poder de las computadoras más potentes a lo largo del tiempo.

MW, escala logarítmica



Fuentes: Análisis de Exponential View; TOP500 / Green500; Epoch AI; OpenAI; Oracle.

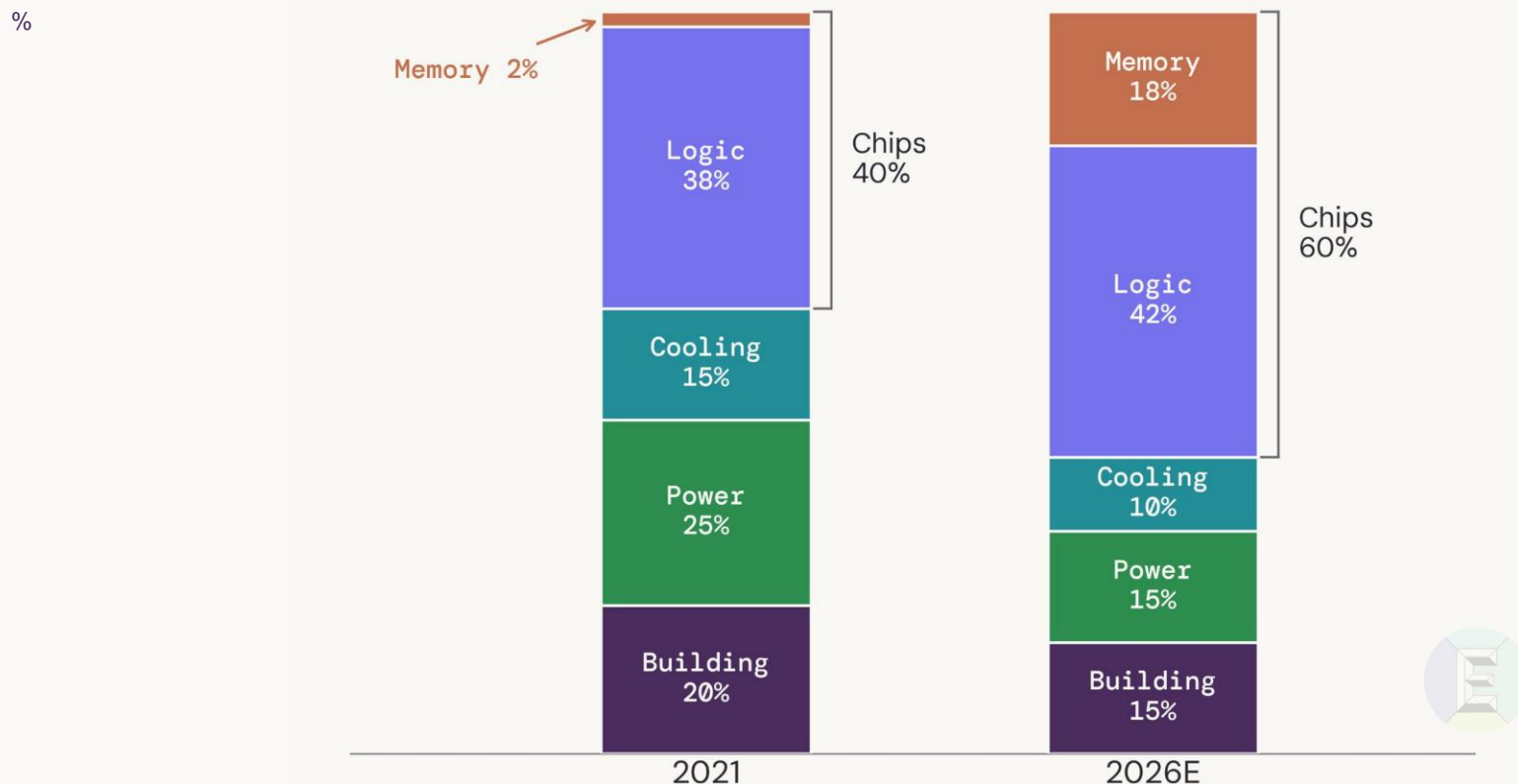
Nota: La construcción completa de Rainier es un objetivo.



1 | Demanda: Es real, grande y rápida

Actualmente, la memoria y la capacidad de procesamiento consumen la mayor parte de cada dólar invertido en la construcción de un centro de datos.

Participación en el coste total de construcción de centros de datos por componente, 2021 vs 2026E



- Cada nuevo dólar permite comprar más silicio y menos hormigón: la proporción del gasto en chips ha aumentado un 50 % (del 40 % al 60 %).
- La memoria es el factor que más ha variado, pasando de un error de redondeo del 2% a aproximadamente el 18%.

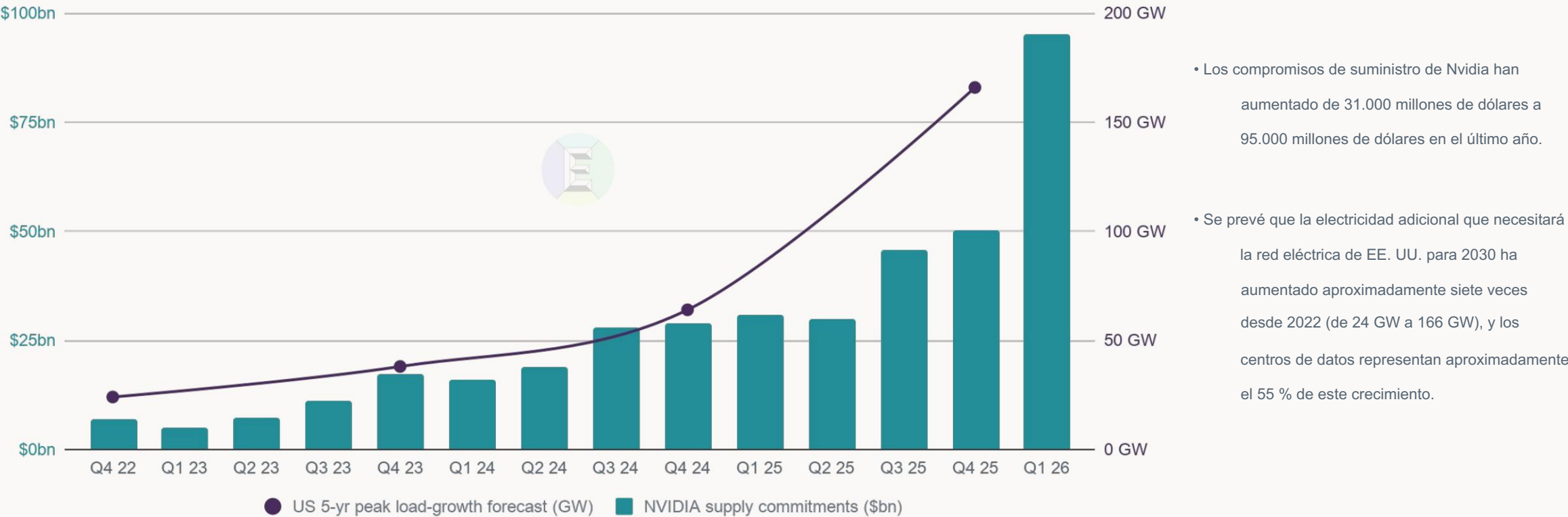


1 | Demanda: Es real, grande y rápida

Esto conlleva un aumento de los compromisos en materia de computación y energía.

Compromisos de computación y energía

Compromisos de suministro de Nvidia (miles de millones de dólares, eje izquierdo) y crecimiento de la demanda en EE. UU. (GW, eje derecho)



Fuentes: Análisis de Exponential View; Grid Strategies; Documentos presentados por Nvidia.



2 | Economía:

Lo grande sigue siendo pequeño y temprano.

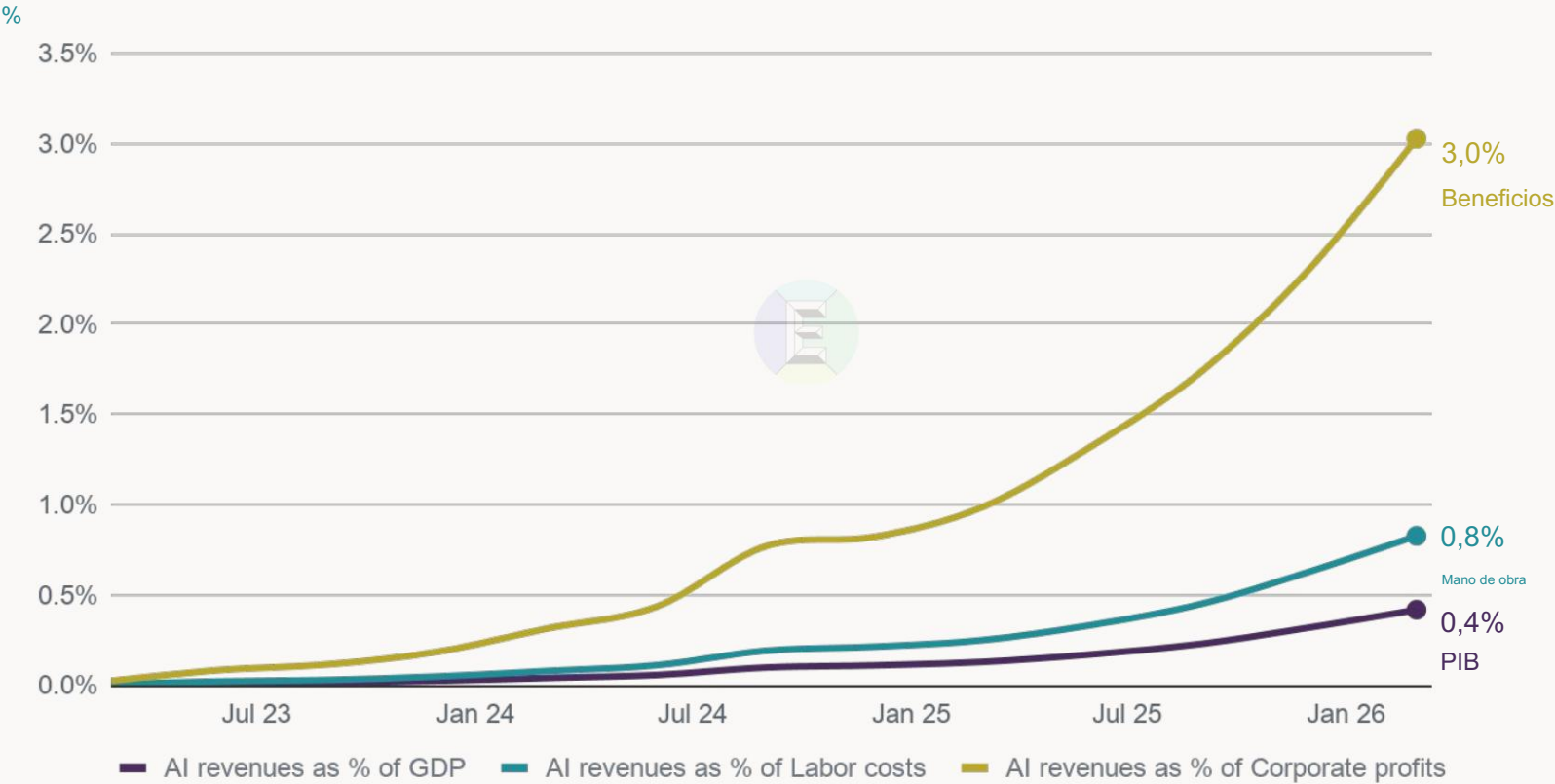
Incluso para las empresas que más invierten, la IA es un error de redondeo en su cuenta de resultados. Aún es pronto para sacar conclusiones. Las iniciativas se han centrado en la eficiencia y el ahorro de costes, aunque la combinación está cambiando. Además, los ingresos medidos podrían subestimar los beneficios sociales, ya que los consumidores informan de ventajas que aún no se reflejan en los datos.



2 | Economía: Lo grande sigue siendo pequeño, y temprano

En comparación con el PIB, los ingresos de la IA siguen siendo una cifra insignificante.

Ingresos globales de IA (excluyendo China), en relación con el PIB de EE. UU., los costos laborales y las ganancias corporativas.



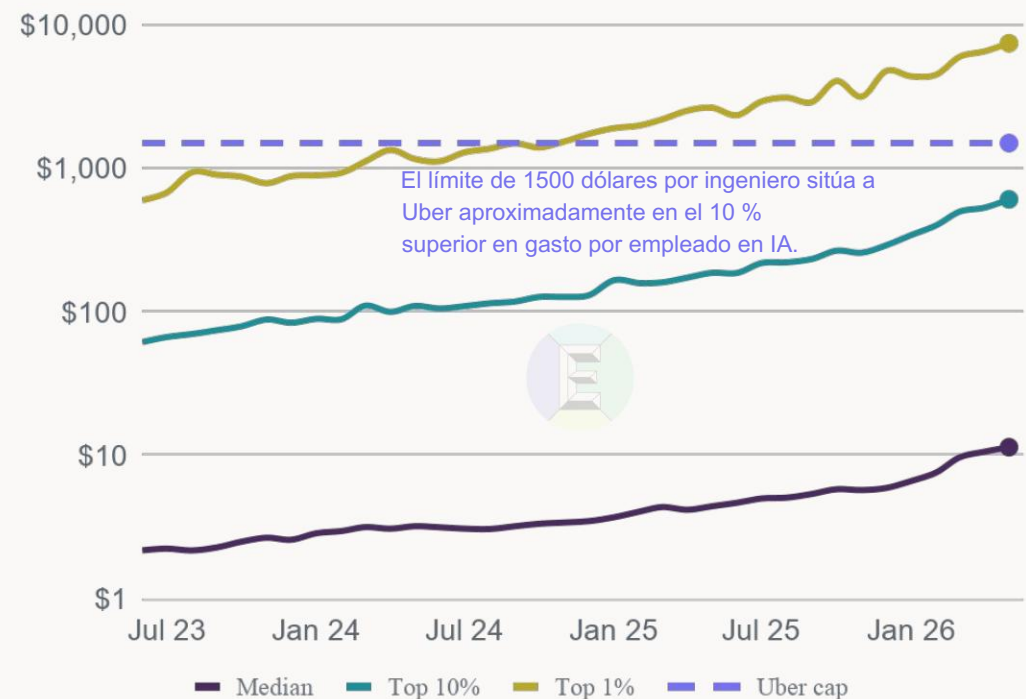
- Aún ínfimo: Los ingresos de la IA equivalen al 0,42 % del PIB estadounidense. (frente al 9,4% del sector de TI).
- Incluso tomando como referencia generosas los beneficios empresariales, la cifra es 32 veces mayor que todos los ingresos de GenAI.
- Aún es pronto: los ingresos por IA en relación con el PIB han aumentado 3 veces en comparación con el primer trimestre de 2025 (0,13 %) y 10 veces en comparación con el primer trimestre de 2024 (0,04 %).



2 | Economía: Lo grande sigue siendo pequeño, y temprano

A nivel empresarial, el gasto en IA sigue siendo relativamente pequeño: por ejemplo, los 1500 dólares por ingeniero de Uber apenas afectan a sus resultados.

Gasto en IA por empleado al mes, escala logarítmica, clientes de Ramp frente a Uber cap



Fuentes: Análisis de Exponential View; Ramp Economics Lab (n = 70.000 empresas estadounidenses); documentos presentados por Uber.

Nota: El 1 % superior, el 10 % superior y la mediana se definen según el nivel de inversión en IA en comparación con la base de clientes de Ramp. La cifra de Uber representa un límite máximo por ingeniero, comparado con la inversión en IA por empleado de Ramp.

Gasto en IA de Uber (límite máximo) frente a partidas de pérdidas y ganancias en dólares/año (gasto en IA %), escala logarítmica, frente al año fiscal 2025.



Al igual que con las tecnologías de propósito general anteriores, algunas mejoras pueden escapar a la medición del PIB.

Impactos de la IA

excedente del consumidor

Valor que llega directamente a las personas a un precio casi nulo. Se vende poco, por lo que el PIB subestima el beneficio para el consumidor, por ejemplo:

- Reemplazo gratuito de software y compras de servicios.
- Aprendizaje, ocio y comodidad

Casos históricos

1880-1920: Iluminación eléctrica

La luz se abarató en un 99,97%: con el salario de una hora se puede comprar aproximadamente 40.000 veces más. Los precios no registraron esta ganancia.

≈ \$0

impacto directo
en el PIB

Nordhaus (1996)

2000-2020: Bienes digitales gratuitos

La búsqueda gratuita, las enciclopedias y los mapas han desplazado a los servicios de pago. Solo la búsqueda supone un coste aproximado de 17.500 dólares al año por persona.

≈ \$0

impacto directo
en el PIB

Brynjolfsson y otros (2019)

Excedente del productor

Valor inherente a los bienes y servicios vendidos. Se realizan y registran más transacciones en el PIB, por ejemplo:

- Funcionalidades habilitadas por IA que impulsan los ingresos
- Lanzamientos más rápidos (valiosos)
- Márgenes de las empresas de servicios

1850-1870: Vapor

Fábricas mecanizadas y ferrocarriles, de modo que un solo trabajador podía producir y transportar mucho más al mercado.

+0,4

PIB anual por persona

Artesanía (2004)

1980-2000: Automatización

Las máquinas herramienta programables reducen la mano de obra en cada unidad, aumentando la productividad por trabajador.

+0,37

PIB anual por persona

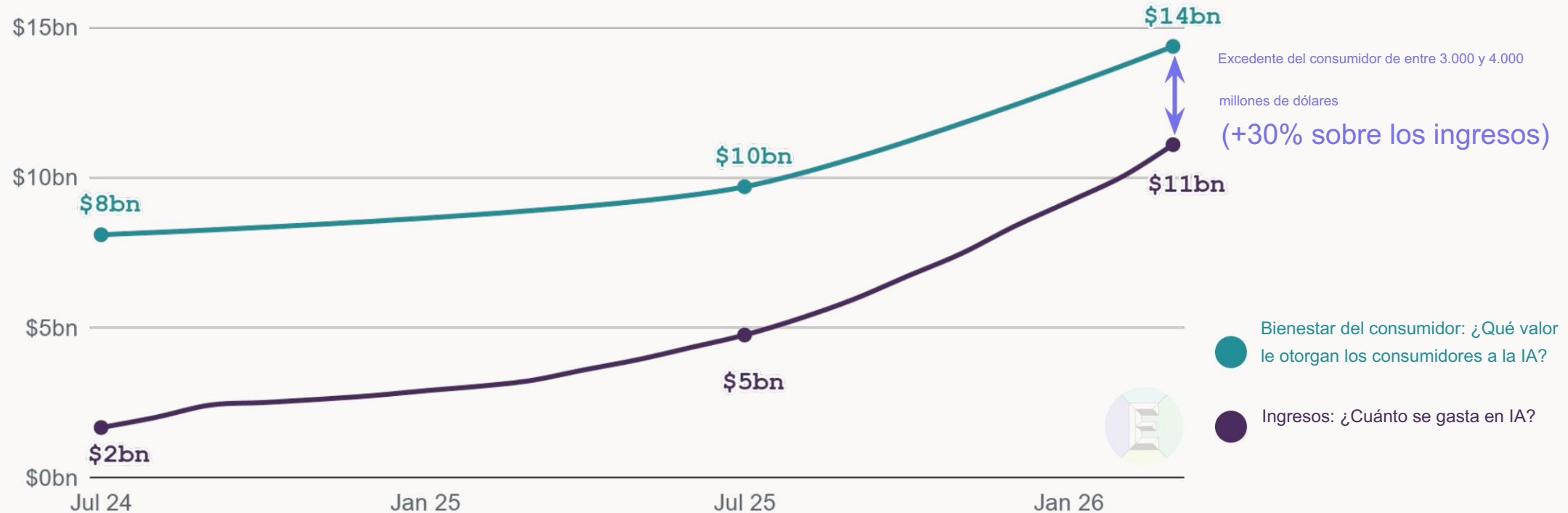
Graetz y Michaels (2018)



2 | Economía: Lo grande sigue siendo pequeño, y temprano

El PIB conoce el precio de todo, pero el valor de nada.
El valor económico de la IA supera los ingresos medidos.

Ingresos mensuales de GenAI frente al bienestar del consumidor estadounidense
miles de millones de dólares/mes, julio de 2024 - marzo de 2026



Fuentes: Análisis de Exponential View; Laboratorio de Economía Digital de Stanford

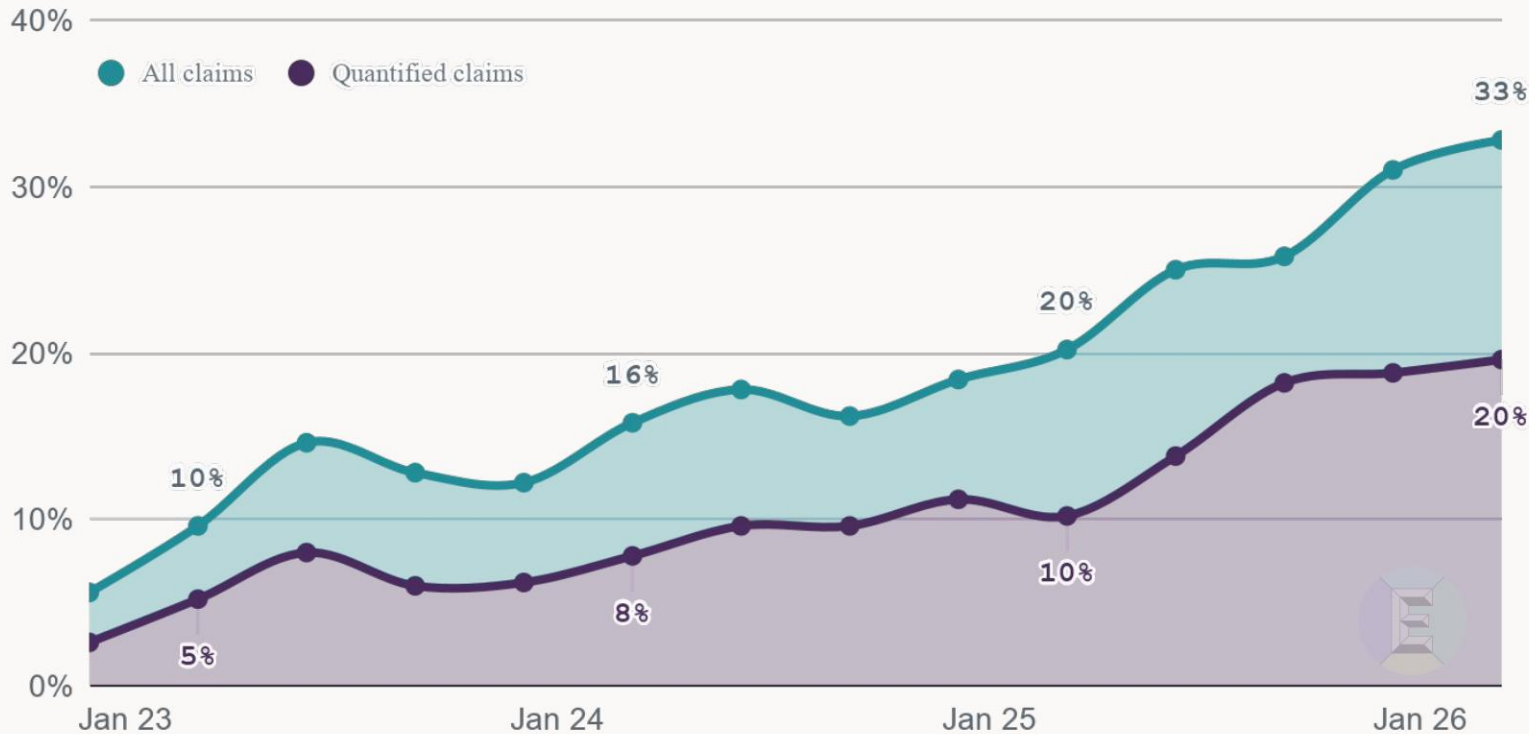
Nota: El valor del beneficio se determinó a partir de las respuestas a la pregunta: "¿Renunciaría al acceso a todas las herramientas de IA como ChatGPT, Gemini, Claude o Copilot durante un mes a partir de mañana por la mañana a cambio de [USD]?". Los ingresos incluyen el gasto global (excluyendo China) de consumidores y empresas.



Las empresas públicas están reportando un mayor impacto de la IA genética.

Empresas que afirman que la IA tiene un impacto en las teleconferencias sobre resultados

S&P 500, cuarto trimestre de 2022 - primer trimestre de 2026



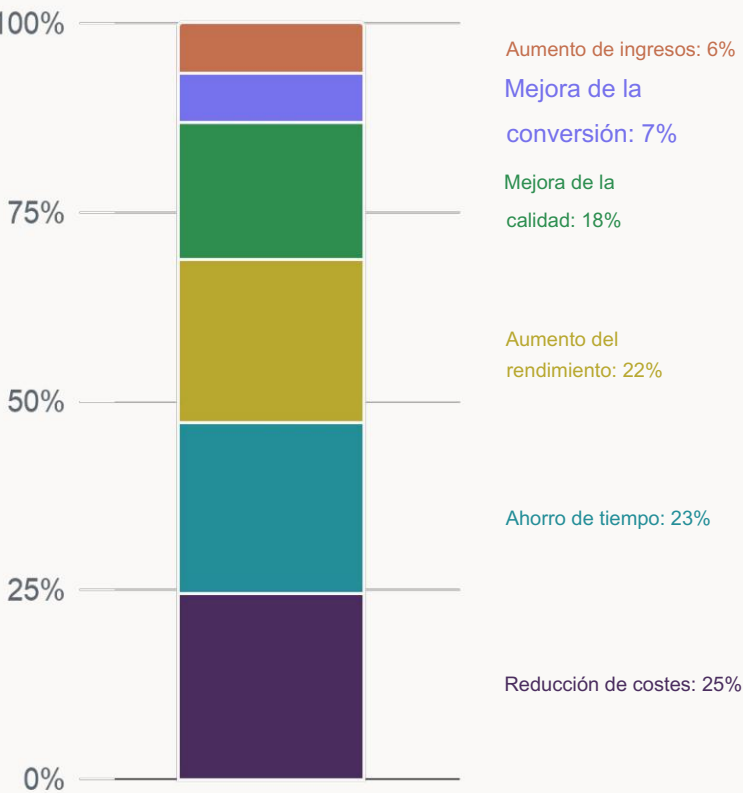
- Mayor interés: Las empresas ven la IA como una oportunidad para mejorar sus ganancias.
Hemos registrado un aumento de 3 a 4 veces en las menciones del impacto de la IA en el índice S&P 500 desde 2023.
- Póngale un número: 50-60% de
Las reclamaciones ya están cuantificadas, pero aún está por verse cuán importantes y significativas serán para los resultados financieros de las empresas.
- Sigue siendo una minoría: La mayoría de las empresas aún no han informado sobre un impacto cuantificado derivado del uso de la IA.



Siete de cada diez afirmaciones de GenAI se centran en el ahorro de costes o la eficiencia.

Resultados de IA declarados

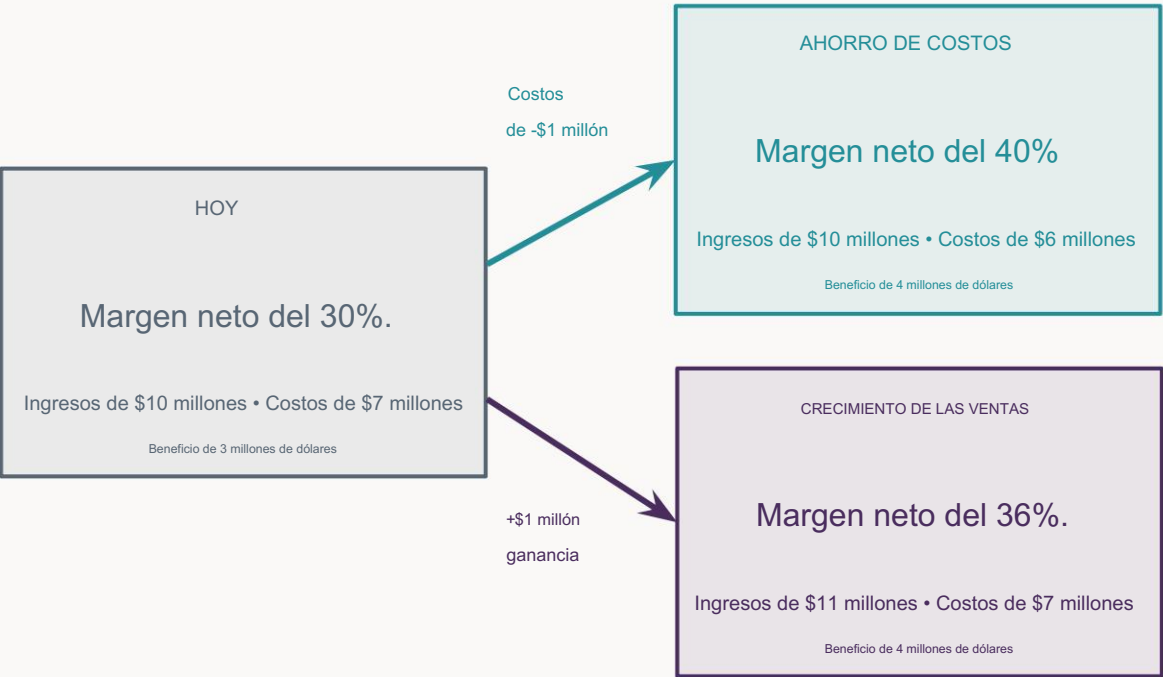
S&P 500, cuarto trimestre de 2022 - primer trimestre de 2026



Fuentes: Análisis de Exponential View; teleconferencias sobre resultados.
Nota: Es posible que las cifras no sumen el 100% debido al redondeo.

Por qué los proyectos iniciales priorizan la eficiencia: un ejemplo ilustrativo.

El mismo impacto puro de 1 millón de dólares tiene **mayores ganancias** frente al crecimiento de las ventas.

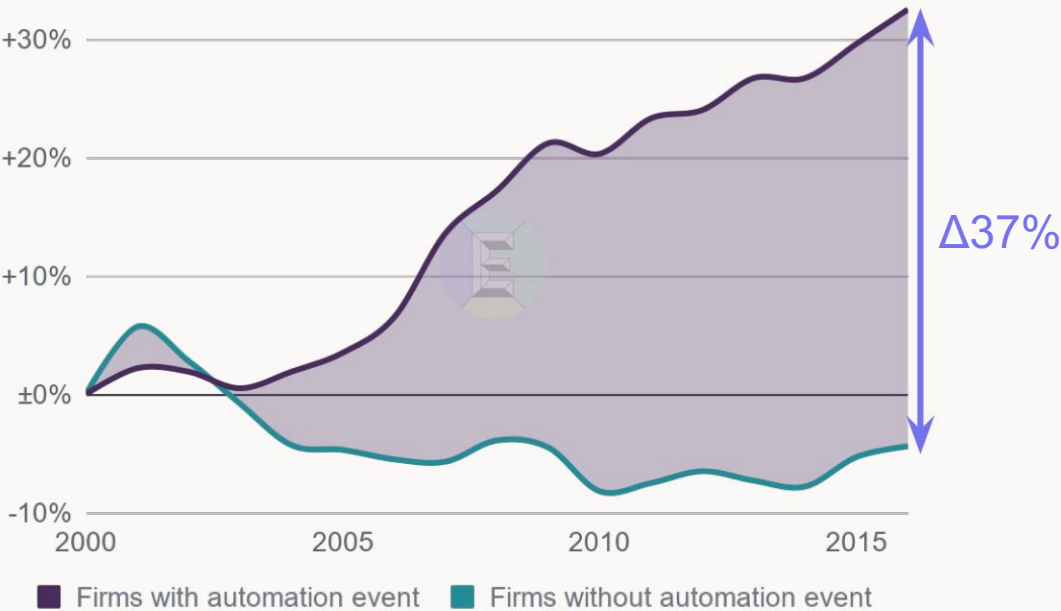


Nota: A efectos ilustrativos, los ingresos se suman sin coste alguno.
Si se suman los costes de ventas, el crecimiento de los márgenes se vería aún más frenado.

2 | Economía: Lo grande sigue siendo pequeño, y temprano

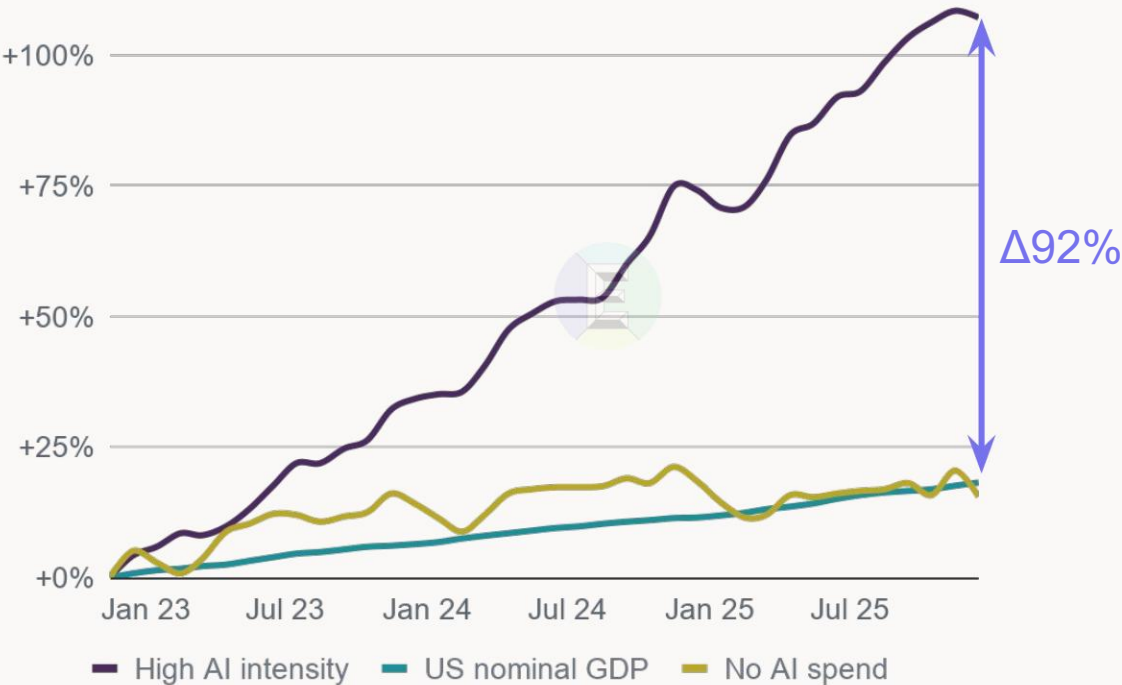
Al igual que en oleadas anteriores, los primeros en adoptar las nuevas tecnologías están superando a sus pares.

El caso de estudio histórico:
Cambio porcentual en el empleo a nivel de empresa con y sin automatización
(2000-2016)



Fuentes: Análisis de Exponential View; Bessen, Goos, Salomons y Van den Berge (2020): "¿Qué sucede con los trabajadores en las empresas que se automatizan?".

La economía de la IA en la actualidad:
Crecimiento de los ingresos con alto uso de IA frente a sin uso de IA:
variación porcentual desde noviembre de 2022



Fuentes: Análisis de Exponential View; Ramp Economics Lab (n = 70.000 empresas estadounidenses).
Nota: Alta intensidad = 25 % de las empresas que más invierten en IA por porcentaje de ingresos.





3 | Gastos de capital:

La mayor inversión en infraestructura tecnológica de la historia está dando sus frutos (por ahora).

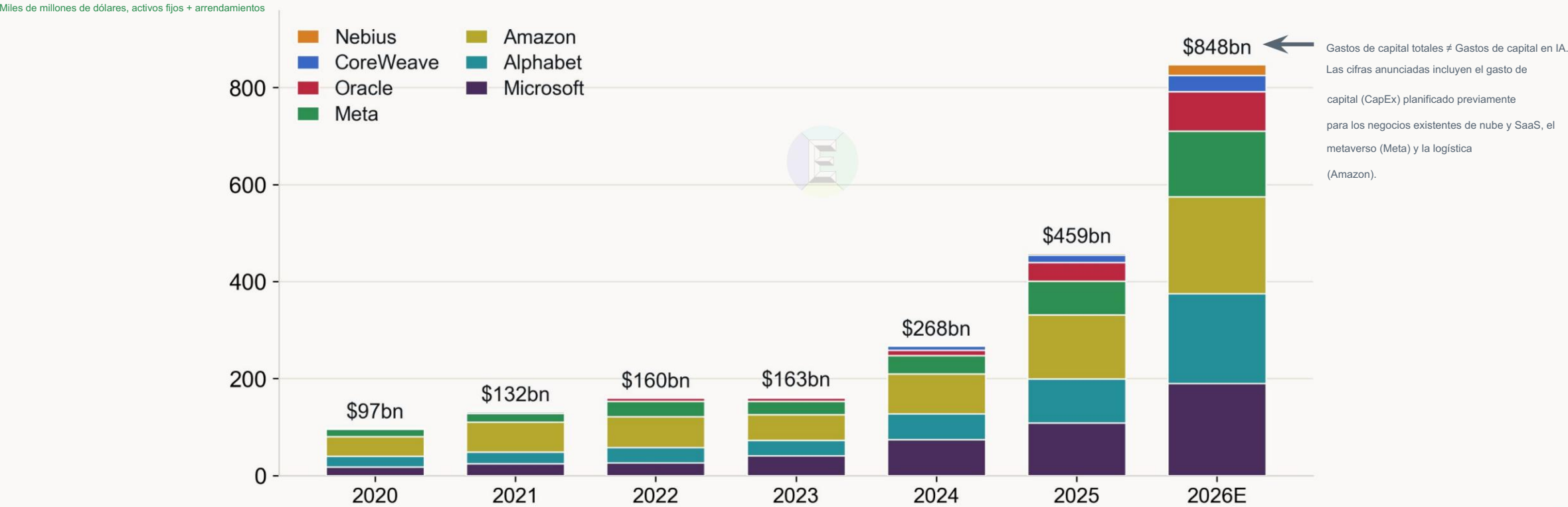
Los proveedores de servicios en la nube a gran escala y las plataformas de computación en la nube de nueva generación se han comprometido a invertir 2 billones de dólares en gastos de capital acumulados hasta 2026, lo que ejerce presión sobre el crecimiento de los ingresos para amortizar la deuda, especialmente a medida que aumenta la financiación con capital externo. Esta situación económica marca la pauta para la financiación de los centros de datos y la producción de tokens.



3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

El gasto de capital (CapEx) de los hiperescaladores y la neocloud alcanza los 2 billones de dólares acumulados hasta 2026E.

Gastos de capital de hiperescaladores y neocloud



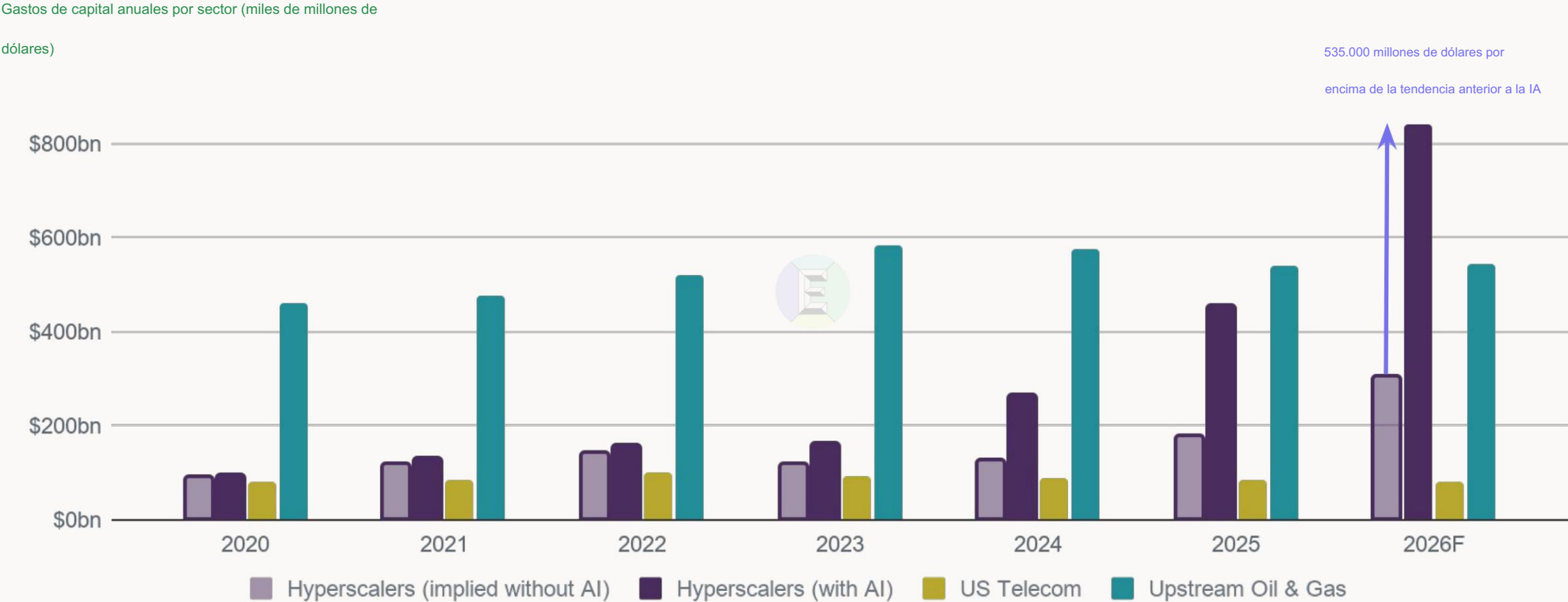
Fuentes: Análisis de Exponential View; documentos presentados por la empresa.

Nota: El año 2026 se basa en valores de referencia. Oracle utiliza una distribución 5/12:7/12 basada en los valores reportados para el año fiscal 2026 y los valores de referencia para el año fiscal 2027.



3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

El gasto de capital vinculado a la IA añade 535.000 millones de dólares por encima de la tendencia anterior a la IA para 2026E.



Fuentes: Análisis de Exponential View; Resultados y previsiones de Hyperscaler; Previsiones de telecomunicaciones y operadores de EE. UU.; Agencia Internacional de Energía.

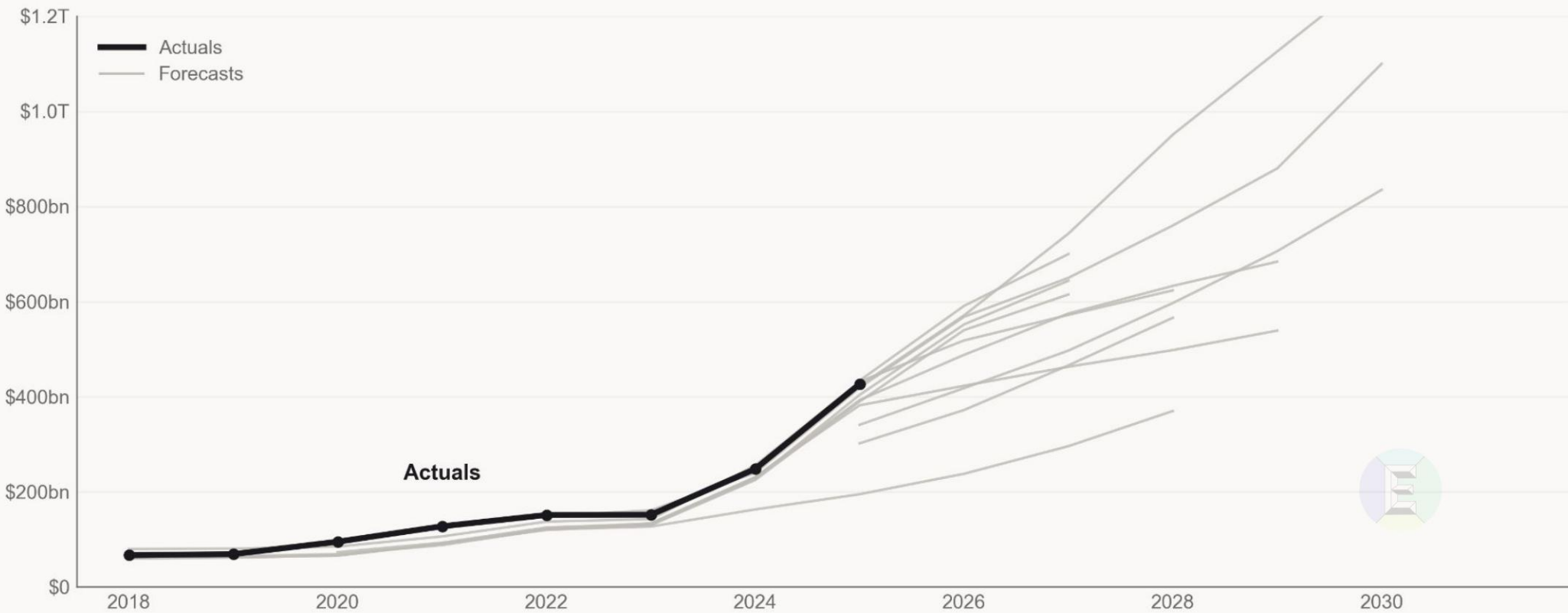


3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Las previsiones han seguido la curva de gastos de capital al alza.

Previsiones de gastos de capital (CapEx) para hiperescaladores/infraestructura de IA por analista y fecha de previsión.

\$/año



Fuentes: Análisis de Exponential View; Barclays, Citi, Goldman Sachs, JP Morgan, Morgan Stanley, New Street Research, SemiAnalysis, UBS; informes de empresas.

Nota: Los analistas utilizan cestas de inversión ligeramente diferentes (las cinco grandes empresas de hiperescala frente a una infraestructura de IA más amplia). Los datos reales aquí presentados corresponden a las inversiones de capital en efectivo (adquisiciones de propiedades y equipos) de Microsoft, Alphabet, Amazon, Meta, Oracle, CoreWeave y Nebius.



3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

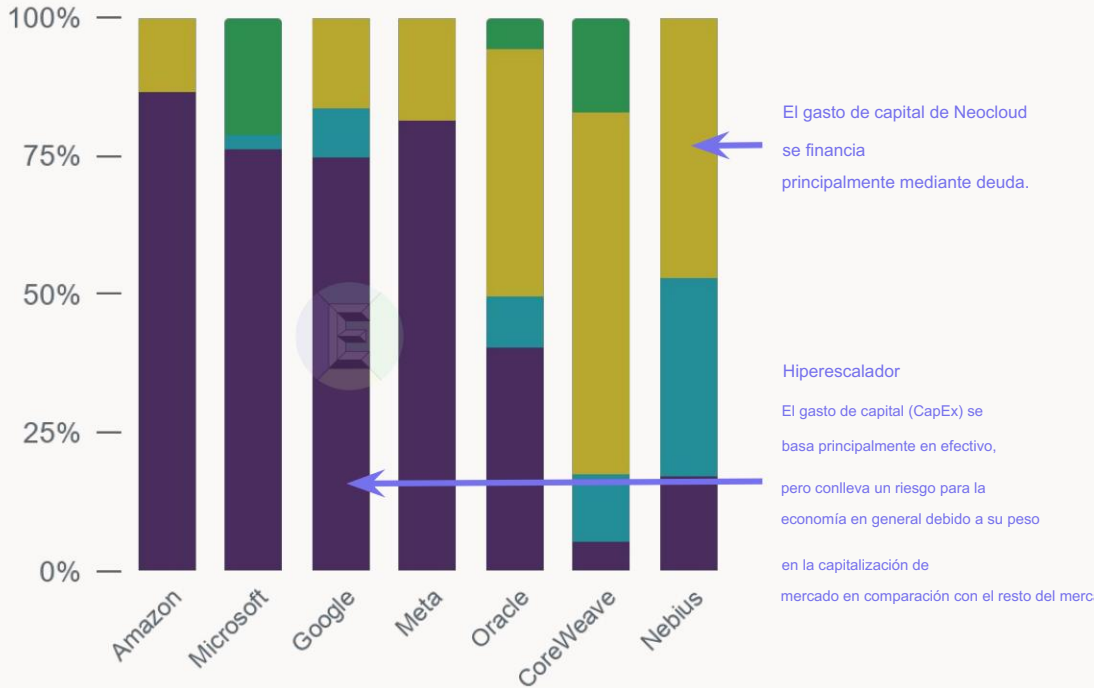
El dólar destinado a la infraestructura de IA marginal se financia cada vez más desde el exterior.

Gastos de capital (CapEx) de hiperescaladores y neocloud por fuente de financiación
\$, 2020-2026E



Gastos de capital por fuente de financiación

%, total estimado para el período 2020-2026



● flujo de caja operativo ■ Arrendamientos ■ Deuda (nueva neta) ■ Equidad ■ Dinero

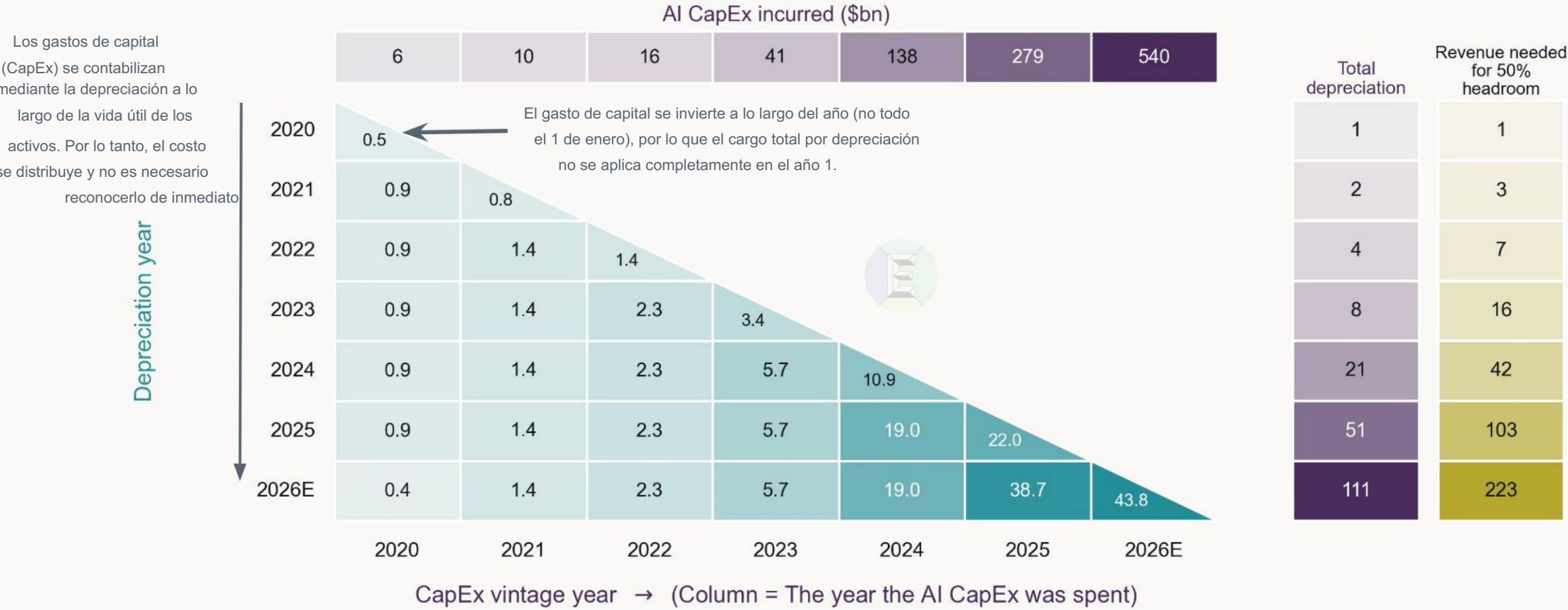
Fuentes: Análisis de Exponential View; documentos presentados por la empresa.

Nota: La deuda se calcula después de deducir los reembolsos (no la emisión bruta) e incluye todos los instrumentos de deuda (bonos, pagarés comerciales, etc.).



3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

El cargo por depreciación estimado para 2026 se acerca a los 111 mil millones de dólares.



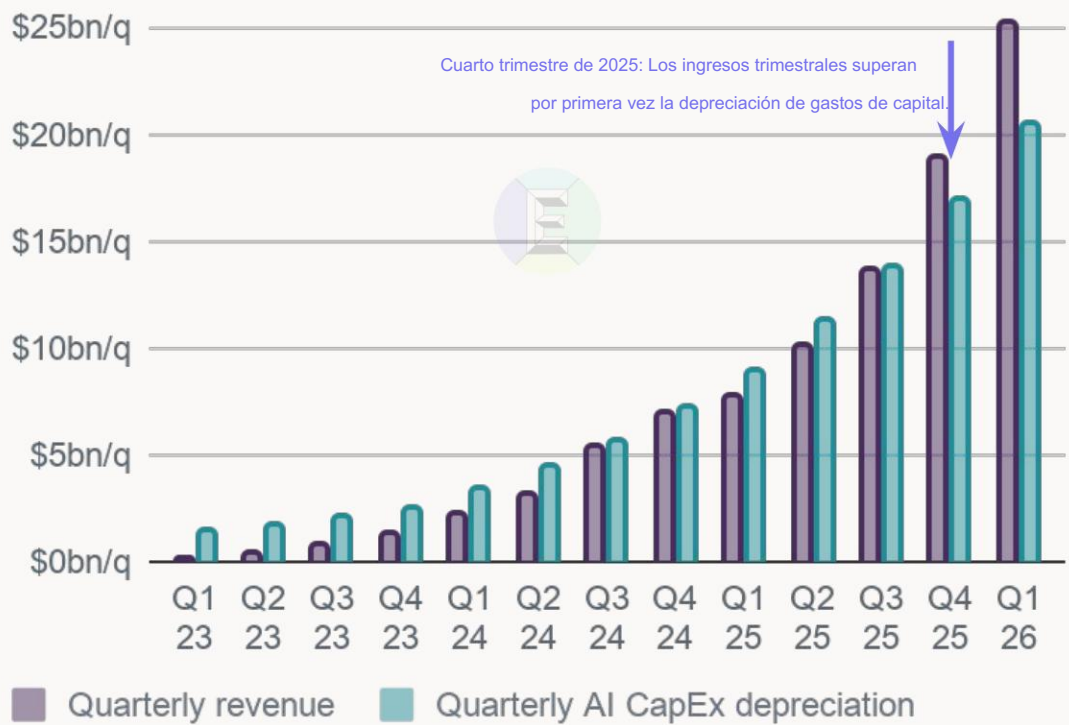
Fuente: Análisis de Exponential View.
Nota: Los equipos informáticos se deprecian a lo largo de 6 años y los edificios a lo largo de 14 años. Los valores de ingresos requeridos excluyen los gastos operativos. El margen de maniobra es la parte de los ingresos que excede la necesaria para cubrir el gasto de depreciación.

3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Los ingresos cubren los gastos corrientes, pero aún no la factura acumulada.

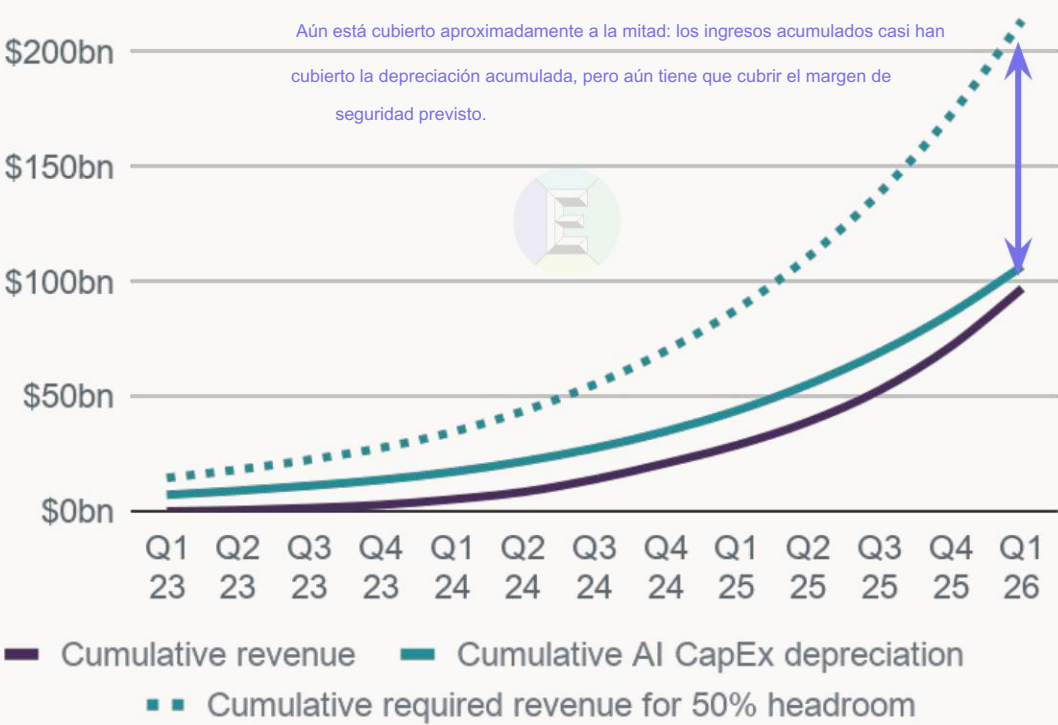
Ingresos trimestrales por IA y depreciación de gastos de capital

Miles de millones de dólares por trimestre, solo hiperescaladores y neoclouds



Depreciación acumulada de ingresos por IA y gastos de capital

Miles de millones de dólares, solo hiperescaladores y neoclouds



Fuentes: Análisis de Exponential View; documentos presentados por la empresa.

Nota: Meta contribuye al gasto de capital de la industria, pero las iniciativas se centran en el aumento de la publicidad, por lo que no se reconocen como ingresos puros de GenAI, o actualmente tienen una monetización directa mínima (por ejemplo, el asistente Meta AI, Muse Spark).

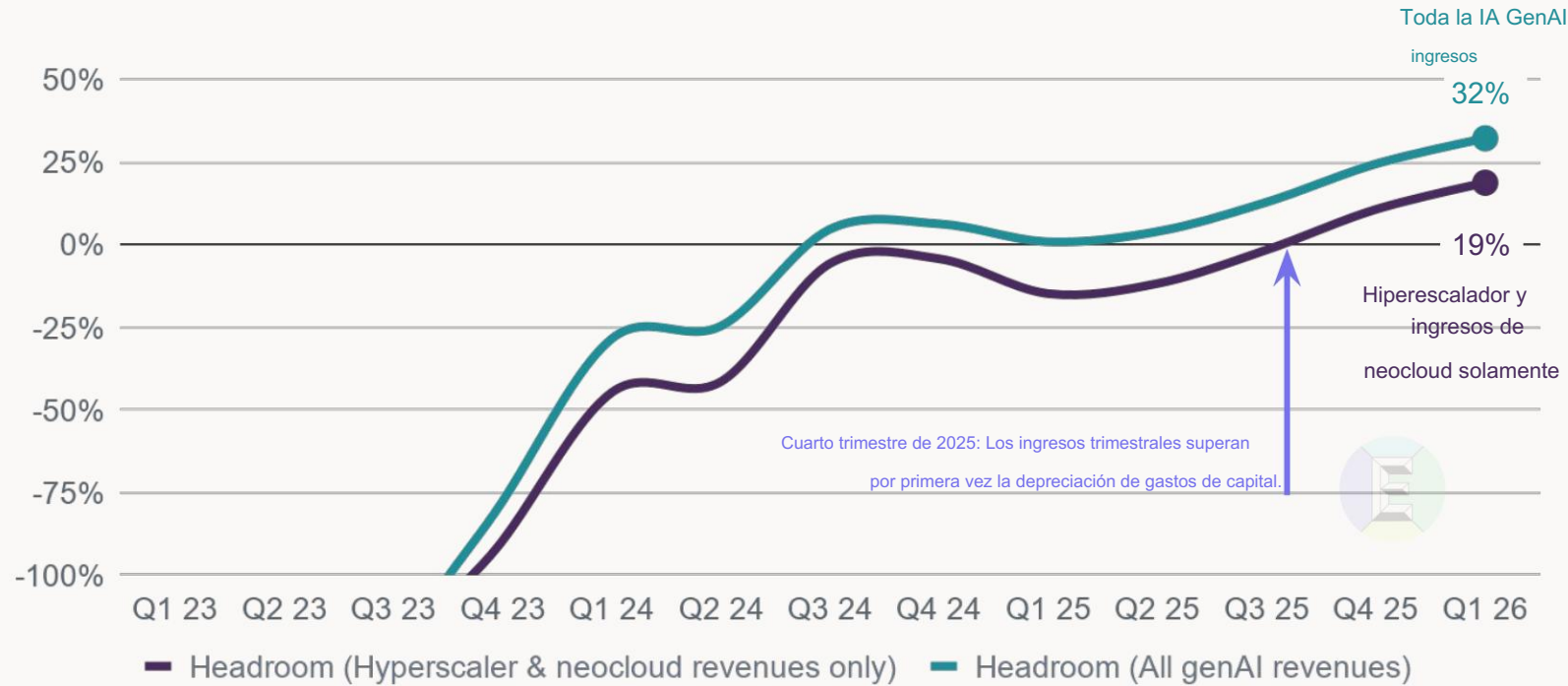


3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Los ingresos por infraestructura de IA acaban de superar el obstáculo de la depreciación de hoy.

Margen de maniobra tras la depreciación trimestral de gastos de capital.

$$\% = (\text{Ingresos} - \text{Depreciación}) \div \text{Ingresos}$$



- Los ingresos de GenAI ahora cubren los Depreciación trimestral de la infraestructura de IA. El margen de crecimiento del primer trimestre de 2026 alcanzó el 19 % para los ingresos de hiperescaladores/neocloud y el 32 % para todos los ingresos de GenAI.
- La cobertura sigue siendo escasa. La depreciación absorbe aproximadamente el 81 % de los ingresos de GenAI de los hiperescaladores/neocloud y el 68 % de los ingresos totales de GenAI antes de los costes adicionales.
- La siguiente prueba es incremental. cobertura. A medida que las inversiones de capital comprometidas en IA entren en servicio, la base de depreciación aumentará. El crecimiento de los ingresos, la utilización y los precios deben seguir aumentando de forma acumulativa o el margen de maniobra se reducirá nuevamente.

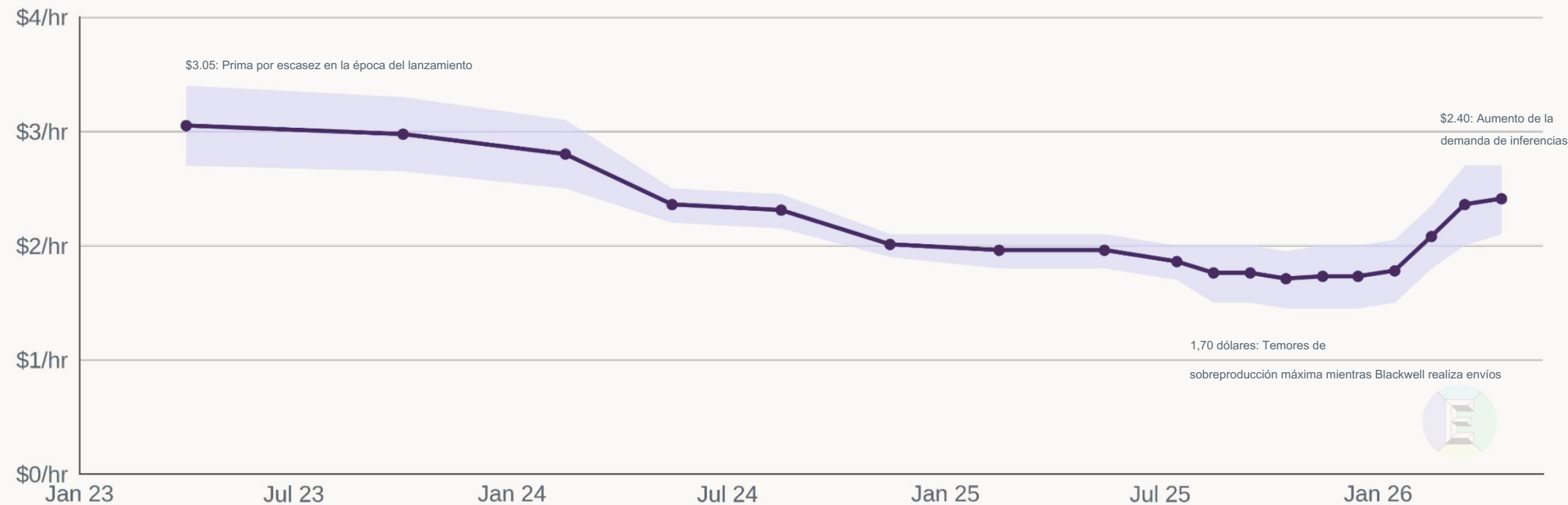


3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Los precios de los alquileres sugieren que la demanda está absorbiendo la oferta existente.

Precio del contrato de alquiler de 1 año para H100

\$/hora/GPU. H100 es la GPU más líquida y más negociada en el mercado de comerciantes: una señal útil.



3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

La economía de los centros de datos supone un obstáculo para la fijación de precios de los tokens.

Coste anual de poseer y operar 1 GW de capacidad de IA: 7.900 millones de dólares.

Costes de capital: 7.000 millones de dólares al año · 89%



Gastos operativos: 900 millones de dólares al año · 11%



Fuentes: Análisis de Exponential View; Epoch AI; SemiAnalysis.

Nota: Modelo ilustrativo. 1 GW de capacidad de TI (6,7k GB200 NVL72 sistemas, 480k GPU), costo de propiedad por Epoch AI (mayo de 2026), anualizado incl. costo de capital, vida útil de TI de 6 años. Salida de tokens de SemiAnalysis InferenceX: FP4, 8k-in/1k-out, 50 tokens/seg/usuario, 65% de utilización (abril de 2026). El rango de salida de tokens refleja un aumento de rendimiento de +10-25% de la decodificación especulativa. Los modelos de peso abierto no incurrir en tarifa de licencia de modelo. La columna de peso cerrado agrega una tarifa de licencia del 25% del precio combinado de Kimi de \$1.29.

- Kimi K2.5 (1T)
- Modelo de clase Kimi bajo licencia cerrada (ilustrativo)

DIVIDIR 7.900 millones de dólares al año POR LA PRODUCCIÓN DE TOKENS:	
÷ TOKENS POR GW / AÑO	
75-85	cuatrillón
= COSTO DEL PROVEEDOR DE INFERENCIA	
\$0.10	\$0.42
Tarifa de licencia de \$0.32 (aproximadamente el 25% del precio de venta combinado de \$1.29)	
PRECIO REQUERIDO PARA UN 50-75% DE BRUTO MARGEN:	
\$0.20 – \$0.40	\$0.84 – \$1.68
VALOR REQUERIDO PARA EL CLIENTE PARA UN RETORNO DE LA INVERSIÓN DEL 25 %	
\$0.25 – \$0.50	\$1.05– \$2.10

3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Los rendimientos brutos de alquiler sugieren que la vida útil se extiende más allá de los seis años.

Rendimiento de la GPU con un 50% de utilización

%, excluye gastos operativos



Fuentes: Análisis de Exponential View; Silicon Data.
Nota: Rendimiento = (Tasa bajo demanda x 50% de utilización x 8760 horas) ÷ precio de lista original.

3 | Inversión de capital: La mayor inversión en el sector tecnológico está dando sus frutos (por ahora).

Una mayor vida útil de la GPU amplía el margen de maniobra.

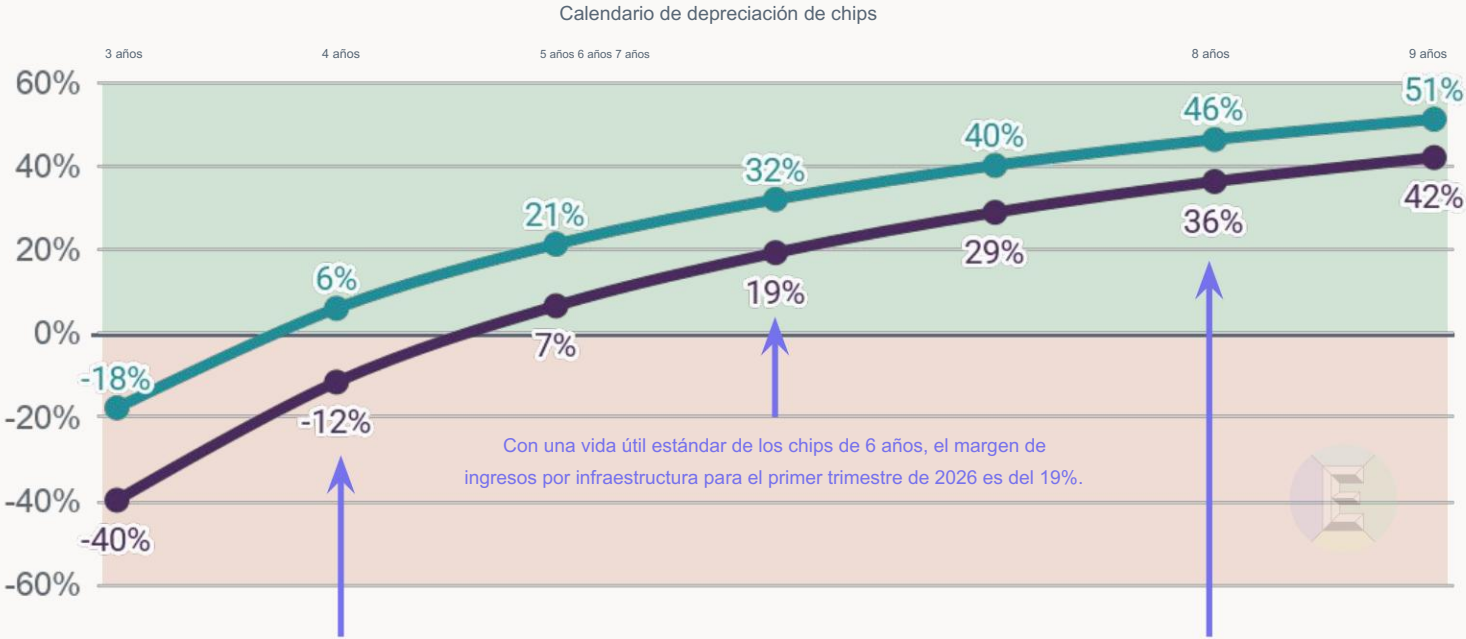
Rango de margen de maniobra después de la depreciación de gastos de capital por chip.
Primer trimestre de 2026, cronogramas de 3 a 9 años, % = (Ingresos – Depreciación) ÷ Ingresos

Mark Zuckerberg

Conferencia telefónica de Meta sobre los resultados del tercer trimestre de 2025

"...el peor escenario posible sería que, en la práctica, hubiéramos preconstruido algo durante un par de años, en cuyo caso, por supuesto, habría algunas pérdidas y depreciación, pero nos adaptaríamos y lo usaríamos con el tiempo."

La sobreconstrucción puede ser una apuesta por una mayor depreciación de los chips.



Si la vida útil del chip es más corta, los ingresos no cubren el CapEx (esto requeriría una inversión inicial).
(Las compras del H100 se están volviendo obsoletas hoy en día)

Prolongar la vida útil de los chips aumenta los márgenes de beneficio:
El uso de chips durante 8 años (tan antiguos como los T4) genera margen de infraestructura hasta el 36%

Fuentes: Análisis de Exponential View; Meta.
Nota: Los edificios se deprecian a lo largo de 14 años como si fueran constantes.





4 | Fichas:

¿La unidad de valor para la economía de la IA?

El volumen de tokens crece 14 veces al año, impulsado por la carga de trabajo de los agentes y una demanda altamente elástica. La fijación de precios basada en tokens ha hecho que esto sea especialmente relevante, pero también representa una oportunidad para que la industria atribuya y evalúe el resultado del consumo de tokens.



“La entrada son electrones, la salida son tokens.
En el medio está Nvidia.”

– Jensen Huang

“Los tokens, las unidades fundamentales de
datos que procesan nuestros modelos...”

– Sundar Pichai

¿La economía de GenAI es una economía de fichas?

Más o menos.

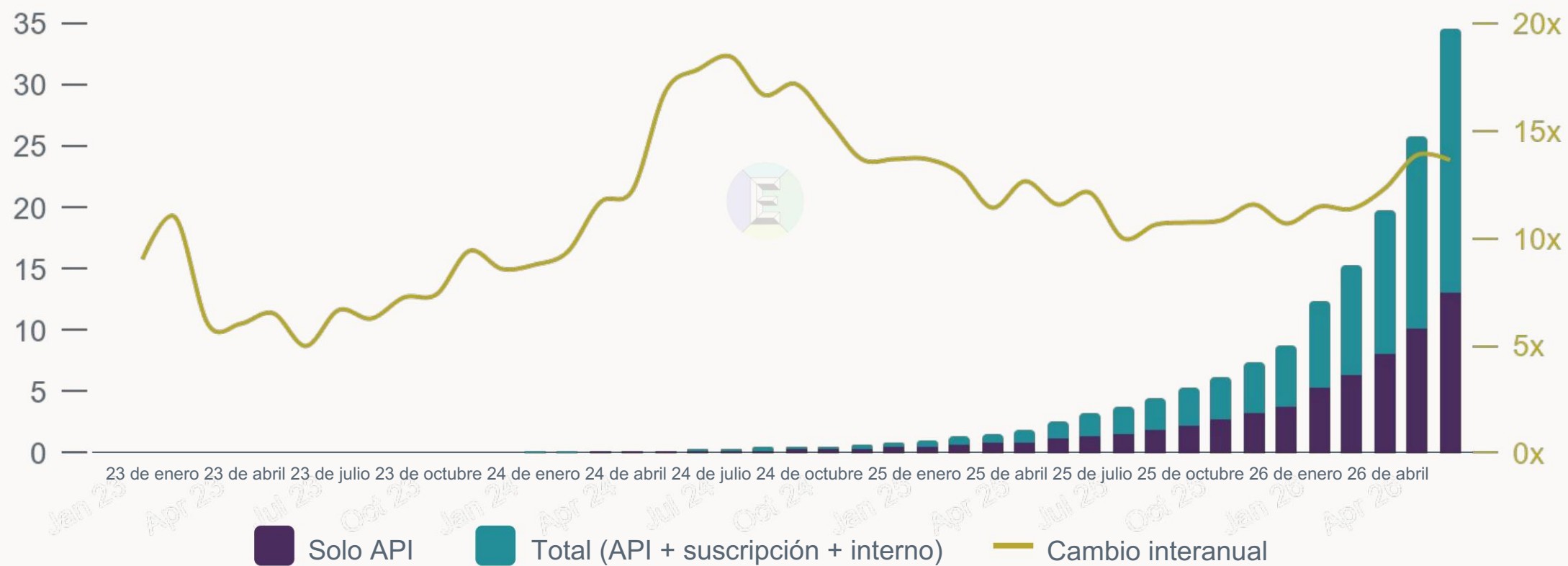


4 | Tokens: ¿La unidad de valor para la economía de la IA?

El volumen global de tokens supera los 30 TC/mes, con un crecimiento interanual de 14 veces.

Tokens de inferencia procesados

Cuatrillones de tokens por mes (eje izquierdo), múltiplo de tasa de crecimiento (eje derecho)



Fuentes: Análisis de Exponential View.
Nota: Global, inc. China



4 | Tokens: ¿La unidad de valor para la economía de la IA?

La transición del chat a los agentes está multiplicando el uso de tokens.

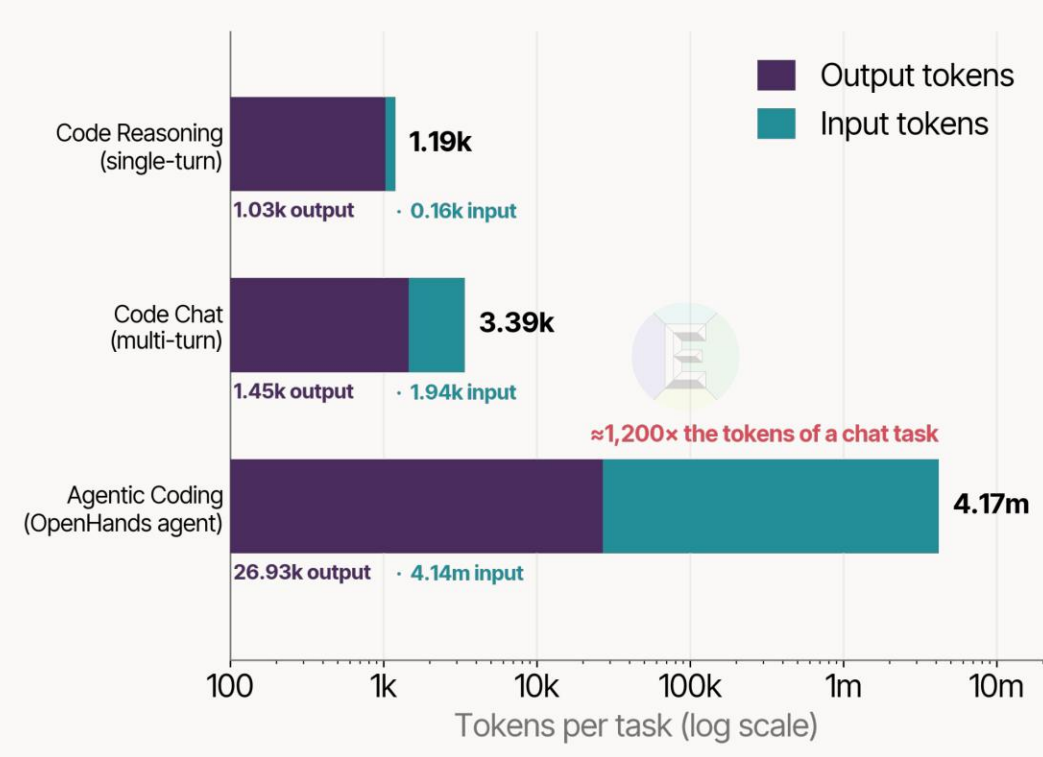
Densidad de coordinación de agentes

% de uso de herramientas por solicitud, OpenRouter



Consumo de tokens por tarea

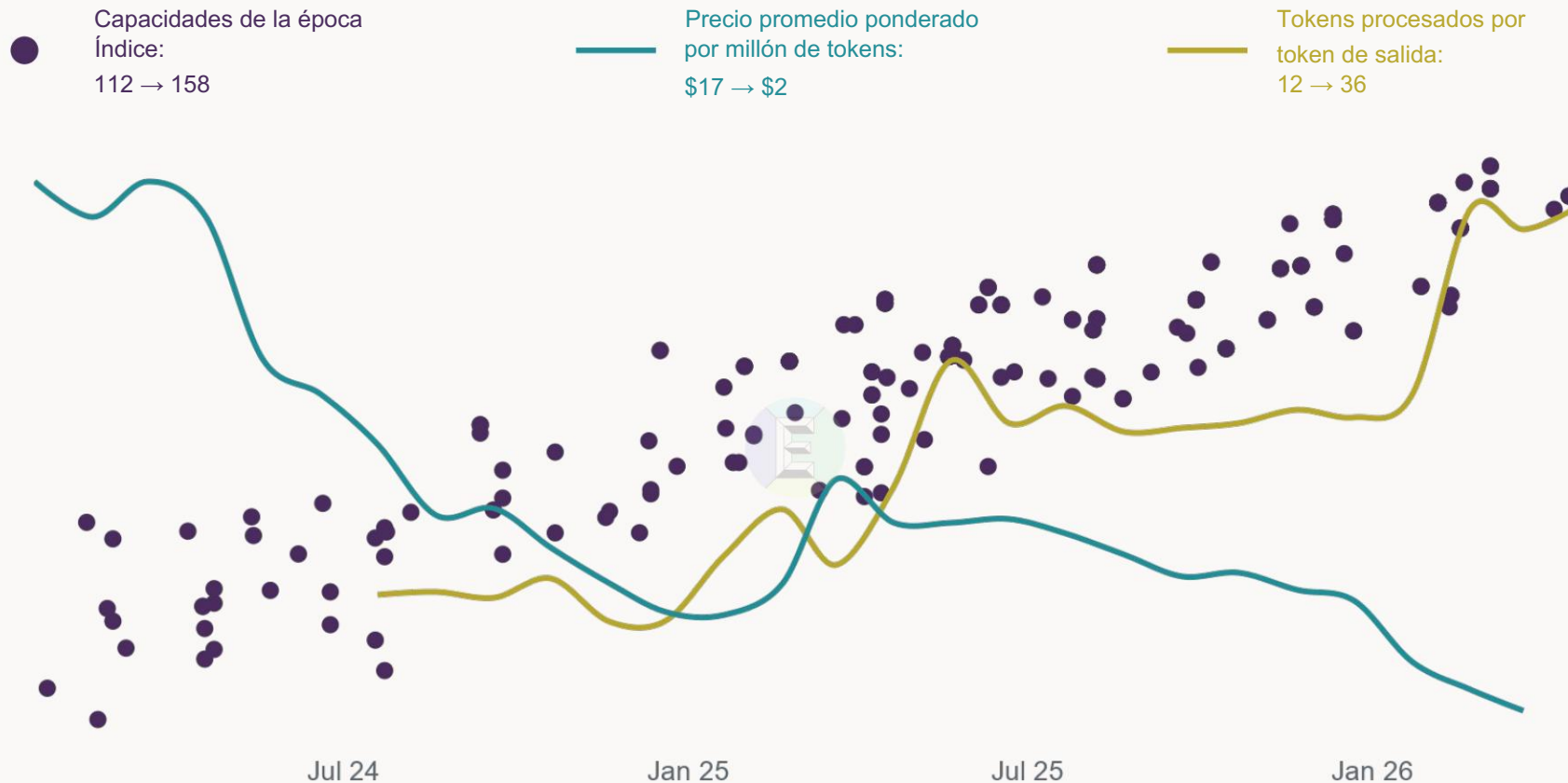
Promedio de tokens consumidos por tarea, escala logarítmica



Fuentes: Análisis de Exponential View; OpenRouter; Bai, Huang, Wang, Sun, Mihalcea, Brynjolfsson y Pentland 2026.

4 | Tokens: ¿La unidad de valor para la economía de la IA?

Los tokens más baratos y los mejores modelos amplifican la demanda.

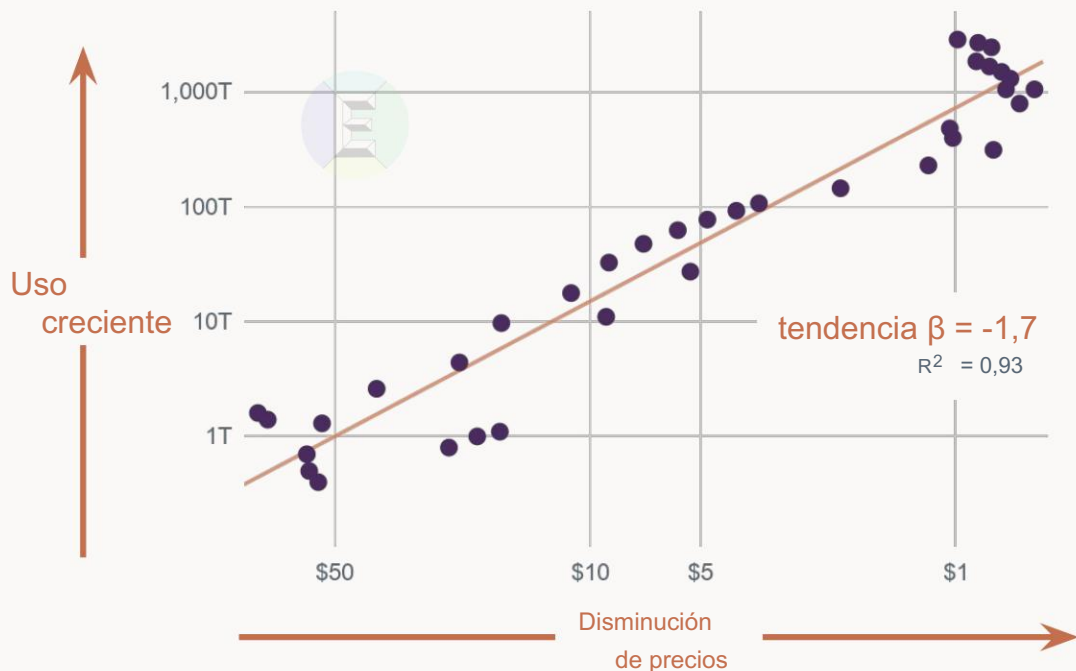


- Modelos más capaces que puedan cubrir una gama más amplia de tareas económicamente útiles, aumentando así su valor y uso.
- El volumen de tokens está aumentando a medida que
Los modelos de razonamiento gastan más tokens en “pensar”.
- La bajada de precios fomenta un mayor uso y hace viables aplicaciones que antes no resultaban rentables.



La demanda de tokens parece elástica: a medida que bajan los precios, el uso crece más rápido.

Elasticidad de precios de Google (precio promedio vs. volumen) \$/millón de tokens vs. billón de tokens/mes, escala logarítmica doble, 2023-2026



Sundar Pichai, Google: I/O 2025: "...procesábamos

9,7 billones de tokens al mes. Ahora, más de 480 billones, 50 veces más".

Precio -97% | Volumen 50x

Sam Altman, OpenAI: "Tres observaciones", 2025

"El coste de utilizar un determinado nivel de IA disminuye aproximadamente diez veces cada 12 meses, y los precios más bajos conllevan un uso mucho mayor".

Precio -90% | Volumen ↑

Tan Dai Volcengine / ByteDance, 2025

"El uso diario de tokens de Doubao superó los 50 billones este mes, frente a los 4 billones de diciembre de 2024."

Precio -50% | Volumen 12x

En todos los proveedores, la magnitud de la elasticidad es de aproximadamente 1,2 a 1,8 : cada reducción de precio del 10 % genera entre un 12 % y un 18 % más de tokens, lo que a su vez incrementa el gasto total en tokens.



4 | Tokens: ¿La unidad de valor para la economía de la IA?

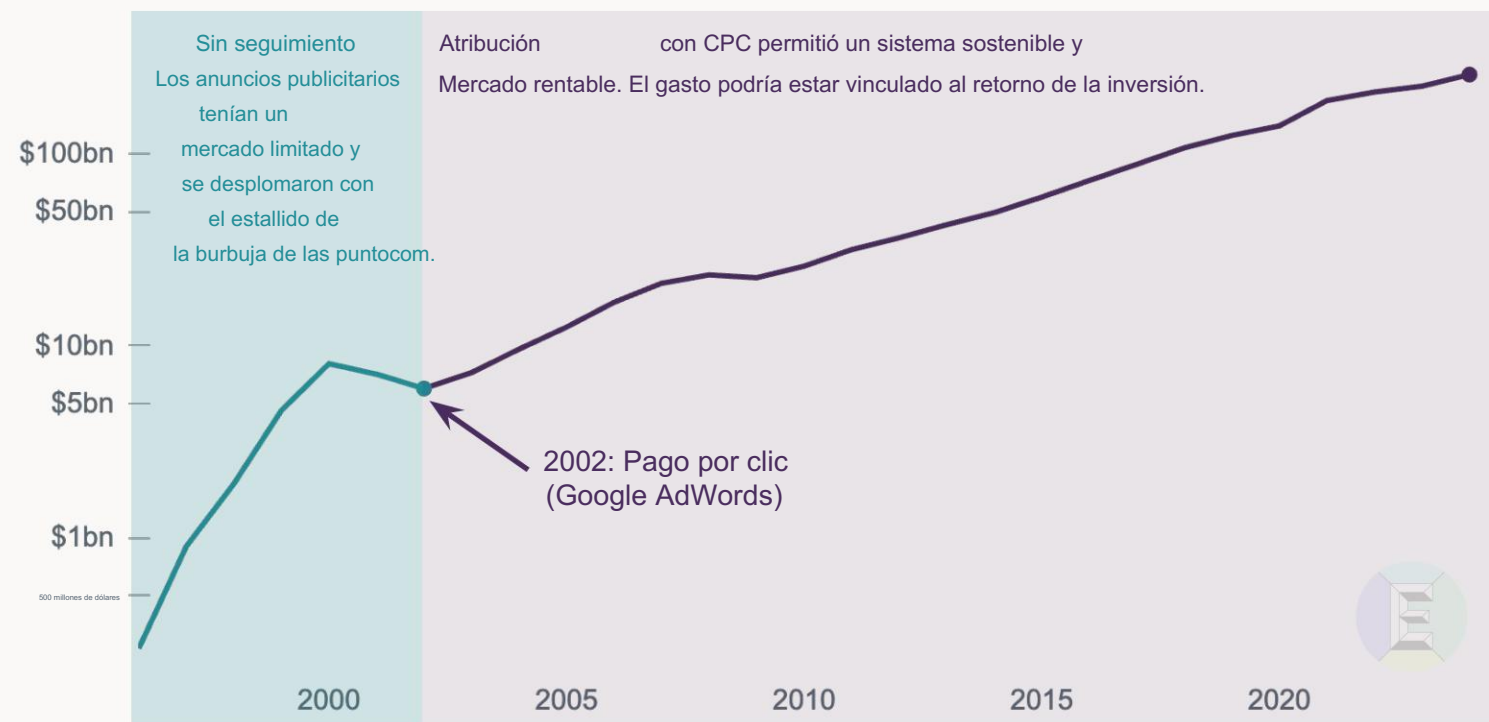
La fijación de precios basada en tokens es el momento de "pago por clic" de la IA.

Evolución de los modelos de precios de la IA

- 1 **GRATIS**
No reconocido en los presupuestos
- 2 **SUSCRIPCIÓN**
Precios basados en asientos sin vinculación a un valor específico.
- 3 **PRECIOS BASADOS EN TOKENS**
El uso medido permite (y requiere) la atribución a proyectos.

Ingresos anuales por publicidad digital

Miles de millones de dólares al año, escala logarítmica, 1996-2024



4 | Tokens: ¿La unidad de valor para la economía de la IA?

Cada generación aumenta la producción de tokens por gigavatio de capacidad.

Tokens producidos por gigavatio de capacidad del centro de datos

Billones de tokens por GW al mes



- Cada gigavatio permite comprar más tokens de producción cada mes.
- Esto permite un superciclo, incluso cuando las limitaciones físicas ejercen presión sobre la IA economía.
- Impulsado por:
 - a) Laboratorios que logran un mayor Eficiencia mediante modelos más pequeños y un mejor servicio.
 - b) Ganancias de hardware, aunque de movimiento más lento (por ejemplo, Hopper → Blackwell → Rubin).
 - c) Cambio en la distribución de la carga de trabajo hacia la inferencia frente al entrenamiento.

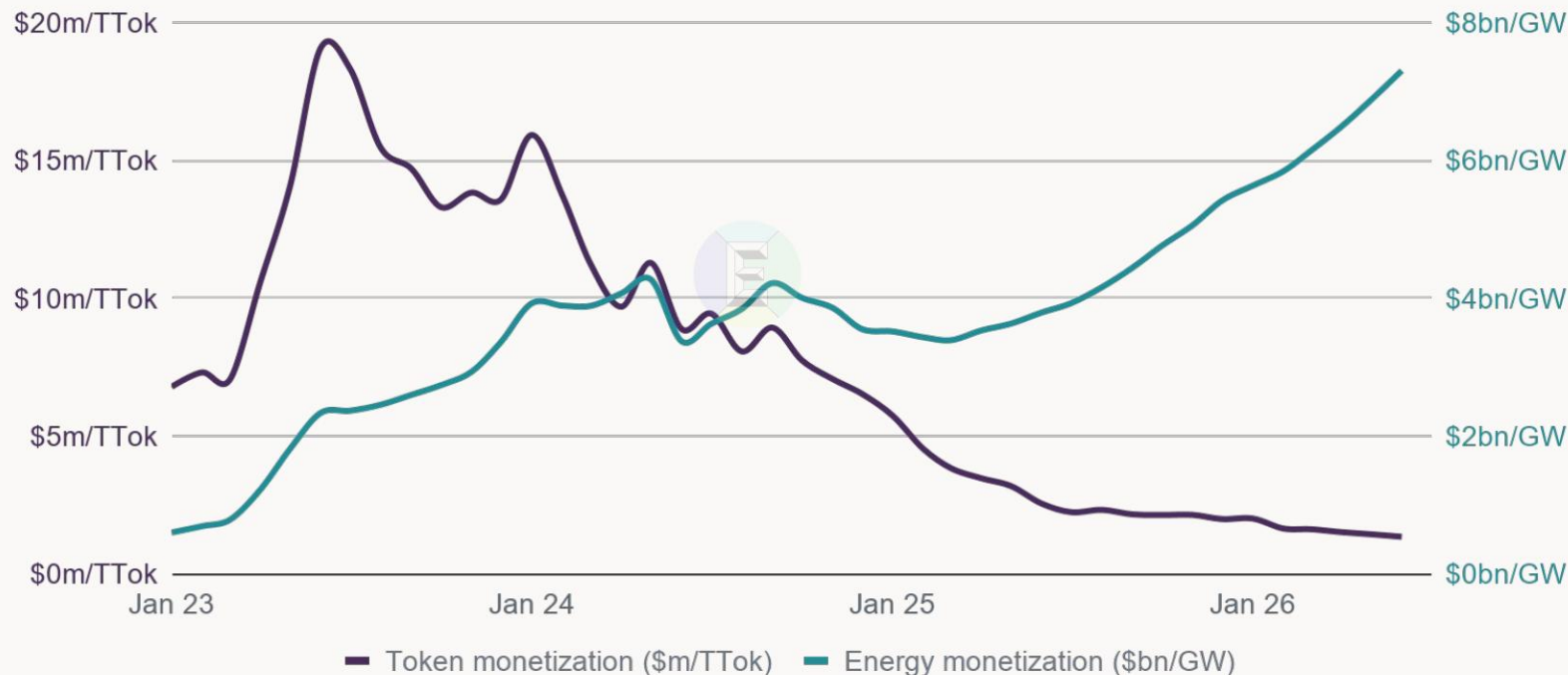


4 | Tokens: ¿La unidad de valor para la economía de la IA?

Esta eficiencia está aumentando la monetización por GW de capacidad mientras que los ingresos por token disminuyen.

Ingresos generados por tokens y capacidad del centro de datos

Millones de dólares/TTok (eje izquierdo), miles de millones de dólares/GW (eje derecho)



- Ingresos por billón de tokens ha caído desde su máximo de 2023, reflejando la caída de los precios.
- Las mejoras en la eficiencia provocan una bajada de los precios de los tokens, que se ve más que compensada por una mayor demanda.
- Los ingresos del sector por GW de capacidad de centros de datos superaron los 7.000 millones de dólares/GW.



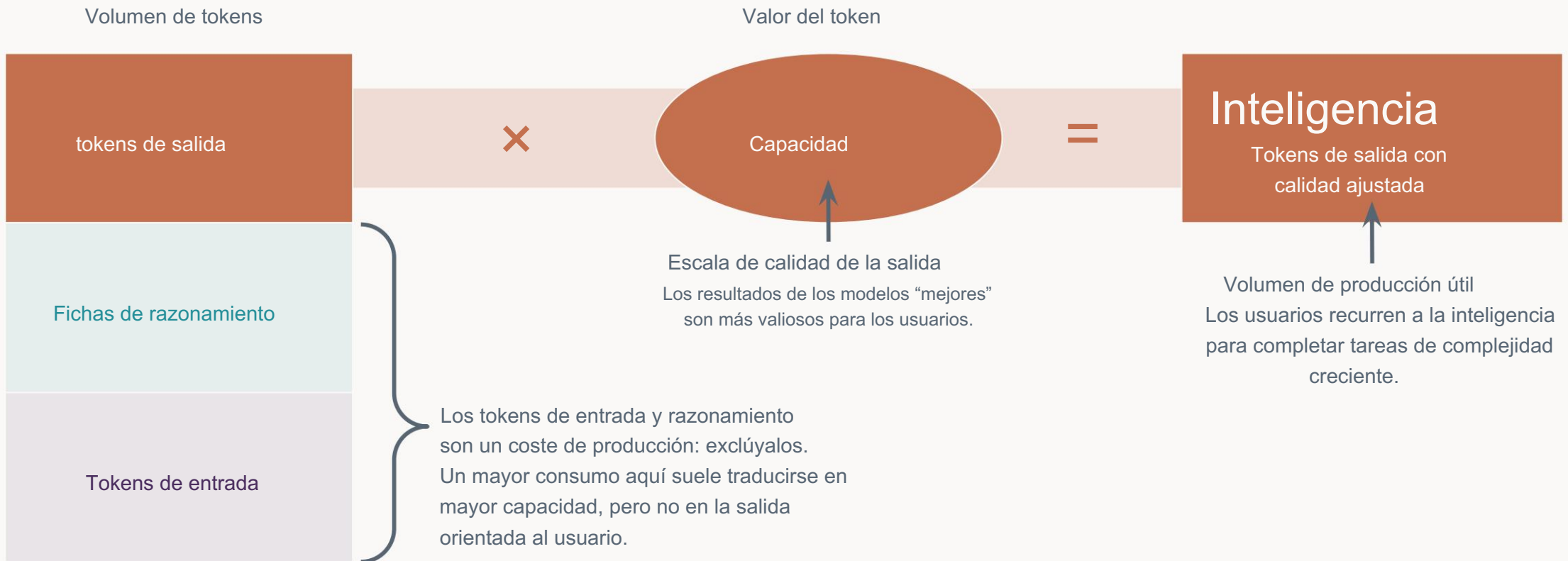
4 | Tokens: ¿La unidad de valor para la economía de la IA?

Los tokens son la métrica de facturación de la IA, pero aún no son una unidad de valor.



4 | Tokens: ¿La unidad de valor para la economía de la IA?

Los tokens ajustados a la calidad son los que más se aproximan a una unidad de valor utilizable.

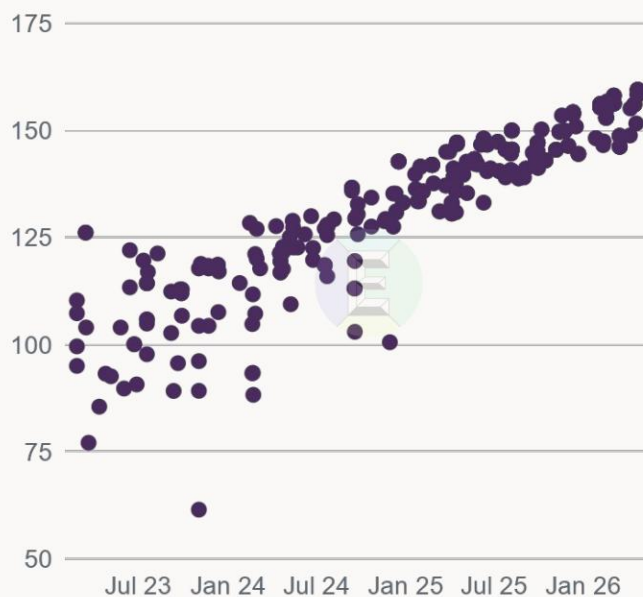


4 | Tokens: ¿La unidad de valor para la economía de la IA?

Independientemente de la medida que elija, la tendencia es al alza.

Índice de Capacidades de Época

Puntuación: GPT-5 = 150,
Claude 3.5 Soneto = 130



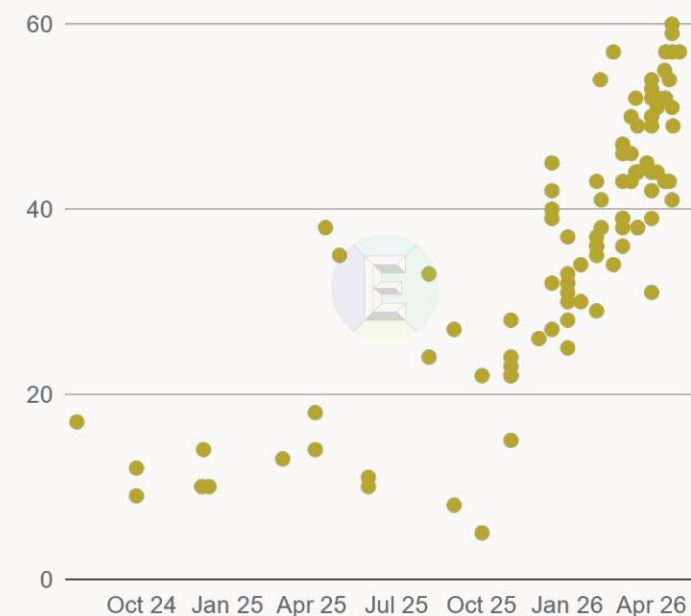
Horizonte de tareas de METR

Duración de la tarea humana con una tasa de
éxito del modelo del 50 %, escala logarítmica



Inteligencia artificial de análisis Índice

Puntuación: 0-100



Fuentes: Análisis de Exponential View; Epoch AI; METR; Artificial Analysis.

Nota: Las tareas de METR consisten principalmente en tareas de ingeniería de software, aprendizaje automático y ciberseguridad.

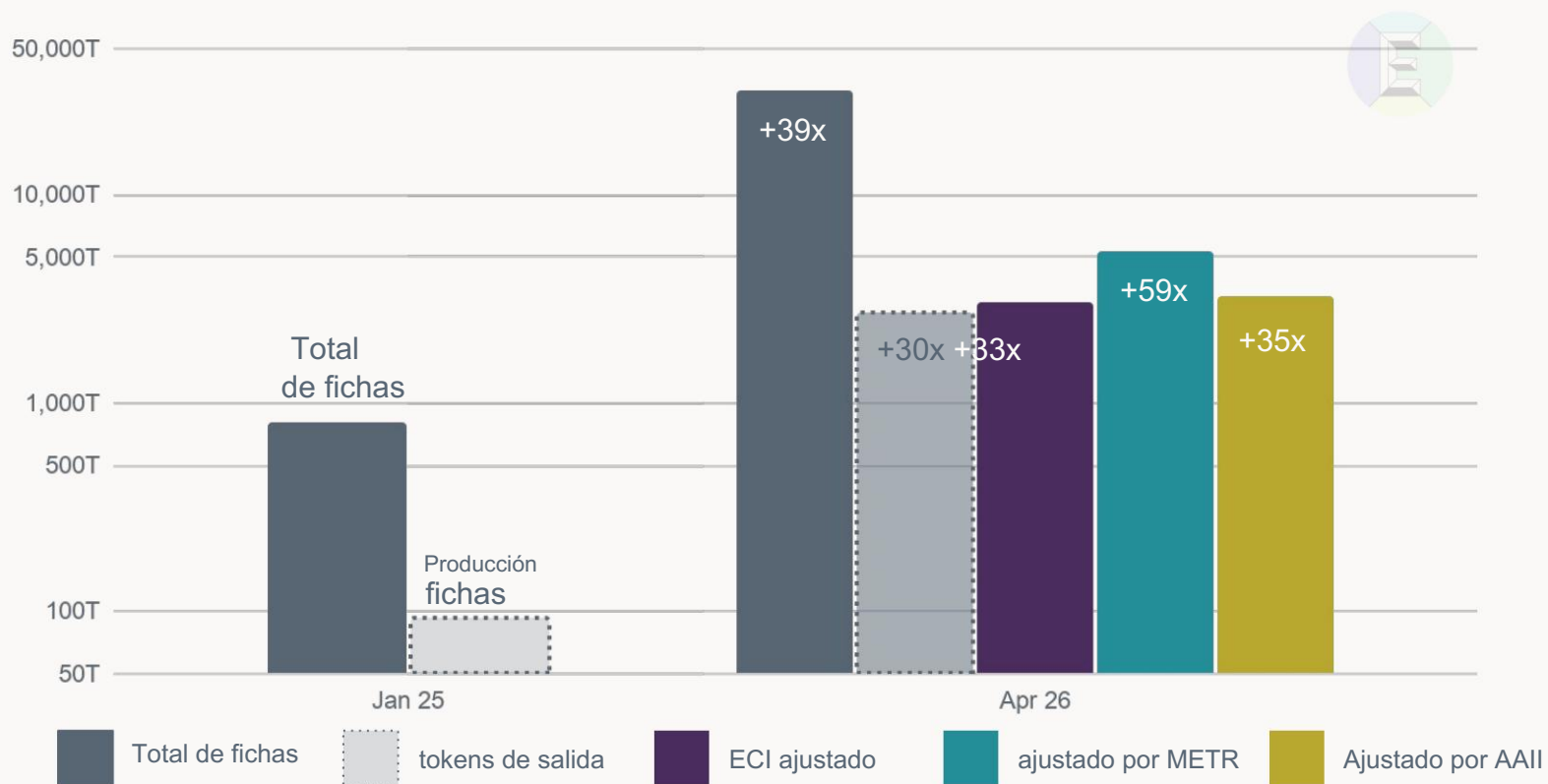


4 | Tokens: ¿La unidad de valor para la economía de la IA?

La producción ajustada a la calidad se mantuvo al ritmo del crecimiento del volumen bruto.

Tokens: Total, producción y calidad ajustada

Billones de tokens al mes, enero de 2025 frente a abril de 2026



- Los tokens de salida sin procesar crecieron más lentamente que el total de tokens (30x frente a 39x): Aumento del volumen dedicado a la entrada y al razonamiento.

- Amplia dispersión en la puntuación mejora (33x-59x):
Diferentes medidas y diferentes
Las escalas de puntuación implican que solo podemos extraer conclusiones orientacionales.

- Los tokens de salida ajustados a la calidad parecen estar creciendo: tanto los volúmenes de producción como las capacidades son superiores a los de enero de 2025.

Fuentes: Análisis de Exponential View; OpenRouter; Epoch AI; METR; Artificial Analysis.

Nota: Los índices ECI, METR y AAIL se han ajustado al promedio de las puntuaciones de enero de 2025 con fines comparativos.





5 | Pila:

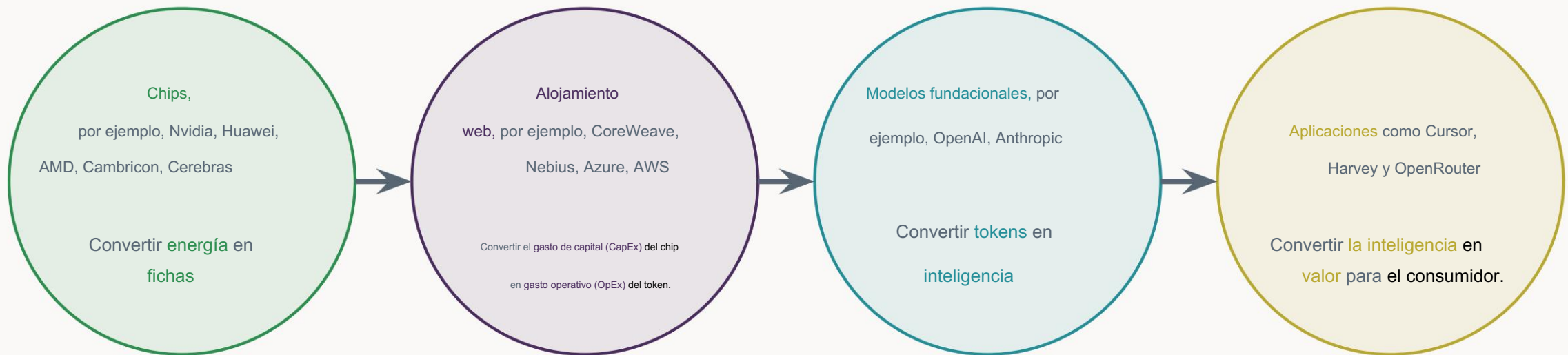
Donde se captura el valor

Los ingresos están concentrados, pero las aplicaciones y los modelos están ganando cuota de mercado. Los laboratorios mantienen sus precios solo mientras controlan la vanguardia; el valor económico de la vanguardia de ayer se desvanece rápidamente ante la falta de competencia. La acumulación de márgenes en toda la cadena de valor dependerá de los avances técnicos, que trascienden la dinámica competitiva.



5 | Pila: Donde se captura el valor

La pila convierte el capital y la energía en trabajo cognitivo.

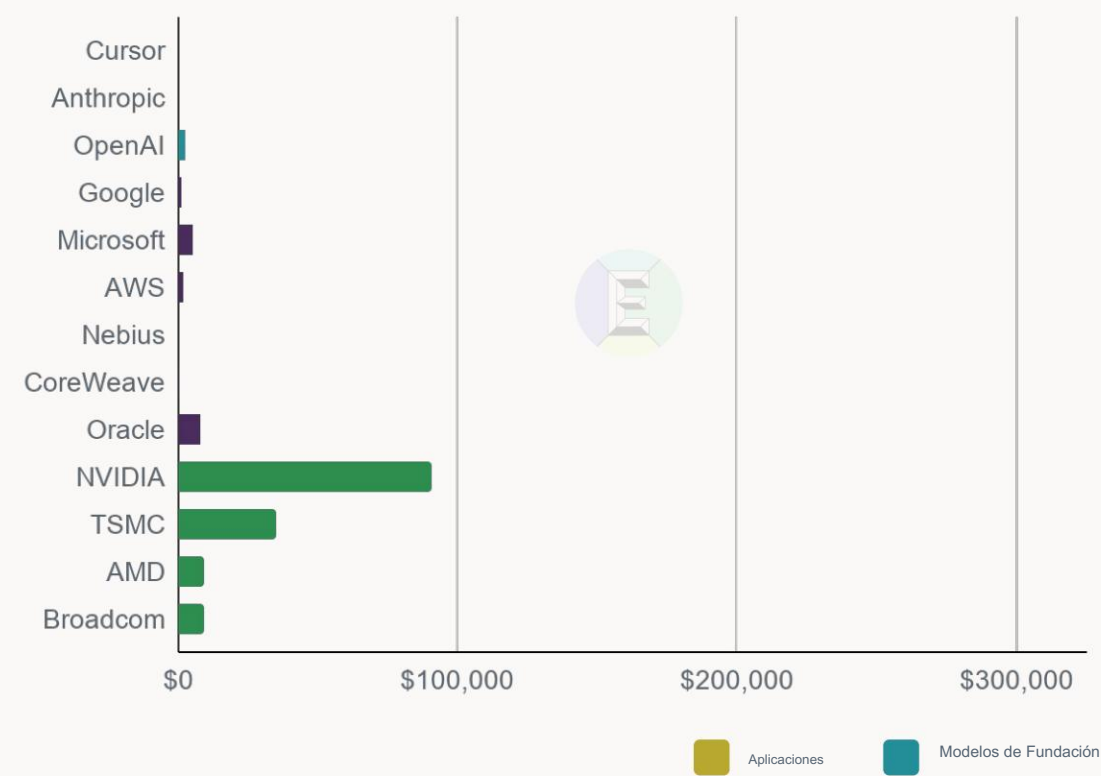


5 | Pila: Donde se captura el valor

Actualmente, los ingresos están concentrados, pero la composición está cambiando.

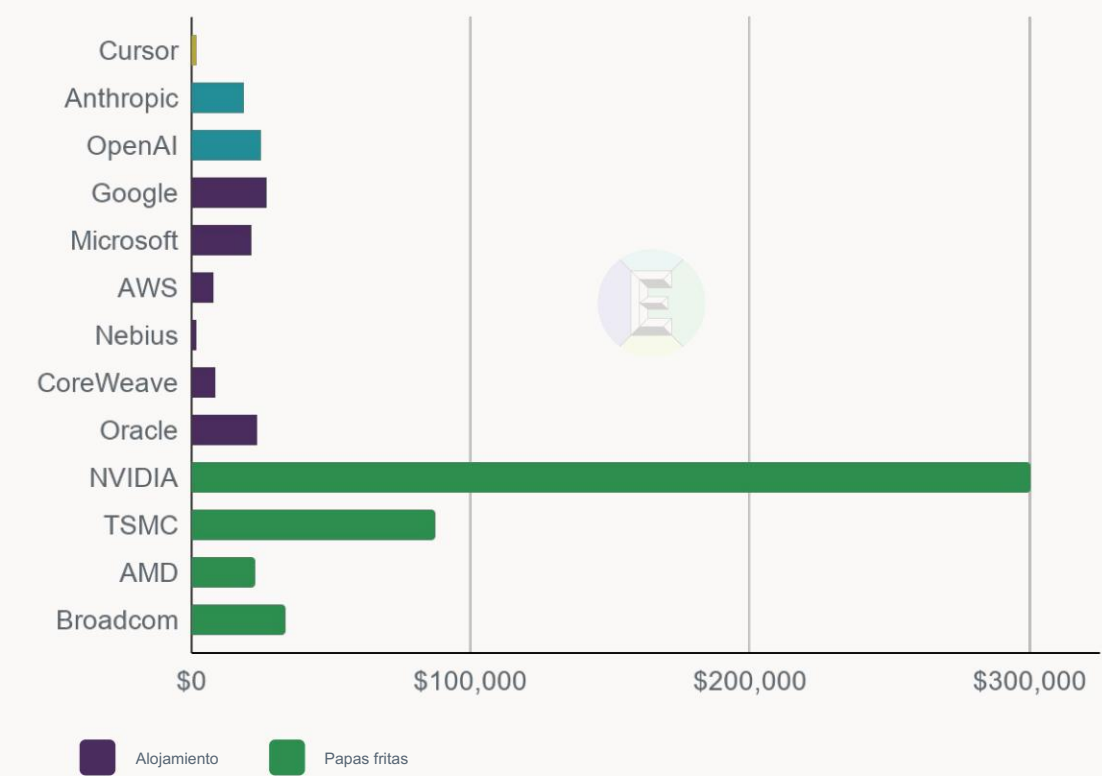
Ingresos anualizados de GenAI

Millones de dólares, primer trimestre de 2024



Ingresos anualizados de GenAI

millones de dólares, primer trimestre de 2026



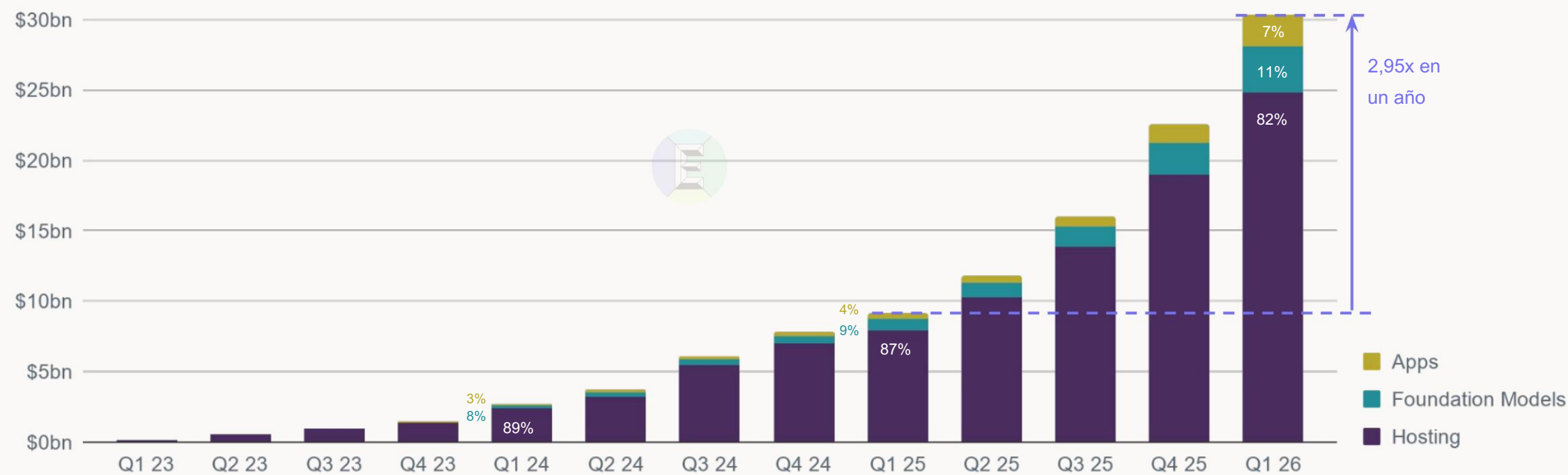
Fuentes: Análisis de Exponential View; documentos presentados por la empresa.
Nota: Los ingresos no están sujetos a ajuste por deduplicación.



5 | Pila: Donde se captura el valor

El valor se está desplazando hacia arriba en la pila, hacia las aplicaciones y los modelos.

Ingresos trimestrales de GenAI, deduplicados por capa para reducir la doble contabilización: miles de millones de dólares por trimestre.



Fuente: Análisis de Exponential View.
Nota: Datos globales excluyendo China. No incluye chips (que la capa de alojamiento contabiliza como gastos de capital).
Es posible que las cifras no sumen el 100% debido al redondeo.

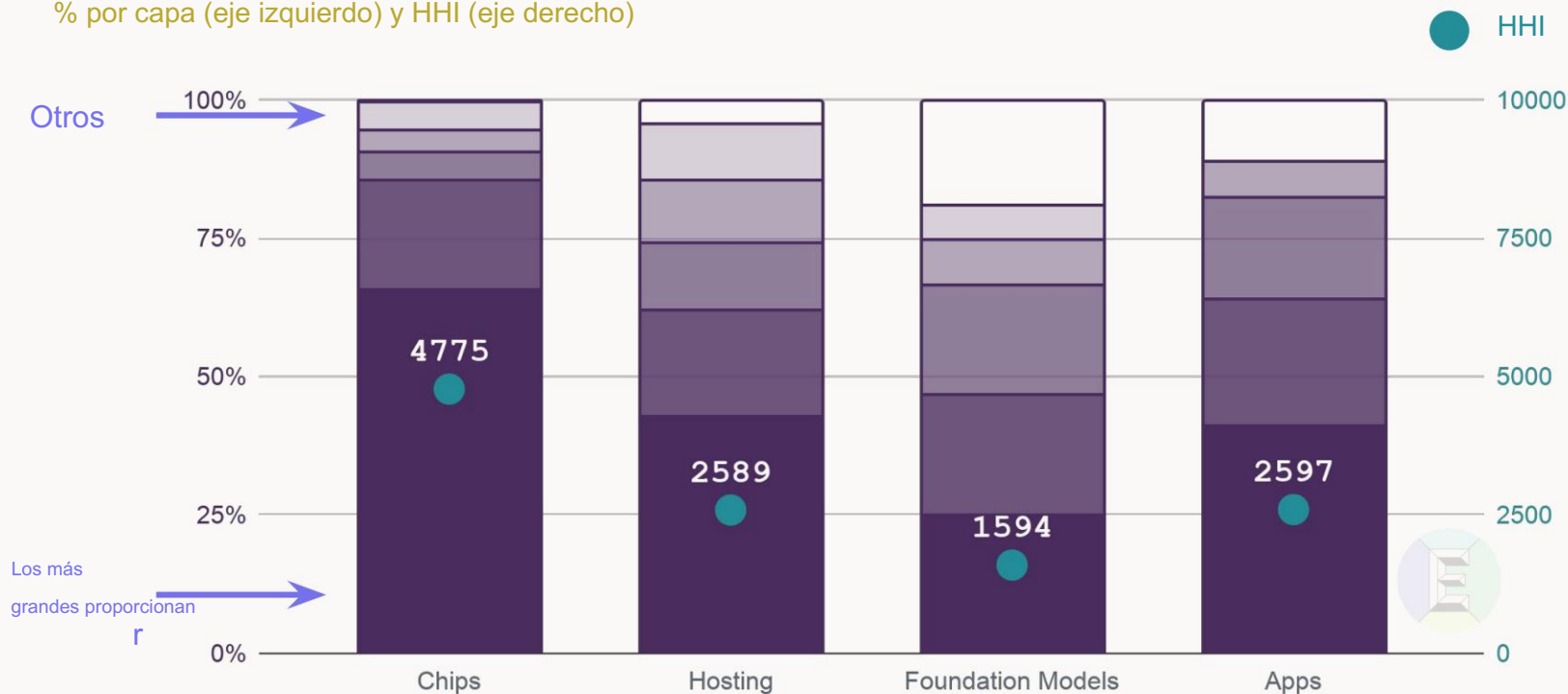


5 | Pila: Donde se captura el valor

El poder de fijación de precios depende de la presión competitiva, no de la posición en el mercado.

Cuota de mercado de las empresas líderes e índice Herfindahl-Hirschman (HHI)

% por capa (eje izquierdo) y HHI (eje derecho)



- Los proveedores de la cadena de suministro fijan precios para los tokens de manera que capturen todo el margen disponible, en ausencia de competencia en la cadena de suministro.
- Nvidia es el mayor proveedor de chips, pero la integración vertical de AWS y Google en la fabricación de silicio personalizado podría reducir su protagonismo.
- Los ingresos de FM se concentran en OpenAI y Anthropic, pero los modelos de ponderación abierta ofrecen una competencia de bajo costo para tokens de calidad.

Fuente: Análisis de Exponential View.

Nota: El alojamiento y las aplicaciones se definen por la proporción de ingresos. Los modelos de fundación se definen por la proporción de tokens (debido a la competencia de ponderación abierta). Los chips se definen por su participación en la capacidad de cómputo (H100-eq) debido a la integración vertical de chips hiperescalares dedicados. El índice Herfindahl-Hirschman es una medida de la concentración del mercado.

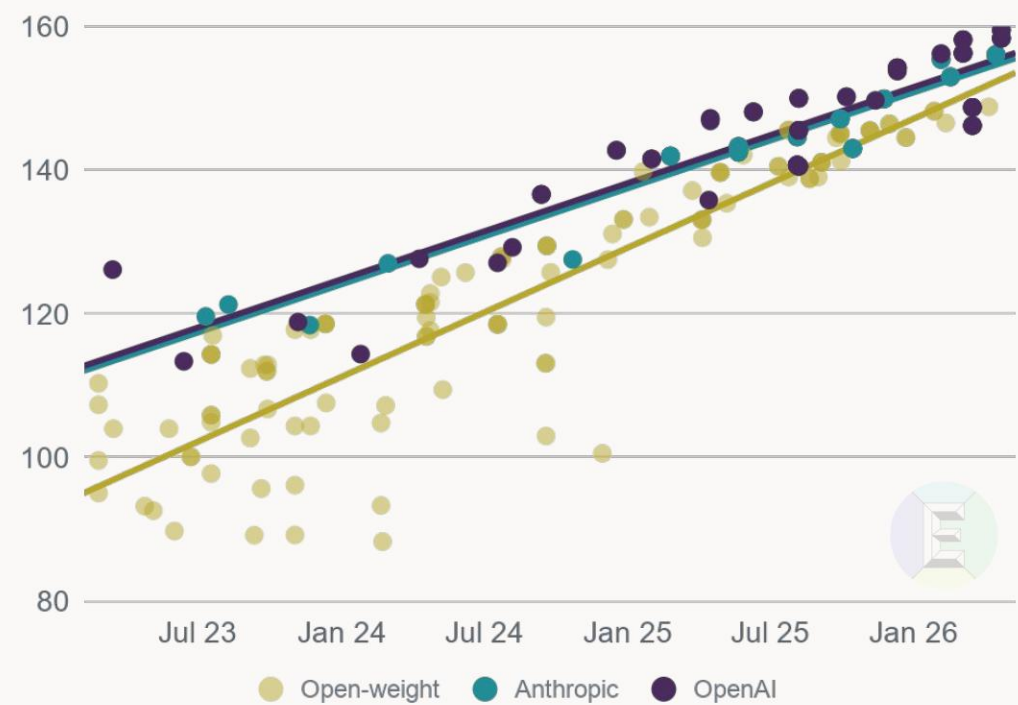


5 | Pila: Donde se captura el valor

Por ahora, Frontier Labs puede justificar sus precios elevados.

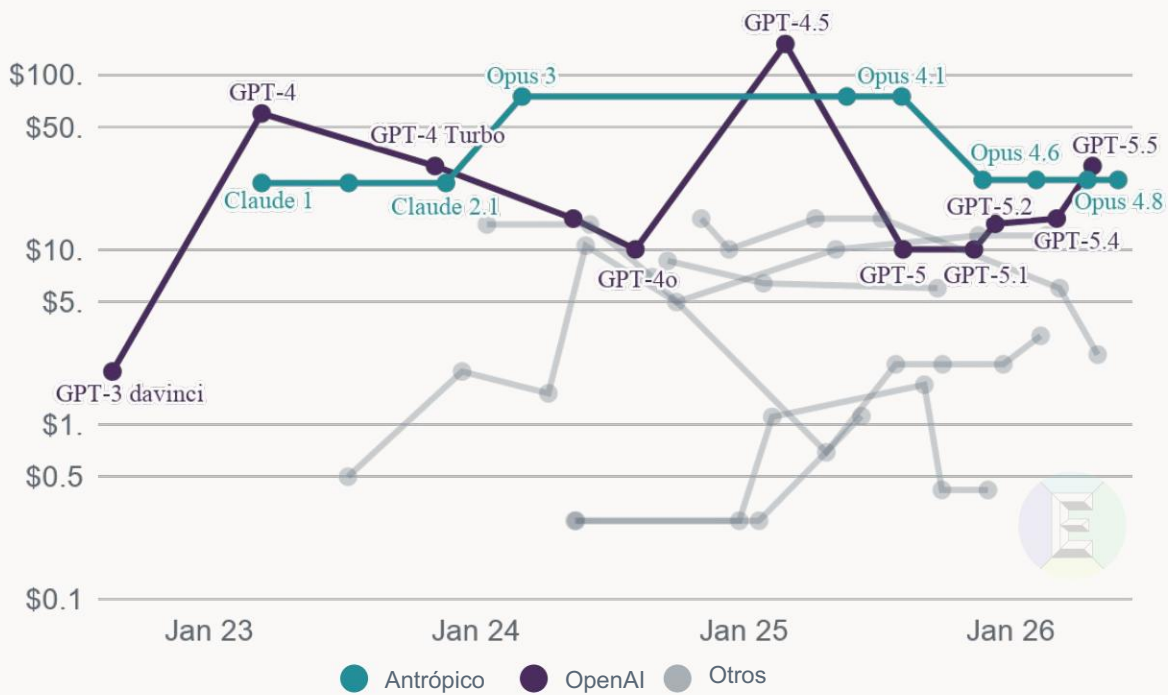
Índice de Capacidades de Época

Puntuación en el momento del lanzamiento: OpenAI/Anthropic/peso abierto



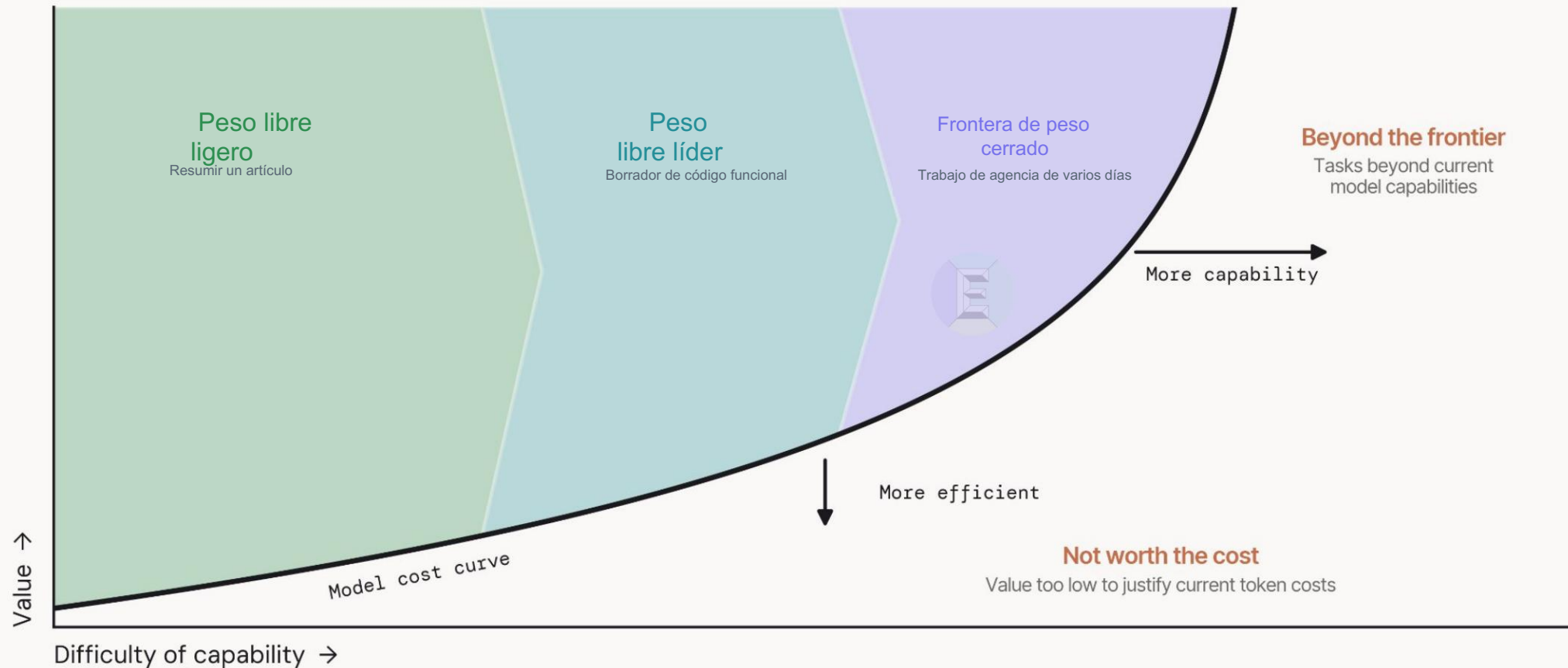
precio de salida

\$/millones de tokens de salida



5 | Pila: Donde se captura el valor

Los laboratorios deben superar la mercantilización de la posición abierta para mantener el margen.

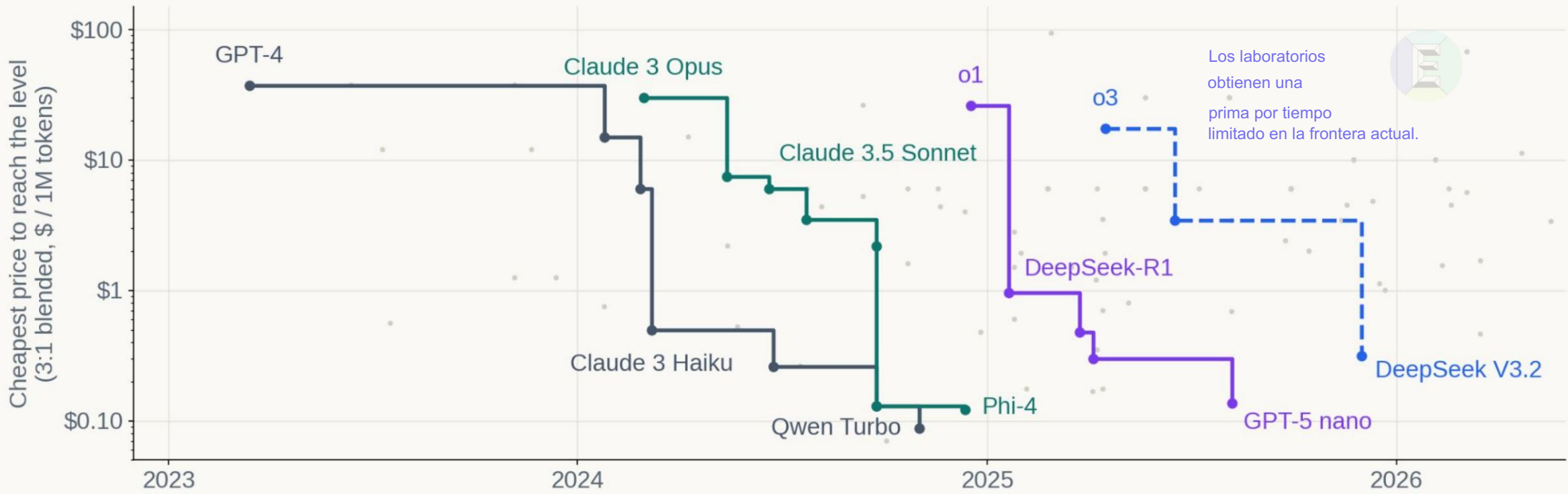


5 | Pila: Donde se captura el valor

La frontera del año pasado se está convirtiendo rápidamente en un bien de consumo.

Precio por frontera de capacidad

Precio combinado en dólares, agrupado por rendimiento en GPQA Diamond (ciencia a nivel de doctorado).

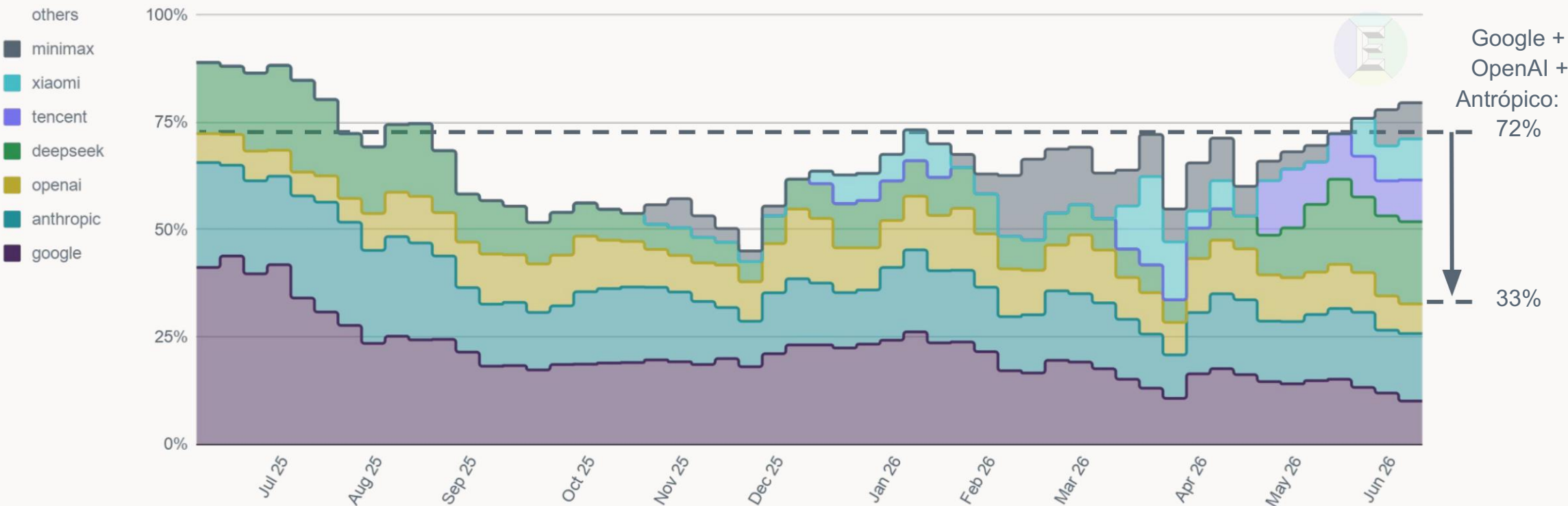


5 | Pila: Donde se captura el valor

Entre los usuarios de OpenRouter que se autoseleccionan, la participación de tokens se está desplazando hacia la ponderación abierta.

Reparto semanal de tokens de OpenRouter

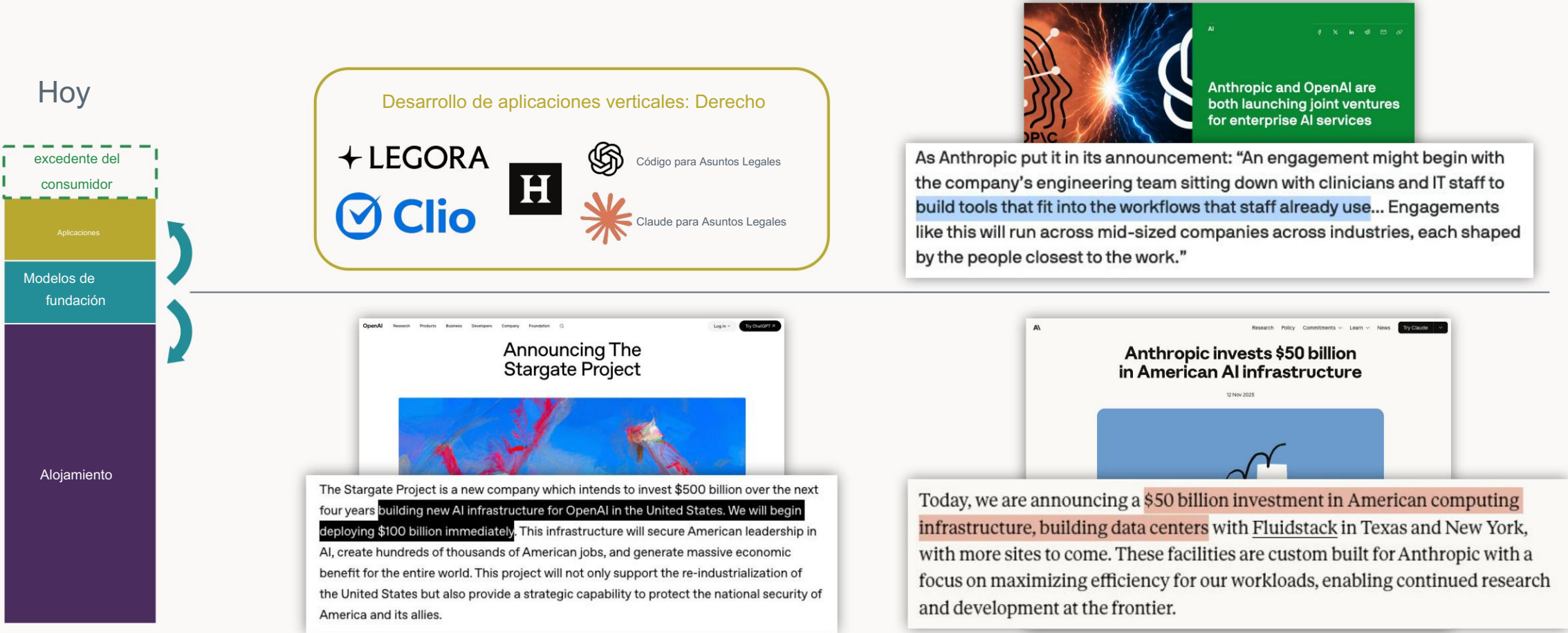
% por autor del modelo



Fuentes: Análisis de Exponential View; OpenRouter.
Nota: Si bien OpenRouter no representa una muestra representativa del mercado, sus datos muestran el comportamiento de los usuarios que optan por el "enrutamiento basado en modelos".

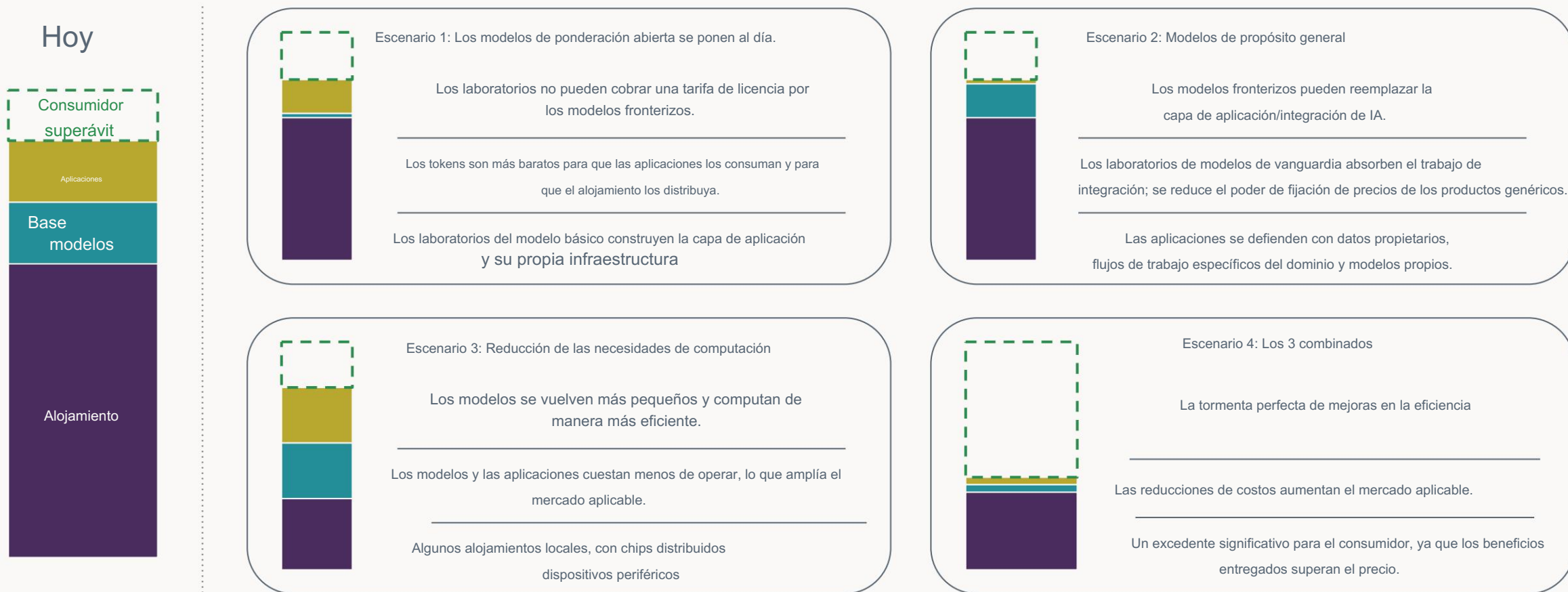


Ante la presión de los precios, los laboratorios se centran en aplicaciones e infraestructura.



5 | Pila: Donde se captura el valor

Comprime cada capa y los consumidores se quedan con el excedente.



Demanda real y elasticidad precio de la demanda:
La reducción de costes impulsa el mercado y genera un mayor excedente para el consumidor.



La demanda de IA está más respaldada por los ingresos que cualquier otro cambio de plataforma anterior.

La clave de la inversión reside en determinar si la caída de los precios puede generar un volumen de tokens suficiente para obtener un retorno sobre la inversión de capital.



Metodología: Cómo contabilizamos los ingresos, el gasto de capital y los tokens.

Lo que incluimos y lo que excluimos

- Contabilizamos los **ingresos** en cada nivel de la pila: aplicaciones (suscripciones y software con prioridad en la IA), API de modelo base y computación y nube de IA vendidas como servicios independientes.
Cada capa representa el gasto real de un cliente de pago. Una empresa que opera en varias capas (por ejemplo, proveedores de modelos básicos con aplicaciones para clientes) distribuye sus ingresos entre ellas. Contabilizamos los ingresos globales fuera de China y excluimos los chips y el hardware (un coste para la capa de computación, no un pago del cliente), las funciones de IA en el software heredado y el aumento de la publicidad derivado de la IA (principalmente Alphabet y Meta). También excluimos el gasto de capital y la financiación de las mediciones de ingresos.
- El **CapEx** se contabiliza como la parte atribuible a la IA del gasto en infraestructura de los siete constructores de pilas (hiperescaladores y neoclouds), incluyendo tanto el efectivo como los activos fijos (PP&E) y arrendamientos.
- Los **tokens** se contabilizan como cada token procesado, de entrada y de salida, en todos los principales proveedores y plataformas.

Cómo eliminamos los duplicados y por qué.

- Los **ingresos** se contabilizan en cada capa, pero nunca se suman entre ellas: un gasto de \$100 en la aplicación que envía \$60 a un proveedor de modelos que a su vez gasta \$30 en alojamiento en la nube para inferencia se atribuye (en línea con el valor añadido) \$40/\$30/\$30 para sumar los mismos \$100 sin doble/triple conteo, lo que de otro modo daría como resultado una cifra errónea de \$190.
- El **CapEx** se contabiliza una sola vez, en el balance de la entidad que realmente posee el activo. Calcule que se arrienda conjuntamente o que es alquilado por un proveedor del modelo de fundación desde
Un proveedor de servicios en la nube a gran escala reside con el propietario-operador, no también con el arrendatario.
- Para **tokens**: cuando la inferencia es atendida por una fundición que ejecuta el modelo base de otra empresa, esos tokens se atribuyen una sola vez, al modelo que realmente se ejecuta, por lo que
Los modelos ofrecidos a través de fundiciones no se contabilizan dos veces.

Fuentes

Todas las cifras se elaboran **de abajo hacia arriba a partir de fuentes primarias y especializadas**, y se triangulan con estimaciones y aproximaciones de alto nivel. **Los ingresos y el gasto de capital** se basan en los informes de la empresa (SEC 10-K, 10-Q y 8-K) y las divulgaciones de los ejecutivos, verificadas mediante la atribución en la nube (por ejemplo, Azure y Bedrock) cuando los ingresos de una empresa privada aparecen en las cuentas de una empresa pública. Nuestros sistemas escanean y rastrean diariamente las fuentes disponibles para crear, mantener y mejorar la amplitud y profundidad de nuestros datos, con **atribución de fuentes y clasificación de confianza**. Este conjunto de datos de alta calidad se utiliza para construir **modelos financieros completos** (incluido el estado de resultados) para las grandes empresas y modelos basados en impulsores para las empresas más pequeñas.

La participación de CapEx en la IA se desglosa por empresa y se concilia con los ingresos de los proveedores de chips (silicio), el coste de fabricación, la composición del segmento y los análisis de los analistas.

Los volúmenes de tokens se concilian a partir de declaraciones ejecutivas, muestreo y análisis de terceros, volúmenes de tráfico y se contrastan con los ingresos y la capacidad de procesamiento disponible.



Autores



Azeem Azhar
Fundador



Guillermo Gildea
Gerente de producto



Hannah Petrovic, Doctora en Filosofía
Investigador sénior



Nathan Warren
Investigador sénior



Marija Gavrilov
Director general

Agradecemos sus comentarios y contribuciones en
aieconomy@exponentialview.co

Para solicitudes de asesoramiento y consultas institucionales,
póngase en contacto con helen@exponentialview.co



Sigue nuestro análisis en

Exponential View (www.exponentialview.co)

Suscríbase a Exponential View para recibir nuestras investigaciones en su correo electrónico cada semana.

Exponential View

🧠 Why AI isn't showing up on your bottom line

A framework to understand your firm's AI transformation

AZEEM AZHAR AND NATHAN WARREN
MAY 27, 2026 · PAID

♥️ 275 💬 22 ↻ 49

Share ...

I had tea with a senior exec at a well-known public tech company last month. She has about a thousand engineers working for her, and nearly every one of them works with Claude Code. They are producing more lines of code, submitting more pull requests, getting more done. Productivity is up for individuals, but she doesn't see proportional gains at the organization level. As she put it to me: "one plus one plus one plus one equals one-and-a-half."

She is not alone. Uber's COO Andrew Macdonald [went on record](#) this week saying that the relationship between AI investment and results is not there yet:

I think maybe implicitly there is more that is getting shipped, but it's very hard to draw a line between one of those stats and, 'Okay, now we're actually producing 25% more useful consumer features.'

AI has delivered something. I have felt it; my team has felt it; most users have felt it, which is why we keep returning and using more of it. Two years ago, only a [dozen Anthropic customers](#) were spending over \$1 million a year on Claude¹; today, [more than 1,000](#) do. More impressively still, Anthropic's average corporate customer increased their spend by [a factor of five in the past year](#).

But in more than three years since ChatGPT's release, only 27% of executives say AI has met their ROI expectations. What do we make of the other 73%? Could their

