



Oscar Obando
CHAVES

¿Quién paga tu suscripción de IA?

El dato viral dice que te subsidian con miles de dólares.
La aritmética dice otra cosa, porque la mayoría que paga
veinte dólares financia a una minoría intensiva. Quién
gana, quién pierde y qué pasa cuando esto termine.

Oscar Obando Chaves

oscarobando.ai

Agosto de 2026

Documento de distribución libre

El número que casi todos citan mal

En junio de este año la firma de análisis SemiAnalysis compró cada plan de OpenAI y de Anthropic y los exprimió con tareas de programación de larga duración hasta agotar los límites semanales, y el resultado dio la vuelta al mundo en una lectura que conviene revisar despacio porque contiene un error de interpretación que cambia por completo la conclusión.

Las cifras que publicaron son llamativas y están verificadas, porque un plan ChatGPT Plus de veinte dólares consumido al tope equivale a unos **700 dólares** a precios de la interfaz de programación, el plan ChatGPT Pro de doscientos dólares llega a unos **14.000** y el Claude Max de doscientos alcanza unos **8.000**, multiplicadores de setenta y de cuarenta veces respectivamente.

El error de lectura aparece cuando esos números se presentan como el costo que asume la empresa, porque **equivalente a precios de interfaz de programación y costo real de cómputo** son dos cosas distintas, ya que el precio de lista de la interfaz ya incorpora margen, y el propio análisis parte del supuesto explícito de que ambas empresas operan con un setenta y cinco por ciento de margen bruto en ese producto.

Aplicando ese supuesto, los catorce mil dólares equivalentes del plan Pro corresponden a unos **3.500 dólares** de costo real de servir, cifra que sigue siendo un subsidio enorme para quien lo consume al tope y que resulta muy diferente de la afirmación que circuló, según la cual la empresa gasta catorce mil dólares por cada suscriptor de ese plan.

Fuentes. SemiAnalysis, publicación del 10 de junio de 2026 en X y modelo Tokenomics. Cobertura de TechJack Solutions y Gate News. El detalle numérico completo del modelo está tras muro de pago, de modo que los decimales dependen de la fidelidad de la cobertura secundaria.

El margen depende del promedio, no del usuario

La observación más útil de todo el análisis aparece en una frase que la cobertura pasó por alto, porque el margen de un plan de suscripción es función de la utilización promedio de su base de usuarios y no una constante que se pueda atribuir a cada suscriptor por separado.

La analogía que usó el analista **Ed Zitron** describe la mecánica con precisión, porque un gimnasio funciona gracias a que la mayoría de sus socios paga y no va, y él mismo lo formuló en números al señalar que si dos mil personas pagan veinte dólares mensuales sin apenas costar nada, bastan tres personas gastando catorce mil dólares para consumir cada dólar de esa recaudación.

Los puntos de equilibrio publicados ordenan la conversación, porque un plan ChatGPT Plus deja de ser rentable alrededor del **11,4 por ciento** de utilización, los planes Claude Pro y Max de gama media rondan el **20 por ciento**, y en la gama alta el umbral baja al **10 por ciento** en Claude Max de doscientos y al **5,7 por ciento** en ChatGPT Pro, de modo que quien usa su plan de forma ocasional genera margen positivo y amplio.

La evidencia sobre la distribución real apunta en la misma dirección, porque cuando Anthropic anunció sus límites semanales en julio de 2025 dijo que afectarían a **menos del 5 por ciento** de sus suscriptores según los patrones de uso vigentes, y citó el caso de un usuario que había consumido decenas de miles de dólares en un plan de doscientos, de manera que el subsidio profundo existe y está concentrado en una minoría muy pequeña.

Fuentes. E. Zitron, "AI's Brokenomics", wheresyoured.at (junio de 2026). Anthropic, anuncio en X del 28 de julio de 2025 sobre límites semanales, efectivos el 28 de agosto de 2025. SemiAnalysis (junio de 2026).

Vender tokens deja plata, la empresa igual pierde

Buena parte del debate público se resuelve al separar dos afirmaciones que suenan parecidas y describen realidades distintas, porque una cosa es el margen bruto de servir consultas y otra muy distinta es la rentabilidad de la empresa que las sirve.

En la capa de inferencia el negocio gana dinero, y el cálculo más transparente lo publicó el ingeniero **Sean Goedecke** en junio de este año estimando energía, refrigeración y amortización de las tarjetas gráficas para llegar a un costo cercano a **un dólar por millón de tokens**, frente a precios de lista que arrancan en cuatro dólares y medio para los modelos menores y llegan a varias veces esa cifra en los de frontera.

El piso de mercado más verificable lo aporta **DeepSeek**, que publicó sus propias cifras de costo y beneficio de inferencia y que al distribuir los pesos de su modelo de forma abierta no puede sostener rentas artificiales, porque cualquier competidor podría servir el mismo modelo más barato, de modo que su margen funciona como referencia difícil de inflar.

La pérdida de las empresas viene de otro lugar, y los números reportados lo muestran con claridad, porque OpenAI cerró el primer trimestre de este año con **5.700 millones** de ingresos y **39 por ciento** de margen bruto, subiendo desde el treinta y tres por ciento de 2025, mientras proyecta pérdidas que podrían llegar a los catorce mil millones en 2026 por entrenamiento, compromisos de capacidad de cómputo y nómina, todos ellos costos que viven por debajo de la línea del margen bruto.

Fuentes. S. Goedecke, "AI inference is obviously profitable" (26 de junio de 2026). DeepSeek, publicación de costos de inferencia en GitHub (1 de marzo de 2025). The Information (16 de junio de 2026), sobre documentos a accionistas de OpenAI. Las cifras financieras provienen de filtraciones a prensa, ya que ambas empresas son privadas y no publican estados auditados.

LAS CUATRO AFIRMACIONES

Qué es cierto y qué se repite sin revisar

La frase de que las empresas de inteligencia artificial subsidian a sus usuarios mezcla cuatro cosas distintas, y ordenarlas en una tabla resuelve casi toda la discusión.

Afirmación	Veredicto	Sustento
Pierden dinero en cada consulta	Falso	La inferencia opera con márgenes altos, con un costo cercano a un dólar por millón de tokens.
Pierden con el suscriptor de veinte dólares	Falso	Ese tramo está muy por debajo de su punto de equilibrio de utilización.
Pierden con la capa gratuita	Cierto	Costo estimado sin ingreso directo, hoy parcialmente compensado con publicidad desde febrero.
Pierden con el usuario intensivo	Cierto y profundo	Márgenes fuertemente negativos en la minoría que consume el plan al tope.
Pierden como empresa	Cierto en OpenAI	Por entrenamiento, capacidad comprometida y nómina, no por servir consultas.

La tabla deja a la vista quién financia el sistema, y la respuesta incomoda a las dos posiciones del debate, porque el suscriptor que paga veinte dólares y hace cinco preguntas al día resulta la fuente del margen y no el problema del modelo de negocio.

Fuentes. *SemiAnalysis* (junio de 2026). *S. Goedecke* (junio de 2026). *OpenAI*, publicación oficial sobre publicidad (16 de enero de 2026), con anuncios activos desde el 9 de febrero de 2026 en las capas gratuita y Go.

La tarifa plana se está retirando del mercado

El dato más accionable de todo este análisis está en lo que las empresas ya hicieron durante los últimos meses, porque la migración desde la tarifa plana hacia el cobro por uso medido dejó de ser una hipótesis y se convirtió en la dirección declarada de la industria.

GitHub Copilot completó el cambio el 1 de junio de este año, cuando todos sus planes pasaron a consumir créditos por token, y algunos usuarios reportaron saltos desde veintinueve dólares hasta cifras de cientos de dólares mensuales una vez que su consumo real quedó expuesto sin el amortiguador de la tarifa plana.

Anthropic recorrió el mismo camino con Claude Fable 5, que pasó a créditos de uso medido el 20 de julio tras dos postergaciones desde la fecha original del 7 de julio, con Max y Team Premium conservando el modelo incluido a la mitad de los límites semanales, y antes había estrenado los límites semanales en agosto de 2025.

El movimiento en sentido contrario también existe y conviene registrarlo, porque **OpenAI** recortó el 30 de julio los precios de sus modelos de menor costo hasta un ochenta por ciento, una señal de que el costo de la inteligencia sigue cayendo con fuerza y de que la comoditización juega a favor de quien compra, de modo que el mercado se está partiendo entre un tramo utilitario cada vez más barato y un tramo de frontera cada vez más medido.

Fuentes. GitHub, anuncio del 27 de abril de 2026 sobre créditos por token efectivos el 1 de junio. Anthropic, cambios de facturación de Claude Fable 5 (julio de 2026). OpenAI, "Advancing the price-performance frontier with GPT-5.6" (30 de julio de 2026).

Lo que esta economía significa para la región

Para América Latina el asunto tiene dos caras que conviene mirar por separado, porque el precio de acceso y la previsibilidad de los costos plantean problemas distintos, y solo el segundo compromete la viabilidad de un negocio construido sobre estas plataformas.

En materia de precio la región recibió una respuesta concreta, porque OpenAI extendió su capa de bajo costo a Sudamérica alrededor de los seis dólares mensuales y la promovió en nueve países de la región mediante una alianza comercial, mientras el plan Plus se mantiene en veinte dólares sin ajuste local, una cifra que en Ecuador representa cerca del **4 por ciento** del salario básico unificado de **482 dólares** vigente desde enero de este año, y que en el plan de doscientos dólares se dispara a más del cuarenta por ciento de ese salario.

El problema serio aparece del lado de la previsibilidad, porque cualquier empresa regional que haya construido su operación sobre planes de tarifa plana está apoyada en un subsidio cruzado que las propias plataformas están desmontando, y los saltos de costo que ya reportaron usuarios de Copilot muestran el orden de magnitud de la corrección cuando el consumo real queda expuesto.

La conclusión práctica cabe en tres decisiones que cualquier organización puede tomar esta semana, porque conviene modelar los costos a precios de la interfaz de programación en lugar de a precios de suscripción, medir la utilización real de los equipos antes de que la mida el proveedor, y probar modelos de pesos abiertos como cobertura, ya que la caída sostenida del costo de inferencia y la guerra de precios trabajan a favor de quien compra con información.

Fuentes. Ministerio del Trabajo del Ecuador, Acuerdo Ministerial MDT-2025-195, salario básico unificado 2026 de 482 dólares. Comunicaciones de OpenAI sobre disponibilidad regional de su plan Go. Reportes públicos de usuarios de GitHub Copilot tras el cambio de facturación.



“If you've got 2000 people who pay \$20 a month but barely cost anything it only takes three people spending \$14,000 to eat every single dollar of that revenue.”

— Ed Zitron, “AI's Brokenomics”, junio de 2026

Oscar Obando Chaves

oscarobando.ai · [linkedin.com/in/oscar-obando-chaves-237495133](https://www.linkedin.com/in/oscar-obando-chaves-237495133)

Agosto de 2026 · Quito, Ecuador

Este documento puede compartirse libremente con atribución.