

Estrategia & Economía de la IA

IA: el precio está claro. ¿Dónde está el valor?

Las empresas empiezan a descubrir una paradoja: cada unidad de inteligencia artificial cuesta menos, pero la factura total crece. El problema ya no es cuánto cuesta un token. Es saber qué resultado económico produce — y, cada vez más, cómo eso está cambiando la forma en que compramos software.

Hay una vieja distinción atribuida a Warren Buffett que resulta especialmente útil para entender el momento actual de la inteligencia artificial:

Precio es lo que pagamos. Valor es lo que obtenemos.

Durante los primeros años de la IA generativa, a las empresas les preocupó relativamente poco esa diferencia. Había que experimentar. Comprar licencias. Abrir APIs. Probar copilots. Construir agentes. Dar acceso a los empleados. Ver qué ocurría.

La factura era todavía pequeña y el aprendizaje justificaba cierta ineficiencia.

Esa etapa empieza a terminar.

La IA está entrando de verdad en los procesos empresariales y, con ella, aparecen nuevas líneas de gasto: licencias, inferencia, tokens, infraestructura, agentes que ejecutan agentes, búsquedas, razonamiento, generación multimodal y supervisión humana.

El CEO empieza entonces a hacer una pregunta bastante elemental:

¿Qué estoy obteniendo a cambio?

Y muchas organizaciones todavía no tienen una buena respuesta. Accenture encontró que apenas el 23% de los ejecutivos C-level declara estar obteniendo valor empresarial amplio y sostenido de sus inversiones en IA. No parece un problema tecnológico. Parece un problema económico.

El precio se ve. El valor, no tanto.

Hay una razón bastante sencilla.

El costo de la IA aparece perfectamente identificado. El CFO recibe una factura. El CIO sabe qué proveedor la generó. Tecnología puede conocer cuántos tokens se consumieron, cuántas llamadas se hicieron o qué modelos se utilizaron.

Pero el valor aparece en otro lugar.

Puede aparecer en ventas porque un comercial prepara una propuesta en treinta minutos en vez de tres horas. En atención al cliente porque aumenta la resolución en el primer contacto. En finanzas porque un cierre requiere menos trabajo manual. En desarrollo porque una funcionalidad llega antes a producción. O en riesgos porque se detecta una anomalía que antes hubiera pasado inadvertida.

El problema es que costo y valor viven en dos contabilidades diferentes. El CFO ve costo, el CIO ve utilización y las áreas de negocio ven resultados, pero esas tres perspectivas rara vez están conectadas. Accenture hizo una encuesta en donde lo cuantifica: el 57% de las empresas cita una débil alineación entre negocio y tecnología como la causa; el 51% señala capacidades insuficientes de datos y analítica; el 50% no logra aislar la contribución real de la IA de otros factores.

Ese es probablemente uno de los problemas más importantes de la adopción empresarial de IA en este momento. No saber cuánto cuesta la IA sería preocupante. Pero saber perfectamente cuánto cuesta y no saber cuánto vale es bastante peor.

El token se está convirtiendo en un factor de producción

El token tiene una característica económicamente interesante. No es solamente una unidad técnica. Empieza a convertirse en un factor de producción.

Cada prompt consume tokens. Cada respuesta también. Cada búsqueda adicional, cada documento incorporado como contexto y cada interacción entre agentes incrementa el consumo.

Cuando una organización pasa de algunos miles de empleados utilizando un chatbot a cientos de workflows operados permanentemente por agentes, la economía cambia por completo. Accenture declara que una de sus plataformas internas procesa aproximadamente 8,7 billones de tokens por semana. La escala ayuda a entender lo que viene.

Y acaba de aparecer incluso una **Tokenomics Foundation bajo la Linux Foundation** — lanzada formalmente el 4 de agosto de 2026, con 30 organizaciones fundadoras entre las que están Accenture, JPMorgan Chase, IBM, SAP, ServiceNow y Oracle— dedicada precisamente a desarrollar estándares y prácticas para gestionar esta nueva economía. Vale una precisión: su foco es la disciplina de FinOps del gasto interno en IA, no un estándar de

cómo un proveedor de software le cotiza un proyecto a un cliente. Son dos conversaciones distintas.

Pero acá conviene evitar otra confusión, más de fondo.

Medir tokens no significa medir valor.

Un millón de tokens utilizados para resolver correctamente cien casos complejos puede ser baratísimo. Diez mil tokens utilizados para generar algo que nadie necesitaba pueden ser carísimos. El token mide consumo. No mide resultado.

Y por eso creo que sería un error repetir con los tokens lo que durante décadas hicimos con las horas de trabajo.

Cambiar horas por tokens no cambia el modelo

Durante buena parte de la historia del software utilizamos la hora como una aproximación del valor.

Proyecto de diez personas durante seis meses. Mil horas de consultoría. Bolsa de horas de mantenimiento. Horas multiplicadas por tarifa.

Era imperfecto, pero funcionaba mientras la capacidad de producir software fuera relativamente escasa.

En un ensayo anterior sostuve que la IA está modificando precisamente esa condición: cuando producir código se vuelve abundante, el valor se desplaza desde escribirlo hacia el criterio, la arquitectura, el conocimiento de dominio y la responsabilidad sobre el resultado. La IA abarata el código convincente. No abarata, todavía, el software gobernable: el que se integra, es seguro, se puede mantener y tiene quién responda por él.

Por eso reemplazar ahora:

horas × precio

por:

tokens × precio

sería una transformación bastante superficial. Seguimos cobrando y midiendo un insumo. Sólo cambiamos el insumo.

Y hay algo más que el consumo tampoco resuelve: **la responsabilidad**. Un cliente no contrata solamente capacidad de cómputo o productividad; también espera que alguien

responda por la arquitectura, la seguridad, la calidad y el funcionamiento del resultado. La tarifa puramente por tokens puede transparentar el consumo, pero deja abierta una pregunta mucho más importante: **¿quién asume el costo cuando el resultado falla?** La verdadera transformación aparece cuando empezamos a pensar:

workflow → resultado → valor económico

y tratamos las horas humanas, los agentes y los tokens simplemente como los factores de producción necesarios para conseguirlo.

La unidad económica debería ser el workflow

Supongamos que un proceso de análisis financiero requería antes tres personas durante quince horas y ahora requiere una persona durante dos horas más unos cientos de dólares de IA. ¿El valor son los tokens que conseguimos ahorrar?

Evidentemente no.

El valor son las horas liberadas, siempre que esas horas puedan convertirse en capacidad productiva, mejores decisiones, menor riesgo o mayor ingreso. Accenture usa exactamente un ejemplo de este tipo: un deck de cierre de trimestre que antes le tomaba a tres analistas de finanzas 15 horas, ahora le toma a una persona 2 horas, a un costo de token de un par de cientos de dólares frente a miles de dólares de tiempo liberado. El retorno no son los tokens ahorrados. Es la capacidad liberada y redirigida hacia trabajo de mayor valor.

Por eso creo que la unidad económica adecuada no debería ser el usuario y tampoco el token. Debería ser el workflow. Para cada proceso relevante una empresa debería poder responder:

¿Cuánto costaba antes? ¿Cuánto cuesta ahora? ¿Qué resultado producía antes? ¿Qué resultado produce ahora? ¿Qué riesgo introdujo o eliminó la IA? ¿Y cuál es, finalmente, el valor económico de la diferencia?

Recién entonces podemos hablar de ROI.

Hay otra sorpresa: los tokens serán más baratos y probablemente gastaremos más

Intuitivamente parecería que la fuerte caída del precio de los modelos debería disminuir la factura empresarial de IA.

Probablemente ocurra lo contrario.

Es una versión moderna de la paradoja de Jevons. Cuando una tecnología se vuelve mucho más eficiente y barata, aparecen usos que antes económicamente no tenían sentido. Como consecuencia, el consumo total puede aumentar en vez de disminuir.

Accenture ilustra esto con un ejercicio de simulación: modeló cómo reaccionarían las organizaciones si el precio de los tokens cayera un 25% hoy, y encontró que solo el 15% conservaría el ahorro; el resto lo reinvertiría en más casos de uso, mayores workloads o modelos más potentes. (Vale la aclaración metodológica: no es una encuesta a empresas reales, sino un modelado conductual con 6.000 perfiles ejecutivos sintéticos generados por IA a partir de datos demográficos y psicográficos reales — una técnica de simulación válida para explorar tendencias, pero que conviene no presentar con el mismo peso evidencial que una medición directa.)

Lo que sí viene de una fuente independiente es una proyección de consumo agregado: Goldman Sachs proyecta que el consumo de tokens podría multiplicarse aproximadamente 24 veces hacia 2030, alcanzando unos 120 cuatrillones de tokens mensuales.

Esto tiene una consecuencia importante para cualquier CEO:

el objetivo no puede ser gastar menos IA.

De la misma manera que una empresa exitosa no intenta minimizar sus gastos comerciales sino maximizar el retorno de cada euro/dólar comercial, tampoco debería intentar minimizar tokens. El objetivo debería ser:

hacer crecer el valor producido por la IA más rápido que su costo.

No toda tarea necesita la mejor inteligencia disponible

Aparece entonces una disciplina económica que hasta hace poco parecía exclusivamente técnica: **elegir qué modelo debe realizar cada tarea.**

Usar el modelo más poderoso para todo puede parecer prudente. Económicamente es absurdo.

Se estima que apenas entre 10% y 20% de las tareas empresariales justifican modelos frontier o near-frontier. Muchas tareas estructuradas pueden resolverse con modelos mucho más pequeños y baratos.

Esto convierte al model routing en una decisión empresarial. Una operación rutinaria y reversible puede utilizar inteligencia barata. Una decisión jurídica, financiera o regulatoria

de alto impacto probablemente justifique modelos más capaces, más controles y mayor supervisión humana. No se trata de usar siempre la IA más barata. Se trata de utilizar la inteligencia económicamente adecuada para el valor y el riesgo de cada tarea.

Un operador de telecomunicaciones citado por Accenture consiguió reducir su gasto anual de IA un 68% después de analizar consumo a nivel de workflow y rediseñar el routing y el trabajo. Y Gartner, de forma independiente, llega a una advertencia convergente: predice que hacia 2028 el costo de IA en programación superará el salario promedio de un desarrollador a medida que el consumo de tokens se dispare — y advierte explícitamente que, sin un modelo operativo de ingeniería gobernado, “los costos pueden escalar más rápido que las ganancias de productividad”. Dos fuentes independientes, con métodos distintos, llegando a la misma conclusión: sin disciplina de asignación, **el gasto en IA crece más rápido que el valor que produce.**

La optimización, nuevamente, no apareció mirando una factura agregada. Apareció entendiendo para qué se utilizaba cada unidad de inteligencia.

Y también está cambiando cómo se compran los servicios tecnológicos

Este cambio no termina dentro de la empresa. También empieza a modificar la relación entre proveedores y clientes.

La investigación que hice sobre modelos comerciales de servicios tecnológicos muestra algo interesante: hasta agosto de 2026 no aparece evidencia sólida de un mercado que esté simplemente sumando “horas humanas + tokens” como dos líneas de una cotización. Lo que sí empieza a aparecer es algo bastante más disruptivo: el consumo, los outputs o directamente los resultados reemplazan a la hora como unidad comercial.

Globant es el caso más concreto. En agosto de 2026 lanzó Glob.AI con una formulación explícita: sus servicios se cobran por output validado o por consumo real, nunca por horas ni por cantidad de profesionales asignados.

Las grandes consultoras indias muestran la misma dirección. Cognizant reporta que el 45% de sus contratos de BPO ya opera bajo modelos outcome-based, mientras TCS confirma una mayor migración hacia compromisos basados en resultados en sus servicios de negocio agénticos. El cambio es real, aunque todavía convive con time & materials y precio fijo en buena parte del mercado.

La hora empieza a perder relevancia como unidad de cotización. El token puede ocupar transitoriamente parte de su lugar como referencia de consumo. Pero sospecho que tampoco será el destino final.

Porque ni el cliente quiere comprar horas ni, en última instancia, quiere comprar tokens.

Quiere comprar resultados

— y quiere saber quién responde si el resultado falla. La pregunta de quién firma el SLA sigue siendo, en última instancia, humana.

Las cinco preguntas que pondría hoy en una reunión de dirección

Si fuera CEO de una empresa cuya factura de IA empieza a crecer, no comenzaría preguntando cómo reducirla.

Comenzaría preguntando:

- 1 ¿En qué workflows estamos gastando realmente nuestra IA?
- 2 ¿Qué resultado empresarial medible produce cada uno?
- 3 ¿Estamos utilizando más inteligencia —y más cara— de la que ese resultado necesita?
- 4 ¿Qué parte del ahorro de productividad estamos capturando y qué parte simplemente se transforma en más consumo?
- 5 ¿Existe alguien en la organización responsable simultáneamente del costo y del valor producido por la IA?

Si esas respuestas no existen, probablemente todavía no tenemos un problema de costos de IA.

Tenemos un problema anterior: no sabemos administrar económicamente la inteligencia que acabamos de incorporar a la empresa.

Del costo por token al valor por resultado

Durante décadas aprendimos a administrar recursos escasos. Personas. Capital. Servidores. Horas de desarrollo.

La inteligencia artificial introduce algo diferente: una forma de inteligencia cuyo costo marginal cae rápidamente y cuya oferta parece crecer todavía más rápido. Eso cambia la pregunta.

El desafío ya no será simplemente conseguir más inteligencia. Será decidir dónde aplicarla, cuánto pagar por ella, y cómo comprobar que el resultado obtenido justifica el consumo — tanto puertas adentro de la empresa como en la forma en que negocia con sus proveedores de software.

Las organizaciones que se limiten a administrar el costo probablemente terminen intentando consumir menos. Las que entiendan la nueva economía intentarán algo bastante diferente: **obtener cada vez más valor por cada euro/dólar de inteligencia que consumen.**

Porque al final los tokens son solamente otra materia prima. Y ninguna empresa construyó una ventaja competitiva duradera solamente por comprar barata su materia prima. La construyó transformándola mejor que sus competidores en algo por lo que un cliente estuviera dispuesto a pagar.

Precio es lo que pagamos. Valor es lo que obtenemos.

Con la IA, finalmente, estamos empezando a descubrir la diferencia.

Notas

1. Accenture, "AI Tokenomics for Enterprise Value" (AI is on your P&L. Most companies are only reading half of it), Manish Sharma, Lan Guan y Surya Mukherjee, 29 de julio de 2026. Fuente de las cifras de 8,7 billones de tokens/semana, 23% de C-level, simulación 25%/15%, rango 10-20% de tareas frontier, y caso del operador de telecomunicaciones.
 2. Goldman Sachs, proyección de consumo de tokens (~24x hacia 2030, ~120 cuatrillones de tokens/mes), citada por Accenture y por el comunicado de lanzamiento de la Tokenomics Foundation (4 de agosto de 2026).
 3. Gartner, comunicado de prensa del 24 de junio de 2026: "AI Coding Costs Will Surpass Average Developer's Salary by 2028 as Token Consumption Surges".
 4. Business Standard, "As AI changes pricing, IT firms see uptick in outcome-based deals", 20 de agosto de 2026 (cifras de Cognizant, TCS y Coforge).
 5. Investigación de mercado propia de Kairosia sobre modelos comerciales de servicios tecnológicos en la era de la IA (agosto 2026), incluyendo el caso Globant / Glob.AI.
 6. Linux Foundation, comunicados sobre el lanzamiento de la Tokenomics Foundation (3 de junio y 4 de agosto de 2026).
-

**ANIBAL CARMONA**anibalcarmona.com

Barcelona · Agosto 2026

Founder [KAIROSIA](#) · [LinkedIn](#)

Anibal Carmona es empresario tecnológico e ingeniero en sistemas. Inició su trayectoria en 1986 con Hitware y, en 1992, se incorporó a Unitech como co-owner. Durante los 34 años siguientes ejerció como Owner, Chairman y CEO, acompañando el desarrollo de la compañía en Argentina, Chile, Perú, Colombia y España. Su trabajo abarcó plataformas tecnológicas para gobierno, justicia, seguridad y organizaciones empresariales, utilizadas por más de diez millones de personas. Entre 2015 y 2019 fue Chairman de CESSI, la Cámara de la Industria Argentina del Software. Allí contribuyó al trabajo conjunto de empresas, universidades, instituciones y organismos públicos en torno a la agenda estratégica del sector, incluida la Economía del Conocimiento. En 2021 fundó Unitech Iberoamericana en Barcelona y lideró su desarrollo en España y Europa hasta abril de 2026, con foco en inteligencia y seguridad, análisis de información, investigaciones y prevención del fraude. Desde octubre de 2025 dirige Kairosia. Trabaja con CEOs, consejos de administración e instituciones en decisiones vinculadas con inteligencia artificial, negocio, tecnología e inversión. Desde abril de 2026 también continúa vinculado a Unitech como Strategic Advisor. Invierte selectivamente como business angel y acompaña a fundadores en cuestiones de estrategia, crecimiento e internacionalización. Sus escritos abordan la inteligencia artificial, la evolución de la industria del software y las decisiones organizacionales que estas transformaciones exigen.