

¿Cuánto cuesta trabajar con **agentes de IA**?

Gobernar la economía del ecosistema agéntico



Juan G. **CORVALÁN**
Mariana **SÁNCHEZ CAPARRÓS**

Corvalán, Juan G.

¿Cuánto cuesta trabajar con agentes de IA? : gobernar la economía del ecosistema agéntico / Juan G. Corvalán ; Mariana Sánchez Caparrós. - 1a ed. - Ciudad Autónoma de Buenos Aires : La Ley, 2026.

Libro digital, PDF

Archivo Digital: descarga y online

ISBN 978-987-03-5249-5

1. Inteligencia Artificial. I. Sánchez Caparrós, Mariana II. Título
CDD 006.3

Autoría



JUAN GUSTAVO CORVALÁN

Doctor en Ciencias Jurídicas y Máster en IA. Speaker en ONU, OEA, MIT y Google Talks. Director del Laboratorio de IA de la UBA (Derecho). Director del Tratado de 4 tomos de IA y Derecho. Autor y coautor de decenas de investigaciones y publicaciones en IA. Co-creador de Prometea, el primer sistema predictivo en la Justicia de Hispanoamérica (año 2017).



MARIANA SÁNCHEZ CAPARRÓS

Doctora en Ciencias Jurídicas y Magíster en Derecho Administrativo. Subdirectora de UBA IALAB. Integrante de la Red de Expertos de UNESCO en Inteligencia Artificial y Estado de Derecho. Cuenta con formación en diseño UX/UI para productos de Inteligencia Artificial de Stanford Online. Lidera proyectos de I+D en inteligencia artificial, innovación y transformación digital.

Índice

I	Introducción: por qué es necesario repensar los costos de utilizar IA generativa	01
II	Tokens y ventanas de contexto: las bases del costo de inferencia	08
	1. El token	
	2. La ventana de contexto	
III	Agentes de IA: el nuevo multiplicador de costos	13
	3. ¿Qué es un agente de IA?	
	4. El agente como multiplicador del consumo	
	5. La misma tarea, dos facturas	
	6. Del precio del modelo al costo de la ejecución	
	7. Plataformas, API y costo por tarea	
	8. La arquitectura define el recorrido económico	
IV	Del precio publicado al costo efectivo	23
	9. Nadie paga el precio de lista	
	10. Dos listas de precios	
	11. El tamaño del contexto	
	12. El caché	
	13. Los lotes	
	14. La residencia de datos	
	15. Las herramientas	
	16. Recapitulemos: el mismo informe, siete precios	
	17. Del costo por token al costo por tarea	
	18. Estrategias de ahorro	
	19. Controlar el consumo	
	20. Lo que pasó en 2026	

V	La primera decisión de gobernanza económica: elegir la arquitectura adecuada	35
	21. Flujo determinista: el modelo no interviene	
	22. Flujo híbrido: el modelo ayuda en un nodo	
	23. Agente: el modelo conduce	
	24. La regla del diseño escalonado	
	25. Un segundo ejemplo: la bandeja de correo	
	26. Recapitulemos: la misma tarea, tres arquitecturas	
VI	Estrategias de optimización y despliegue	43
	27. Dónde corre el modelo	
	28. Qué modelo en cada paso	
	29. Cuánto contexto carga	
	30. Cuánto razona antes de responder	
	31. El caso Coinbase: la respuesta contraria a Uber	
	32. Recapitulemos: cuatro perillas y un límite	
VII	El costo que la factura no muestra	52
	33. El ciclo completo y la parte que no se acelera	
	34. Tres formas de delegar, tres precios	
	35. Las cuatro paradojas, en dinero	
	36. El cuello de botella y el colapso	
	37. Cómo se presupuesta	
VIII	Conclusiones	60
	Lo que el ejemplo mostró	
	Lo que eso significa	
	Lo que pasó cuando no se hizo	
	Lo que queda	
IX	Implicancias para la alta dirección y las áreas de tecnología	63
	Seis preguntas antes de desplegar	
	Para las áreas de tecnología	
	Para la alta dirección	

CAPÍTULO I

Introducción: por qué es necesario repensar los costos de utilizar IA generativa

Una persona abre ChatGPT, Claude o Gemini, escribe una instrucción y recibe una respuesta en pocos segundos. Desde afuera, la operación económica parece bastante simple: un prompt, una respuesta y un costo asociado al intercambio.

Ahora imaginemos otra escena. La persona ya no formula una pregunta concreta. Le pide al sistema que investigue un tema, consulte fuentes, compare documentos, realice cálculos, redacte un informe y revise el resultado antes de entregarlo. La respuesta final podría ocupar las mismas dos páginas que en el primer caso. Sin embargo, detrás de ellas el sistema pudo haber realizado decenas de llamadas al modelo, utilizado herramientas, reprocesado contexto y corregido errores sin que la persona interviniera.

La respuesta visible puede parecer la misma. El proceso que la produjo, no.

Acá comienza el desafío que organiza este trabajo: comprender cuánto cuesta realmente esa respuesta.

El cambio puede comprenderse como el tránsito desde una IA generativa predominantemente reactiva hacia un ecosistema agéntico crecientemente proactivo. En la lógica reactiva, la IA espera una instrucción, procesa la solicitud y devuelve una respuesta. En la lógica agéntica, recibe un objetivo y puede desplegar una secuencia de acciones para alcanzarlo. Puede planificar, seleccionar herramientas, consultar fuentes externas, observar resultados, modificar el curso de ejecución y continuar actuando hasta alcanzar una condición de finalización.

La diferencia no reside solamente en que hace más cosas. Cambia la lógica de interacción: de una IA que principalmente responde a instrucciones pasamos a sistemas capaces de sostener ciclos de ejecución con mayor autonomía. Y cuando cambia la lógica de ejecución, cambia también la lógica del costo.

En 2024 publicamos el trabajo ¿Cuál es el costo de trabajar con inteligencia artificial generativa? Allí analizamos algunas de las principales variables económicas que debían considerarse para incorporar esta tecnología en proyectos y organizaciones: el precio de los tokens, las ventanas de contexto, las diferencias entre proveedores y ciertos criterios básicos para evaluar la viabilidad económica de su utilización¹.

Dos años después, ese análisis necesita ser revisado y ampliado. No solamente porque cambiaron los modelos, aumentaron sus capacidades o se modificaron las estructuras tarifarias. El cambio más importante es otro: cambió la arquitectura mediante la cual utilizamos IA generativa y, con ella, la forma en que se produce el costo.

En una interacción reactiva, el consumo puede asociarse con relativa facilidad a los tokens de entrada, los tokens de salida y el precio del modelo utilizado. En una ejecución agéntica, se distribuye entre múltiples llamadas, herramientas, iteraciones y contextos que pueden volver a procesarse. Esta dinámica se intensifica cuando varios agentes distribuyen tareas, intercambian resultados o revisan el trabajo de otros.

La paradoja es que una respuesta muy breve puede ocultar una ejecución muy extensa. La mayor autonomía no implica solamente nuevas capacidades: también puede introducir un consumo más distribuido, acumulativo y variable.

Este ecosistema no ingresa a las organizaciones de una sola manera. Puede hacerlo mediante plataformas de uso cotidiano que incorporan capacidades agénticas por diseño, como ChatGPT, Gemini, Claude, Grok o Kimi; mediante agentes empaquetados o asistentes de trabajo, como Claude Cwork, ChatGPT Work, Codex y herramientas equivalentes; o a través de agentes diseñados específicamente para determinados procesos e integrados con documentos, bases de datos, APIs, sistemas internos y herramientas externas.

Estas modalidades pueden coexistir dentro de una misma institución. La fragmentación tecnológica produce, así, una fragmentación económica: el gasto puede quedar distribuido entre usuarios, plataformas, modelos, herramientas, agentes y proyectos sin una visión consolidada de cuánto se consume, dónde se consume y qué valor produce ese consumo.

A esta transformación técnica se suma un cambio en los modelos comerciales de acceso. Los planes de suscripción mensual, que durante la primera etapa de expansión de la IA generativa ofrecían modalidades amplias o aparentemente ilimitadas, comenzaron a convivir con límites expresados en tokens, mensajes, créditos, sesiones u otras unidades de consumo.

Durante 2026 esta tendencia se hizo especialmente visible en herramientas de uso intensivo y servicios empresariales, donde distintos proveedores avanzaron hacia modalidades que vinculan más directamente el precio con el consumo efectivo².

Para las organizaciones, la consecuencia es directa. Antes alcanzaba, muchas veces, con preguntar cuánto costaba acceder a la herramienta. Ahora también hay que determinar cuánto cuesta utilizarla. La licencia muestra solo la puerta de entrada.

En este contexto, anticipar y atribuir el costo efectivo de utilizar IA se vuelve más complejo. La relevancia del problema comenzó incluso a generar prácticas y marcos específicos orientados a la gobernanza económica de los tokens. La FinOps Foundation, por ejemplo, aborda este campo bajo la noción de Tokenomics³. Y en el primer semestre de 2026 dejó de ser un problema teórico: empresas con miles de ingenieros y equipos financieros sofisticados agotaron en pocos meses el presupuesto de IA del año entero.

Ahora bien, gobernar económicamente la IA agéntica no puede reducirse a contar tokens. Saber cuánto se consume es indispensable, pero no responde la pregunta organizacional más importante: ¿qué obtenemos a cambio de ese consumo?

Dos sistemas pueden consumir una cantidad similar de tokens y producir impactos radicalmente distintos. Uno puede intervenir en una actividad marginal; el otro, en un proceso crítico capaz de reducir tiempos, evitar errores, ampliar capacidad operativa o mejorar una decisión.

Veámoslo en cuatro niveles:

01	TOKEN CONSUMO TÉCNICO	permite medir el consumo técnico.
02	EJECUCIÓN RECURSOS DEL RECORRIDO	permite reconstruir los recursos utilizados durante el recorrido del sistema.
03	TAREA COMPLETADA COSTO DEL RESULTADO	permite estimar cuánto costó alcanzar un resultado efectivamente utilizable.
04	VALOR PRODUCIDO JUSTIFICACIÓN	permite evaluar si ese costo se encuentra justificado.

La distinción importa porque una ejecución no siempre equivale a una tarea completada. El sistema puede fallar, requerir un nuevo intento, producir un resultado que deba rehacerse o demandar una intervención humana significativa antes de que el producto resulte útil. Y esa intervención humana, la que verifica, corrige e integra lo que el sistema produjo, tiene un costo que ninguna factura de proveedor registra.

Una primera conclusión: un token más barato no garantiza una tarea más barata.

Una organización puede reducir el precio unitario de los tokens y, al mismo tiempo, aumentar el costo total de sus procesos si despliega arquitecturas que realizan más llamadas, acumulan más contexto o ejecutan más iteraciones de las necesarias. Del mismo modo, un modelo más caro puede resultar económicamente conveniente si completa correctamente una tarea con menos pasos y menor intervención humana.

Y hay una segunda conclusión: una tarea más barata tampoco implica una mejor decisión si produce menos valor o aumenta el riesgo.

Esto nos lleva a una de las ideas centrales del trabajo: la economía de la IA agéntica es, en buena medida, una economía de la arquitectura. Elegir entre una automatización determinista, un flujo híbrido o un agente no determina solamente las capacidades del sistema. También condiciona cuántas veces se utilizará el modelo, cuánto contexto procesará, qué herramientas invocará, cuánto costará supervisarlos y qué mecanismos de control serán necesarios.

Por eso, la decisión económica fundamental no comienza con la comparación de modelos ni con la lectura de las tarifas publicadas por los proveedores. Comienza antes: **¿qué arquitectura necesitamos realmente para resolver esta tarea?**

Si una actividad puede resolverse adecuadamente mediante reglas, incorporar IA generativa puede agregar un costo innecesario. Si requiere componentes puntuales de interpretación, clasificación, extracción o generación de lenguaje, un flujo híbrido puede resultar suficiente. Solamente cuando la tarea exige adaptación dinámica, exploración, uso flexible de herramientas o razonamiento iterativo puede justificarse avanzar hacia una arquitectura plenamente agéntica.

La pregunta no debería ser si podemos utilizar un agente, sino cuál es la arquitectura mínima suficiente para alcanzar el resultado esperado con un costo, un riesgo y un nivel de control aceptables.

Desde esta perspectiva, gobernar económicamente el ecosistema exige combinar tres miradas.



01

MIRADA CONTABLE

Contable

Cuánto se consume, dónde se consume y qué recursos explican el gasto.



02

MIRADA DE PRODUCTIVIDAD

Productividad y valor

Qué tarea se completa, con qué finalidad, bajo qué nivel de autonomía, con qué calidad, con qué riesgo y con qué impacto en el resultado.



03

MIRADA COGNITIVA

Cognitiva

Cuánto juicio humano exige cada resultado y qué hace la organización para que ese juicio no se atrofie mientras delega.

Una factura puede decir cuánto gastamos. No puede decir, por sí sola, si el gasto valió la pena.

Por todo ello, este trabajo retoma nuestro análisis de 2024 y lo traslada al ecosistema agéntico. Lo hace con un solo ejemplo, un contrato de cincuenta páginas y un informe de una página, que recorre todos los capítulos y va recibiendo precio en cada uno: lo que cuesta una llamada y lo que cuesta un agente; lo que cambia entre el precio publicado y la factura; lo que decide la arquitectura; lo que ahorran las decisiones de despliegue; y lo que cuesta la parte del trabajo que ninguna factura muestra.

Desde lo metodológico, combinamos documentación técnica y tarifaria de proveedores, literatura académica y especializada, métricas publicadas por terceros y casos de implementación reportados públicamente durante 2026. Las tarifas y condiciones comerciales fueron verificadas al 5 de septiembre de 2026 y se citan con esa fecha; cambian con frecuencia y deben leerse como órdenes de magnitud.

En definitiva, gobernar el costo de la IA agéntica no significa simplemente consumir menos. Significa hacer visible el recorrido que produce cada resultado, comprender qué estamos pagando y decidir qué nivel de autonomía vale la pena financiar.

La pregunta final ya no es cuánto cuesta el token de un modelo. Es cuánto cuesta alcanzar un resultado, bajo qué arquitectura, con cuánto juicio humano y qué valor obtenemos a cambio.

CAPÍTULO II

Tokens y ventanas de contexto: las bases del costo de inferencia

Imaginemos que una persona necesita revisar un contrato, analizar la documentación relacionada y preparar un informe de una página.

Puede delegar esa tarea en Cowork de Claude, en ChatGPT Work, en el modo multiagente de Grok o, si se trata de una actividad técnica, en un entorno como Google Antigravity⁴.

Desde la perspectiva del usuario, la tarea parece simple: adjuntar los archivos, describir el resultado esperado y esperar el informe.

Desde la perspectiva del sistema, en cambio, esa única instrucción puede poner en marcha varias operaciones. El o los agentes deben interpretar el pedido, leer el contrato, identificar cláusulas relevantes, contrastarlas con la documentación, organizar los hallazgos, redactar una primera versión y revisarla antes de entregar el resultado.

El usuario recibe una página. La plataforma, para producirla, puede haber leído varias veces el contrato y la documentación. Y cada nueva lectura forma parte del costo.

Para comprender cómo se forma ese costo hay dos conceptos técnicos fundamentales. Por un lado, los tokens: las unidades con las que el modelo procesa la información. Por otra parte, la ventana de contexto: cuánta información puede mantener disponible en cada llamada.

En una conversación simple, estas variables permiten estimar el precio de un intercambio. En una ejecución agéntica, permiten reconstruir el costo de un recorrido completo.

1. El token

Los modelos de lenguaje no leen un texto del mismo modo que una persona. Antes de procesarlo, lo dividen en unidades más pequeñas denominadas tokens.

Un token puede coincidir con una palabra completa, con una parte de una palabra, con un signo de puntuación o, en algunos casos, con un solo carácter. La división depende del tokenizador utilizado por cada modelo.

Por ejemplo, la frase "Cuánto cuesta trabajar con agentes de IA" tiene 40 caracteres, pero el tokenizador de OpenAI utilizado en el ejemplo la divide en ocho tokens⁵.

Esta fragmentación es importante porque los proveedores no facturan por palabras ni por páginas, sino por la cantidad de tokens que el modelo recibe y genera.

Los tokens de entrada incluyen las instrucciones, los documentos, el historial de la conversación, las respuestas de herramientas y cualquier otra información enviada al modelo⁶.

Los tokens de salida corresponden al contenido que el modelo produce: una respuesta, un resumen, un plan, código, una instrucción para utilizar una herramienta o, en los modelos actuales, el razonamiento intermedio que genera antes de responder.

Ambas categorías suelen tener precios diferentes y, en casi todos los modelos, generar tokens de salida resulta varias veces más costoso que procesar tokens de entrada. El costo se calcula en cada llamada al modelo.

Esa última precisión cambia la lectura económica del ejemplo inicial. Si el contrato y la documentación se envían una vez, se contabilizan una vez. Si el agente los vuelve a incorporar durante la planificación, el análisis y la revisión, esos tokens pueden contabilizarse nuevamente en cada llamada.

Como aproximación práctica para textos en inglés, suele estimarse que un token equivale a unos cuatro caracteres y que 1.000 tokens representan alrededor de 750 palabras. Es una regla útil para una primera estimación, no una medida exacta: la relación cambia según el idioma, el vocabulario, los números, los símbolos y el modelo utilizado⁷.

El idioma también forma parte de la arquitectura económica.

La diferencia entre idiomas no es neutral⁸. Una misma idea puede requerir más tokens en una lengua que en otra. En pleno furor de la IA generativa (2023), con los primeros modelos de lenguaje, algunos autores mostraron que los sistemas comerciales de tokenización pueden generar costos diferentes para contenidos equivalentes según el idioma. La literatura ha denominado a este efecto "tokenizer tax"⁹: no se trata de un impuesto fiscal, sino del mayor costo que pueden enfrentar los hablantes de lenguas no anglófonas como consecuencia de las decisiones de diseño del tokenizador.

Una medición posterior sobre 25 lenguas europeas encontró para el español una fertilidad promedio de 1,46 tokens por palabra, frente a 1,23 en inglés. Según el modelo considerado, el mismo contenido en español podía requerir entre un 6 % y un 31 % más tokens¹⁰.

En síntesis, una estimación organizacional no debería apoyarse únicamente en la cantidad de palabras o páginas. Cuando el volumen es relevante, conviene medir el contenido con el tokenizador del modelo que efectivamente se utilizará y observar el consumo real de las ejecuciones.

2. La ventana de contexto

Si el token es la unidad de consumo, la ventana de contexto define cuántas de esas unidades pueden estar disponibles para el modelo durante una llamada.

Puede imaginarse como la mesa de trabajo del sistema. Sobre esa mesa se colocan la instrucción, el contrato, la documentación relacionada, el historial de pasos anteriores y los resultados de las herramientas. También, en algunos casos, una parte de la respuesta que el modelo está generando.

Una mesa más grande permite desplegar más documentos. No obliga a ponerlos todos.

Una ventana amplia hace posible procesar expedientes extensos, repositorios de código o conversaciones prolongadas. Algunos modelos de frontera se acercan o llegan al millón de tokens de contexto, una capacidad que puede equivaler, de manera aproximada, a cientos de páginas de texto¹¹.

Pero capacidad no significa gratuidad. Toda la información que se incorpora al contexto activo debe procesarse. Una ventana más grande habilita tareas más ambiciosas, aunque también puede aumentar el costo cuando se la llena con información innecesaria o cuando el mismo contenido vuelve a enviarse en sucesivas llamadas.

Además, una ventana amplia no garantiza que toda la información sea utilizada con la misma eficacia. Cuanto más extenso es el contexto, más importante resulta decidir qué contenido debe permanecer disponible y en qué momento.

Cuando el material supera la capacidad del modelo, o cuando mantenerlo completo sería demasiado costoso, el sistema puede detener la ejecución, resumirlo, recuperar solamente los fragmentos relevantes, reutilizar información almacenada en caché o comprimir el historial de la ejecución. Estas técnicas no son detalles menores. Determinan qué información se procesa, cuántas veces se procesa y cuánto cuesta conservarla disponible.

El precio también puede cambiar con la longitud. Algunos proveedores aplican tarifas diferenciadas cuando una solicitud supera determinados umbrales de contexto, de modo que una ejecución extensa no siempre se calcula con el mismo precio unitario que una breve¹². El capítulo IV muestra dónde está ese umbral y cuánto cuesta cruzarlo.

En una interacción reactiva, el contexto suele crecer de manera relativamente visible: cada nuevo mensaje agrega información a la conversación. En un agente, parte de ese crecimiento ocurre fuera de la vista del usuario. El sistema puede incorporar planes, observaciones, resultados de herramientas y versiones intermedias que luego vuelven a ingresar al modelo.

Volvamos al ejemplo. En una conversación simple, el contrato podría procesarse una vez para producir una respuesta. En una ejecución agéntica, puede volver a ingresar en cada etapa de análisis, contraste, redacción y corrección.

Por eso, tokens y ventana de contexto son la base del costo de inferencia. La inferencia es el proceso mediante el cual un modelo de lenguaje entrenado utiliza los patrones aprendidos para generar una respuesta o predicción a partir de una entrada. El entrenamiento es la etapa inicial en la que el sistema procesa volúmenes masivos de datos para ajustar sus parámetros internos. La inferencia ocurre cada vez que el usuario interactúa con la herramienta: cada vez que el sistema lee un documento, procesa una instrucción o redacta un informe, está ejecutando una inferencia.

Ahora bien, todo esto no explica por sí solo el costo final de una tarea agéntica. Para hacerlo es necesario observar cuántas llamadas realiza el sistema, qué información repite, qué herramientas utiliza y cuántas veces decide volver a intentarlo. Ese es el desafío del capítulo siguiente.

CAPÍTULO III

Agentes de IA: el nuevo multiplicador de costos

Volvamos al contrato y a la documentación del capítulo anterior.

Una persona puede pedirle a un modelo que lea los archivos y prepare un informe. El sistema recibe la instrucción, procesa el material y devuelve una respuesta.

También puede delegar el mismo objetivo en Cowork de Claude, en ChatGPT Work, en un entorno agéntico de Grok o, para determinadas tareas técnicas, en Google Antigravity.

Desde la perspectiva del usuario, el resultado esperado no cambia: revisar el contrato, contrastarlo con la documentación y producir un informe de una página. Desde la perspectiva del sistema, el recorrido puede ser completamente diferente.

Un agente puede interpretar el objetivo, organizar un plan, abrir los documentos, localizar cláusulas relevantes, buscar antecedentes, contrastar la información, redactar una primera versión, detectar que falta un dato, volver a consultar los archivos y corregir el informe antes de entregarlo.

El resultado puede ser el mismo, pero el recorrido y el costo no.

Los agentes de IA constituyen una de las formas de uso de los modelos de lenguaje con mayor capacidad para incrementar el consumo de tokens. No necesariamente porque utilicen un modelo más caro, sino porque pueden utilizarlo muchas veces dentro de una misma tarea¹³.

3. ¿Qué es un agente de IA?

La diferencia entre un asistente y un agente no depende solamente de la calidad de la respuesta. Depende de la arquitectura que organiza la interacción.

Un asistente recibe una instrucción y devuelve una respuesta. Si la persona quiere ampliar, corregir o continuar el trabajo, debe formular una nueva instrucción.

Un agente, en cambio, recibe un objetivo y puede sostener una secuencia de acciones para alcanzarlo. Puede planificar, seleccionar herramientas, consultar fuentes, procesar documentos, observar resultados, modificar su estrategia y continuar actuando hasta completar la tarea¹⁴.

La diferencia puede expresarse de una manera sencilla:



Esto no significa que el agente actúe sin límites. Su autonomía es funcional. Opera dentro de un entorno previamente diseñado: dispone de ciertas herramientas, tiene permisos determinados, recibe instrucciones de sistema y puede estar sujeto a límites de tiempo, presupuesto, cantidad de pasos o intervención humana.

La autonomía no elimina la arquitectura. La hace más importante.

Para comprender el costo conviene distinguir tres unidades.



4. El agente como multiplicador del consumo

El mayor consumo de un agente surge del bucle que sostiene su funcionamiento.

En el ejemplo del contrato, una interacción directa podría procesar una vez los documentos y generar el informe. Un agente puede procesarlos al planificar, volver sobre ellos al contrastar cláusulas, incorporar resultados de una búsqueda y utilizarlos nuevamente durante la revisión final.

Cada paso puede producir nuevos tokens de entrada y de salida. También puede agregar información al contexto que deberá conservarse para las etapas siguientes.

Por eso, una respuesta breve no necesariamente es barata. Puede ser la parte visible de una ejecución extensa.

El relevamiento que hicimos desde IALAB permite ponerle números al recorrido¹⁵. Para analizar un expediente de cincuenta páginas estimamos 25.000 tokens de entrada y 3.000 de salida. Con las tarifas relevadas al 24 de junio de 2026, esa única llamada costaba alrededor de un centavo y medio de dólar con DeepSeek V4-Pro, cuatro centavos con Grok 4.3, seis centavos y medio con Gemini 3.5 Flash y veinte centavos con Claude Opus 4.8 o GPT-5.5.

Ahora reconstruyamos la ejecución del agente. Supongamos que hace cinco llamadas y que en cada paso conserva solamente lo necesario de los anteriores, de modo que el contexto crece de manera gradual:

1. Lectura inicial y planificación: 25.000 tokens de entrada y 1.000 de salida.
2. Identificación de las cláusulas relevantes: 27.000 de entrada y 2.000 de salida.
3. Contraste del contrato con la documentación: 30.000 de entrada y 2.000 de salida.
4. Redacción del informe: 33.000 de entrada y 3.000 de salida.
5. Revisión y corrección final: 36.000 de entrada y 1.000 de salida.

Al terminar, el agente procesó unos 160.000 tokens: 151.000 de entrada y 9.000 de salida. Seis veces más que la llamada única.

En dinero, con Claude Opus 4.8 o GPT-5.5 el informe cuesta alrededor de 1 dólar. Con Gemini 3.5 Flash, 30 centavos. Con Grok 4.3, 20 centavos. Con DeepSeek V4-Pro, 7 centavos.

Un informe agéntico con un modelo de frontera cuesta, en números redondos, 1 dólar. Una única llamada, 20 centavos. Es la misma página, producida cinco veces más cara.

El cálculo es estimativo. No incluye reintentos, búsquedas externas, tiempo de ejecución, infraestructura ni revisión humana. Si el agente utiliza esas capacidades, el costo final será mayor.

Anthropic informó que, en sus mediciones¹⁶, un agente utilizaba aproximadamente cuatro veces más tokens que una interacción de chat, y que un sistema multiagente utilizaba alrededor de quince veces más. Nuestro ejemplo arroja un aumento cercano a cinco veces. El resultado se encuentra en un orden de magnitud compatible con esa medición y muestra que la diferencia puede explicarse sin cambiar de modelo.

No se trata de una tarifa universal. Un agente puede necesitar menos pasos y otro muchos más. Gartner¹⁷ estimó en 2026 un rango todavía más amplio: entre cinco y treinta veces más tokens por tarea que un chatbot estándar. La magnitud depende del tipo de tarea, del modelo, de la cantidad de iteraciones y del diseño del sistema. El ejemplo muestra algo distinto: aun utilizando exactamente el mismo modelo, la arquitectura puede cambiar varias veces el costo de alcanzar el resultado.

Un token más barato no garantiza una tarea más barata.

Un proveedor puede reducir el precio por millón de tokens y, al mismo tiempo, una organización puede aumentar su gasto si habilita sistemas que realizan más llamadas, acumulan contexto innecesario o utilizan herramientas sin límites claros.

5. La misma tarea, dos facturas

Un dólar por informe es una cifra abstracta hasta que aparece en una factura. Y aparece en dos facturas distintas según cómo se accede.

Por suscripción, la persona paga un monto fijo y recibe un cupo de uso que se renueva por ventanas de tiempo. Al 5 de septiembre de 2026, tanto Anthropic como OpenAI organizan sus planes individuales en tres escalones: 20 dólares al mes, 100 dólares con cinco veces más uso y 200 dólares con veinte veces más uso. En ambos casos el cupo se mide en ventanas de cinco horas, y en el plan de 20 dólares esa ventana admite, según el proveedor y la extensión de los mensajes, entre diez y cien mensajes. Los propios proveedores advierten que ese número no es fijo: depende del modelo elegido, de los archivos adjuntos y del largo de la conversación.

Una tarea agéntica no vale un mensaje de ese cupo. Vale varios, porque el sistema vuelve a leer, busca y corrige antes de entregar. Anthropic señala que Cowork no tiene un cupo propio: consume del mismo límite que el chat. OpenAI indica que las tareas de ChatGPT Work que se ejecutan en la nube pueden consumir más cupo que un mensaje local. Varios informes seguidos, y la herramienta deja de responder: pide esperar a que se renueve la ventana.

En la suscripción el precio es fijo y el límite es el tiempo.

Los proveedores ya ofrecen una válvula para no esperar. Cuando el cupo se agota, el usuario puede comprar créditos sin cambiar de plan. OpenAI los vende en paquetes de 1.000 créditos por 40 dólares y publica la tarifa en créditos de cada modelo: con GPT-5.6 Sol, 100 créditos por millón de tokens de entrada y 500 por millón de salida. Nuestro informe de 151.000 tokens de entrada y 9.000 de salida cuesta, en esa unidad, unos 20 créditos. El paquete de 40 dólares alcanza para cincuenta informes: 80 centavos cada uno.

Es decir, la misma cifra que muestra la API, calculada desde la pantalla del usuario.

Por API no hay cupo ni espera. Cada informe se cobra al pasar. Cuarenta contratos en el mes son 40 dólares; doscientos contratos son 200 dólares. Si la tarea la resuelve un sistema multiagente que consume quince veces más que un chat, cada informe pasa a costar 3 dólares y los mismos doscientos contratos, 600 dólares.

| En la API el precio es variable y el límite es el presupuesto.

La diferencia no es el precio. El dólar por informe es el mismo. La diferencia es quién pone el techo y en qué unidad se expresa. En la suscripción lo pone el proveedor y se mide en horas. En la API lo pone la organización y se mide en dinero. Si la organización no lo pone, no hay techo.

6. Del precio del modelo al costo de la ejecución

Las plataformas ya muestran, en su propia forma de facturar, que el costo agéntico tiene más de una dimensión.

Anthropic publica un ejemplo de una sesión de una hora con Claude Managed Agents y Claude Opus 5, un modelo posterior al que utilizamos en nuestro relevamiento. La ejecución consume 50.000 tokens de entrada, 15.000 de salida y una hora de tiempo activo. El costo total informado es de 70 centavos de dólar: 25 centavos por la entrada, 37 centavos por la salida y 8 centavos por el tiempo de ejecución. Con reutilización de contexto mediante caché, el mismo ejemplo baja a 52 centavos¹⁸.

Puesto en escala, 70 centavos por hora de agente activo son unos 6 dólares por jornada de ocho horas y unos 110 dólares por veinte jornadas.

El ejemplo es valioso porque separa tres componentes que suelen aparecer mezclados: inferencia, infraestructura y gestión del contexto.

Las herramientas agregan otra capa. La búsqueda web, la consulta de archivos, la ejecución de código o la conexión con servicios externos pueden tener un precio propio. Además, los resultados que devuelven pasan a formar parte del contexto y pueden aumentar los tokens de las llamadas posteriores¹⁹.

La documentación de xAI hace explícita esta dinámica: en una solicitud agéntica, los tokens de entrada se contabilizan de manera acumulativa a través de las distintas inferencias y cada nueva llamada puede incorporar el historial completo generado hasta ese momento. La plataforma permite limitar la cantidad de turnos precisamente porque una investigación más profunda implica mayor latencia y mayor costo²⁰.

Aquí aparece otra diferencia frente a la interacción reactiva. En un chat, el usuario decide cuándo formular una nueva instrucción. En un agente, buena parte de esas decisiones ocurre dentro del ciclo de ejecución.

La organización no habilita solamente capacidad de respuesta o de ejecución de resultados; también abre las puertas a diversos escenarios de consumo.

7. Plataformas, API y costo por tarea

Para la mayoría de los usuarios, la inteligencia artificial aparece a través de una plataforma. Entramos en ChatGPT, Claude o Gemini, escribimos una instrucción, adjuntamos eventualmente un documento y esperamos una respuesta. Este acceso suele organizarse mediante una suscripción, un cupo o créditos. Conocemos el precio del acceso, pero no siempre cuánto costó una tarea concreta.

Conectarse por API significa dejar de hablar con la inteligencia artificial a través de su pantalla y hacer que otro sistema hable con ella directamente.

En el ejemplo del contrato, la diferencia puede verse así.



La diferencia no es solamente técnica. La IA deja de ser un lugar al que vamos a consultar y se convierte en una capacidad que podemos insertar dentro de un proceso.

Si tenemos que revisar un contrato, el trabajo manual funciona perfectamente. Si tenemos que revisar cien, podemos construir otro recorrido:



La persona aparece recién cuando hace falta revisar una excepción, corregir un error o tomar una decisión.

Ahí se entiende la importancia de las APIs para la IA agéntica. Funcionan como una puerta de acceso a la inteligencia del modelo. Nuestro sistema envía el documento, indica qué debe hacerse y define cómo debe devolverse el resultado. Esto puede repetirse cientos o miles de veces sin abrir manualmente un chat.

Usar una plataforma de IA es utilizar un producto.

Conectarse por API es utilizar la inteligencia del modelo como una pieza para construir otro producto, un proceso o un agente.

Esta diferencia también modifica la lógica económica. En una plataforma solemos mirar el precio mensual. Mediante API podemos reconstruir con mayor precisión el consumo de cada ejecución: cuánto texto enviamos, cuánto procesa y produce el modelo, qué herramientas utiliza y cuántas veces lo llamamos.

La API no es necesariamente más barata. Hace visible otra unidad de análisis. Ya no observamos solamente el precio de acceso: podemos estimar cuánto cuesta completar la tarea.

En una arquitectura agéntica, cada paso puede convertirse en una nueva llamada. Por eso, automatizar cien contratos exige responder algunas preguntas: qué modelo utilizamos; cuántas veces lo llamamos; qué información enviamos en cada paso; cuánto razonamiento y qué herramientas habilitamos; y cuánto cuesta completar el proceso entero.

Ahí pasamos de usar IA a diseñar sistemas con IA agéntica.

8. La arquitectura define el recorrido económico

El ecosistema agéntico no introduce solamente un recargo sobre el modelo de costos anterior: cambia la forma en que los costos se producen.

En una interacción directa, el recorrido es relativamente acotado. En una ejecución agéntica, el consumo puede ser acumulativo, variable y parcialmente invisible para el usuario.

Dos sistemas que utilizan el mismo modelo pueden producir costos radicalmente diferentes. Uno puede resolver el análisis del contrato mediante una única llamada. Otro puede planificar, buscar, releer, redactar, evaluar y corregir antes de entregar exactamente la misma página.

La mayor autonomía puede estar justificada cuando esas acciones mejoran la calidad del resultado, permiten trabajar con información cambiante o resuelven una tarea que no podría estructurarse de antemano.

Pero si el mismo resultado puede obtenerse mediante una llamada directa o un flujo más acotado, la complejidad adicional puede aumentar el costo sin producir un valor proporcional.

Por eso, la pregunta económica no es solamente qué modelo utiliza el agente. También importa cuántas veces puede llamar al modelo; qué información conserva y vuelve a procesar; qué herramientas puede utilizar; cuántos intentos puede realizar; qué condiciones detienen la ejecución; y qué intervención humana requiere para producir un resultado utilizable.

Estas decisiones forman el presupuesto real del agente.

Ahora bien, al efecto multiplicador de la arquitectura hay que agregar cómo ese consumo se convierte efectivamente en dinero. Para hacerlo, no alcanza con observar una tarifa nominal por millón de tokens. Hay que considerar las categorías de entrada y salida, el contexto, el caché, las herramientas, la infraestructura y las condiciones comerciales aplicables a cada ejecución.

De eso se ocupa el capítulo siguiente.²¹

CAPÍTULO IV

Del precio publicado al costo efectivo

Volvamos al informe del capítulo anterior. Con Claude Opus 4.8 o GPT-5.5, el agente lo produjo por alrededor de 1 dólar. Esa cifra salió de una cuenta simple: los tokens de entrada por el precio de entrada, los de salida por el precio de salida, dividido por un millón.

Esa cuenta usa el precio de lista. Y el precio de lista es el que casi nadie paga.

Según cómo se organice la ejecución, el mismo informe puede costar la mitad o el doble sin cambiar de modelo. Este capítulo muestra por qué: cinco variables que suben o bajan la cuenta y que, en una arquitectura agéntica, dejan de ser detalles comerciales para convertirse en decisiones de diseño.

9. Nadie paga el precio de lista

Los proveedores publican sus tarifas como precio por millón de tokens de entrada y por millón de tokens de salida. La unidad es cómoda para comparar. También es engañosa, porque sugiere que todos los tokens valen lo mismo.

No es así. Un token de salida cuesta varias veces más que uno de entrada. Un token que el modelo ya procesó y recupera desde caché cuesta una décima parte. Un token enviado dentro de una solicitud muy larga puede costar el doble. Un token procesado en lote, cuando la respuesta puede esperar, cuesta la mitad. Y un token procesado en una región determinada puede tener un recargo.

El precio efectivo de cada token depende, entonces, de su categoría y de las condiciones en que se procesó. Dos soluciones con el mismo modelo y el mismo proveedor pueden pagar precios muy distintos por el mismo volumen.

El precio por millón es el precio de lista. Lo que se paga es otra cosa.

10. Dos listas de precios

Las listas de precios de OpenAI y Anthropic separan lo que se paga siempre (tokens de entrada y salida) de lo que se puede usar para pagar menos (caché y lotes)²²:

MODELO	TARIFAS PRINCIPALES		OPCIONES DE AHORRO	
	TOKENS DE ENTRADA	TOKENS DE SALIDA	TOKENS LEÍDOS DESDE CACHÉ	PROCESAMIENTO POR LOTES
GPT-5.5 OpenAI	5	30	0,50	50 % menos
Claude Opus 4.8 Anthropic	5	25	0,50	50 % menos

Tarifas expresadas en dólares por millón de tokens. Valores vigentes al 5 de septiembre de 2026.

La tabla anterior distingue entre las tarifas principales de una interacción y las opciones que permiten reducirlas. Cada llamada combina **tokens de entrada** (las instrucciones, documentos y demás información que recibe el modelo) y **tokens de salida** (el contenido que este genera), que se cobran a precios diferentes.

Como opciones de ahorro: la **lectura desde caché** permite reutilizar contenidos de entrada previamente procesados a un precio menor; mientras que el **procesamiento por lotes** ofrece un descuento para tareas que no requieren una respuesta inmediata.

Por eso, el costo efectivo no depende solamente del modelo elegido y de la cantidad total de tokens, sino también de su composición y de la modalidad utilizada para procesarlos.

Ahora bien, la tabla no muestra todo lo que aparece en la factura. OpenAI aplica una tarifa mayor cuando la solicitud supera los 272.000 tokens de entrada: la entrada sube a 10 dólares y la salida a 45. Anthropic cobra aparte la escritura en caché (6,25 dólares por millón si el contenido se conserva cinco minutos, 10 si se conserva una hora) y el tiempo de ejecución de sus agentes administrados. Y ambos cobran la búsqueda web a 10 dólares cada mil consultas: un centavo por búsqueda, que se suma a los tokens que esa búsqueda devuelve.

Entonces, para saber cuánto costaría el informe del Capítulo III en producción, hay que considerar cinco factores adicionales: el tamaño del contexto, el uso de caché, el procesamiento por lotes, la región donde se procesa la consulta y las herramientas que el agente invoca. Las cinco secciones que siguen las recorren de a una.

11. El tamaño del contexto

En nuestro ejemplo, la llamada más grande del agente envió 36.000 tokens. Está muy lejos del umbral. Pero supongamos que el expediente no tiene cincuenta páginas sino seiscientas: cada llamada superaría los 272.000 tokens y, en OpenAI, la solicitud entera pasaría a la tarifa de contexto largo. La entrada se duplica y la salida sube la mitad.²³

Un contexto de un millón de tokens es una mesa enorme. No obliga a usarla toda. Cuando se usa, el precio de cada cosa que está sobre la mesa puede cambiar.

La ventana amplia habilita tareas más grandes. También habilita facturas más grandes.

EN POCAS PALABRAS

Con cincuenta páginas, el informe cuesta 1 dólar. Con seiscientas, cada lectura se cobra al doble.

12. El caché

El caché permite reutilizar partes estables del contexto que el modelo ya procesó, en lugar de procesarlas nuevamente en cada solicitud. Puede aplicarse, por ejemplo, a instrucciones, documentos de referencia, definiciones de herramientas o fragmentos del historial que se repiten.

En los agentes resulta especialmente relevante porque suelen realizar múltiples llamadas al modelo conservando una base común de información e instrucciones. Cuando esa parte permanece sin cambios, puede recuperarse desde el caché, lo que reduce el costo de los tokens de entrada y el tiempo de procesamiento.

Volvamos a las cinco llamadas. El contrato tiene 25.000 tokens y el agente lo envió completo cinco veces. Cinco lecturas del mismo documento: 125.000 de los 151.000 tokens de entrada son el contrato repetido.

Con caché, el contrato se paga entero una vez y las cuatro lecturas siguientes se pagan a una décima parte. Con Opus 4.8, la escritura cuesta unos 16 centavos y las cuatro lecturas, 5 centavos en total. Sin caché, esas mismas lecturas costaban 62 centavos.

El informe baja de 1 dólar a unos 55 centavos. El mismo modelo, el mismo recorrido, casi la mitad del precio.

El ahorro depende de que el contenido efectivamente se reutilice. Almacenar en caché algo que no vuelve a leerse cuesta más que no almacenarlo. Por eso el caché obliga a separar lo estable de lo variable: instrucciones, reglas, documentos base de un lado; la consulta puntual, el estado de la tarea y los resultados de herramientas del otro.

| El caché no abarata el contexto. Abarata el contexto que se repite.

EN POCAS PALABRAS

| El agente lee el mismo contrato cinco veces. Sin caché paga las cinco lecturas enteras. Con caché paga una entera y cuatro a una décima parte. El informe baja de 1 dólar a 55 centavos.

13. Los lotes

El procesamiento por lotes permite agrupar varias tareas y enviarlas para que se ejecuten de manera diferida, en lugar de pedir una respuesta inmediata para cada una. Es la modalidad adecuada para trabajos que pueden esperar, y esa demora es justamente lo que hace que se pague menos.

Los cien contratos del capítulo III, por ejemplo, no necesitan respuesta inmediata: pueden procesarse durante la noche.

Ambos proveedores cobran la mitad por las solicitudes que aceptan esperar. Los 100 dólares del lote de contratos pasan a ser 50. Y como el descuento se suma al caché, el informe que ya bajó a 55 centavos queda cerca de 28.

La condición es la latencia. El lote sirve para clasificar, extraer, evaluar o generar en volumen. No sirve para una persona que espera la respuesta en pantalla.

El precio más bajo del catálogo es para quien puede esperar.

EN POCAS PALABRAS

Si los cien contratos se procesan de noche, cuestan 50 dólares en vez de 100. Y si además se usa caché, cada informe queda en 28 centavos.

14. La residencia de datos

Supongamos ahora que el organismo exige que los contratos se procesen dentro de una región determinada. Por norma, por contrato o por política interna.

Los proveedores ofrecen esa opción y algunos la cobran. En OpenAI, el procesamiento regional tiene un recargo del 10 %. El informe de 1 dólar pasa a 1,10. El de 28 centavos, a 31.

No es una variable técnica ni una optimización. Es un costo de cumplimiento: si la organización debe garantizar dónde se procesan los datos, el recargo no se elige. Se paga.

La comparación de precios se hace bajo las condiciones que la organización necesita, no bajo las que el proveedor publica primero.

EN POCAS PALABRAS

Si los contratos tienen que procesarse en una región determinada, todo lo anterior cuesta un 10 % más. El informe de 1 dólar pasa a 1,10.

15. Las herramientas

Hasta acá el agente trabajó solamente con los documentos. Supongamos que en la tercera llamada hace una búsqueda web para verificar un antecedente.

La búsqueda cuesta un centavo. Ese es el precio visible. Después, los 3.000 tokens que devuelve entran al contexto y viajan en cada una de las llamadas siguientes: 6.000 tokens más de entrada, 3 centavos. Y si el resultado lleva al agente a una lectura adicional del expediente, otra llamada completa.

Tres búsquedas y una relectura, y el informe de 1 dólar pasa a 1,30 sin que nadie haya cambiado de modelo.

Las herramientas actúan en tres niveles. Tienen precio propio. Generan tokens: la definición de la herramienta, la llamada y el resultado. Y disparan trabajo posterior, porque el agente decide sobre lo que la herramienta le devolvió. Un agente con muchas herramientas y sin reglas de uso puede multiplicar el costo sin mejorar el resultado.

La herramienta ejecuta una acción. Pero también produce contexto, y el contexto se paga.

EN POCAS PALABRAS

Cada búsqueda cuesta un centavo, pero lo que la búsqueda trae viaja en todas las llamadas siguientes. Tres búsquedas y una relectura, y el informe pasa de 1 dólar a 1,30.

16. Recapitulemos: el mismo informe, siete precios

El informe es el mismo. El modelo es el mismo. Lo que cambia es cómo se lo procesa.

CÓMO SE PROCESA EL INFORME ²⁴		COSTO EN USD
En lote y con caché		0,28
Con caché		0,55
Al precio de lista (capítulo III)		1,00
Con procesamiento regional		1,10
Con tres búsquedas web y una relectura		1,30
Con búsquedas, en región con recargo		1,45
Expediente de 600 páginas, con búsquedas y recargo regional		>2,00

7x

De 28 centavos a más de 2 dólares. Ninguna de esas diferencias aparece en la lista de precios. Todas aparecen en la factura.

El modelo es el mismo. La arquitectura, no.

17. Del costo por token al costo por tarea

Las cinco variables muestran cuánto cuesta cada llamada. Todavía no dicen cuánto cuesta completar la tarea.

Una tarea puede requerir varias ejecuciones. Puede fallar y repetirse. Puede producir un resultado que una persona deba corregir. Cada uno de esos consumos es parte del costo real, aunque no aparezca en la cuenta de una ejecución exitosa.

Esta forma de medir ya se utiliza para comparar modelos. Artificial Analysis publica una métrica de costo por tarea: el costo promedio ponderado de ejecutar las evaluaciones que integran su Índice de Inteligencia, considerando tokens de entrada, lecturas y escrituras de caché, tokens de razonamiento y tokens de respuesta.²⁵ La métrica muestra algo que el precio por millón oculta: dos modelos con la misma tarifa pueden necesitar cantidades muy distintas de tokens para resolver lo mismo.

Con personas pasa lo mismo, y lo entendemos sin esfuerzo. Un abogado con veinte años de experiencia cobra más la hora. Lee el contrato una vez, va directo a las tres cláusulas que importan y entrega un dictamen que no hay que rehacer. Uno recién recibido cobra la mitad, lo lee tres veces, consulta, redacta un borrador y el socio lo corrige. Por hora, el segundo es más barato. Por tarea, el primero.

El precio por token es el honorario por hora. El costo por tarea son los honorarios totales. Y nadie contrata al socio del estudio para fotocopiar un expediente: para eso, el más barato es la elección correcta.

Acá la analogía tiene un límite, y conviene marcarlo antes de que induzca a error. La experiencia de un abogado se acredita con años de trabajo; la capacidad de un modelo no se lee en su precio. El precio por token refleja el tamaño del modelo, el cómputo que consume y la posición comercial del proveedor. No refleja cuán bien resuelve una tarea concreta.

Un modelo de frontera puede necesitar más pasos que uno intermedio para clasificar correos o extraer fechas de un contrato, porque esa tarea no exige lo que el modelo caro sabe hacer mejor. Y dos modelos con la misma tarifa pueden consumir cantidades muy distintas de tokens para el mismo trabajo: uno va directo, el otro razona de más, reintenta o produce respuestas más largas de lo necesario. La medición de Artificial Analysis muestra precisamente eso.

El modelo más caro no es el más inteligente. Es el más caro. Cuál es el más inteligente para una tarea solamente se sabe midiendo cuántos pasos necesita para completarla.

Por eso, un modelo más caro por token puede ser más barato por tarea si resuelve con menos pasos, menos contexto y menos corrección humana. Y un modelo barato puede ser caro si necesita tres intentos.

La unidad económica relevante es la tarea completada, no la llamada ejecutada.

18. Estrategias de ahorro

Las variables anteriores sugieren por sí mismas las decisiones que reducen el costo.

01	CACHÉ CONTEXTO ESTABLE	Reutilizar el contexto estable mediante caché.
02	LOTES SIN URGENCIA	Procesar en lote lo que no necesita respuesta inmediata.
03	MODELO ADECUADO CÓMPUTO JUSTO	Un modelo económico para clasificar y extraer; uno de frontera solo para lo que lo exige.
04	CONTEXTO NECESARIO SOLO LO NECESARIO	Una ventana grande es capacidad disponible, no una obligación de llenarla.
05	MEDICIÓN REAL EN PRODUCCIÓN	Los costos estimados en el diseño rara vez coinciden con los observados en producción.

Ninguna de estas medidas corrige una arquitectura mal elegida. Sirven para evitar consumo innecesario dentro de una solución que ya se justifica. La optimización empieza por la arquitectura y sigue por estas decisiones. El capítulo V se ocupa de la primera; el VI, de las segundas.

19. Controlar el consumo

Todo lo anterior supone que la organización puede ver cuánto consume. En una conversación simple alcanza con mirar la factura del proveedor. En una arquitectura agéntica, no: un mismo agente llama varias veces al modelo, usa herramientas, procesa documentos y sostiene contexto durante muchos pasos. Si la organización solo ve el total a fin de mes, no puede saber qué decisión de diseño lo produjo.

Las herramientas de control de consumo hacen cuatro cosas. Muestran cuántos tokens se consumen, en qué modelo, en qué paso y con qué herramienta. Atribuyen ese consumo a un área, un proyecto, un usuario o un caso de uso. Fijan límites antes de que el gasto ocurra: un tope mensual por clave de API, un presupuesto por proyecto, una alerta al 80 %. E intervienen cuando el límite se supera: detienen, degradan a un modelo más barato o escalan a una persona.

Un agente sin límite de pasos, sin presupuesto máximo y sin trazabilidad puede seguir iterando aunque cada paso aporte cada vez menos. El problema no es técnico. Es que la organización habilitó una capacidad de gasto variable sin decidir quién puede gastar, cuánto, para qué y cuándo debe detenerse.

El control también permite distinguir gasto de valor. Un flujo caro no es necesariamente ineficiente: puede justificar su costo si acorta un proceso crítico o mejora una decisión. Pero esa evaluación solo es posible si el costo puede vincularse con el resultado. Sin atribución, la organización ve una factura. Con atribución, ve una inversión que puede evaluar.

El costo de la IA agéntica se gobierna antes del despliegue, con límites, y durante la operación, con medición.

20. Lo que pasó en 2026

Todo lo anterior podría leerse como una advertencia teórica. No lo es. En el primer semestre de 2026 varias empresas grandes descubrieron, con la factura en la mano, lo que este capítulo describe.

El caso más visible fue Uber. A fines de 2025 la empresa había armado su presupuesto de IA para 2026 y salió a promover el uso de herramientas de programación agéntica entre sus ingenieros. Lo hizo con entusiasmo: publicó tablas de posiciones internas que rankeaban a los ingenieros según cuánto usaban Claude Code²⁶. La herramienta se extendió entre unos 5.000 ingenieros mucho más rápido de lo que los modelos financieros habían previsto. El costo mensual por ingeniero rondaba los 150 a 250 dólares; los usuarios intensivos, entre 500 y 2.000. El propio CTO, Praveen Neppalli Naga, contó que gastó 1.200 dólares en una sesión de dos horas²⁶.

En abril, cuatro meses después de empezar el año, el presupuesto anual estaba agotado. Naga lo reconoció públicamente y dijo que estaban "de vuelta al pizarrón" con sus supuestos²⁷.

La herramienta no falló. Los ingenieros no la usaron mal. La usaron para lo que fue diseñada: agentes en paralelo, refactorizaciones grandes, generación automática de pruebas. Desde la productividad, el despliegue fue un éxito. Desde las finanzas, una fuga²⁸.

En mayo, el presidente y COO de Uber, Andrew Macdonald, dijo que era difícil trazar una línea entre el consumo creciente de Claude Code y las innovaciones para el cliente: "ese vínculo todavía no está". Quizás se estaba entregando más, pero nadie podía mostrar que la empresa produjera un 25 % más de funciones útiles²⁹.

En junio, Uber fijó un tope: 1.500 dólares mensuales por empleado por cada herramienta de codificación, con excepciones aprobadas caso por caso.³⁰ En agosto, el CTO anunció que la empresa había bajado su costo por token y declaró terminada "la era del tokenmaxxing": la próxima etapa no se caracterizaría por quién gasta más tokens sino por cómo se usan con mayor eficiencia³¹.

Uber no fue la única. Microsoft revocó las licencias de Claude Code de sus desarrolladores meses después de habilitarlas. Y un ejecutivo de Salesforce lo resumió así: la economía de los tokens es más abstracta y opaca que cualquier cosa que la industria haya administrado a esta escala, y exige un músculo operativo distinto del que se construyó para la nube.³²

Leído con las categorías de este trabajo, el caso de Uber no tiene nada de misterioso. Ocurrieron cinco cosas a la vez:

01

PRESUPUESTO

Presupuesto desactualizado

El presupuesto se armó con la lógica de la licencia fija en el momento en que las herramientas pasaban a cobrar por consumo. Los presupuestos de 2026 se cerraron en el otoño de 2025, antes de que las capacidades agénticas se dispararan; quedaron obsoletos antes de empezar a ejecutarse³³.

02

INCENTIVOS

Incentivo sin contrapeso

La organización creó un incentivo a consumir sin ponerle un presupuesto al lado: la tabla de posiciones funcionó, el presupuesto no la sobrevivió.

03

RESPONSABILIDAD

Responsabilidad partida

Los equipos que impulsaban la adopción no eran los que administraban el gasto.

04

CONTROLES

Controles ausentes

Ninguna de las cuatro funciones del apartado anterior estaba en su lugar: no había visibilidad por persona, ni atribución por proyecto, ni límite previo, ni intervención automática. El tope de 1.500 dólares llegó dos meses después de agotado el presupuesto.

05

VALOR

Medición del valor

La pregunta es exactamente la pregunta de la introducción de este trabajo. Tokens no es lo mismo que tareas, y tareas no es lo mismo que valor. Uber podía medir cuántos tokens consumía. No podía medir qué obtenía a cambio.

EN POCAS PALABRAS

Presupuesto, incentivos, responsabilidad, controles y valor: cinco piezas que, cuando no se alinean, hacen que la adopción no sea sostenible.

Queda una última lección, y es la que cierra el círculo con el capítulo III. Uber bajó su costo por token y lo celebró. Pero un token más barato no garantiza una tarea más barata: cuando el precio unitario baja, el consumo tiende a subir, y el gasto total puede terminar siendo mayor. Los economistas lo llaman paradoja de Jevons, y es el riesgo que ahora enfrenta la empresa³⁴.

Nada de esto es exclusivo de las grandes tecnológicas. Una encuesta a 372 empresas encontró que solo el 15 % pronostica sus costos de IA con un margen de error del 10 %, y casi una de cada cuatro se equivoca por más de la mitad³⁵. Uber es noticia porque es Uber. La dinámica es la misma en un organismo público con doscientos empleados.

El presupuesto de tokens no se descubre en la factura. Se decide antes, se mide durante y se ata al valor que produce.

Y la primera decisión, la que condiciona todas las demás, es qué arquitectura habilitar. De eso trata el capítulo siguiente.

CAPÍTULO V

La primera decisión de gobernanza económica: elegir la arquitectura adecuada

Volvamos a los cien contratos. Sabemos que un agente los revisa por unos 100 dólares con un modelo de frontera, que con caché y lote pueden ser 28, y que con herramientas y recargos pueden pasar de 200. Todavía no nos preguntamos algo anterior: si hace falta un agente.

La decisión más importante en el diseño de una solución con IA generativa no es qué modelo elegir ni qué proveedor contratar. Es cómo estructurar el flujo de trabajo. Esa decisión fija cuántas veces se llama al modelo, cuánto contexto procesa, qué herramientas invoca y cuánto control hace falta para supervisarlos. Es decir, fija el presupuesto antes de que exista.

Las formas de automatización que precedieron a la IA generativa no desaparecen con la llegada de los agentes. Siguen siendo indispensables y, en muchos casos, son más económicas, más predecibles y más sencillas de auditar³⁶. La IA agéntica amplía el repertorio de arquitecturas disponibles. No lo reemplaza.

Para este análisis distinguimos tres: el flujo determinista, el flujo híbrido y el agente. La pregunta que las separa es una sola: ¿quién diseña el camino?

Antes de comparar modelos hay que decidir si el modelo conduce, ayuda o no interviene.

21. Flujo determinista: el modelo no interviene

En un flujo determinista, el camino lo diseña una persona y el sistema lo recorre sin desviarse. Si se verifica la condición X, se ejecuta la acción Y; si no, la acción Z. El universo de comportamientos posibles queda definido de antemano y permanece estable durante la operación.

Con los cien contratos:

Entra el contrato → el sistema verifica plazo, partes, firmas y montos contra reglas → si cumple, lo aprueba; si no, lo deriva → informe.

No hay modelo de lenguaje. No hay tokens. El costo se concentra en desarrollar las reglas, mantenerlas y operar la infraestructura, y es fijo y previsible. Su ventaja de gobernanza es igual de clara: cada resultado se explica por la regla que lo produjo, y se puede probar de manera exhaustiva antes de desplegar. Funciona con la lógica de un reloj mecánico: preciso mientras las condiciones sean las previstas; frente a un evento no contemplado, no improvisa, se detiene³⁷.

Ese es también su límite. La regla solo ve lo que se le dijo que vea. No puede leer una cláusula ambigua, interpretar una redacción inusual ni decidir qué hacer con un documento que no encaja en ninguna rama.

Es la arquitectura correcta para tareas repetitivas, estables y formalizables: validar requisitos, calcular plazos, controlar completitud, derivar según criterios cerrados, generar alertas.

Si alguien puede escribir la regla, no hace falta el modelo.

EN POCAS PALABRAS

Los cien contratos cuestan cero tokens. Lo que cuesta es escribir las reglas una vez y mantenerlas.

22. Flujo híbrido: el modelo ayuda en un nodo

El flujo híbrido conserva la estructura determinista y le agrega el modelo de lenguaje en los puntos donde las reglas no alcanzan: donde hay que leer, interpretar, clasificar, extraer o redactar.

Entra el contrato → reglas ordenan y filtran → el modelo lee y extrae las cláusulas relevantes → reglas comparan con los criterios → informe para revisión humana.

El modelo aparece una vez, en un nodo, y no decide el camino. Quién diseñó el flujo definió de antemano cuándo se lo consulta, qué información recibe, con qué instrucciones y qué se hace con su salida. El modelo clasifica, resume o extrae dentro de una ruta que le es dada, no de una que elige.

Con los cien contratos, una llamada por contrato. Con un modelo económico como Gemini 3.5 Flash, unos 6,50 dólares los cien. Con Claude Opus 4.8, unos 20. Y el número es previsible, porque el diseño fija cuántas veces se llama al modelo: cien contratos son cien llamadas, no cien recorridos distintos.

Esta arquitectura no es nueva. Es, con otra tecnología, la que construimos en 2017 con Prometea en la Fiscalía General Adjunta ante el Tribunal Superior de Justicia de la Ciudad de Buenos Aires. El nodo de inteligencia artificial era un sistema de aprendizaje supervisado construido con modelado de tópicos, bolsa de palabras y expresiones regulares: leía las actuaciones previas del caso, las comparaba con miles de sentencias y dictámenes anteriores y predecía la solución aplicable, con un 96 % de acierto promedio. A partir de esa predicción, árboles de decisión automatizaban la generación íntegra del dictamen: preguntas cerradas que descartaban ramas hasta armar el documento, como en el control de autosuficiencia del expediente. La persona revisaba, editaba y firmaba; la competencia seguía siendo del titular del órgano³⁸.

Las cifras de aquel despliegue anticipan las de este libro. Mil dictámenes sobre derecho a la vivienda requerían 174 días de trabajo; con Prometea, 45. Mil dictámenes de empleo público, de 83 días a 5. El control de requisitos formales de mil expedientes, de 160 días a 38³⁹. El modelo predictivo hacía una sola cosa, en un solo lugar del flujo. Las reglas hacían el resto. Ocho años después, con modelos generativos en el nodo, la lógica es la misma.

La ventaja de gobernanza del híbrido es que la estructura determinista sigue definiendo los pasos, los puntos de intervención humana, los límites de uso del modelo y las condiciones bajo las cuales una salida se revisa, se rechaza o se escala. La trazabilidad y la atribución de responsabilidad se conservan.

El híbrido captura el valor del modelo sin asumir el multiplicador del agente.

EN POCAS PALABRAS

**El modelo hace una sola cosa, en un solo lugar, y las reglas hacen el resto.
Cien contratos, entre 6,50 y 20 dólares, y se sabe de antemano.**

23. Agente: el modelo conduce

En el agente, la relación se invierte. Las personas definen el objetivo y los límites; el modelo diseña y ejecuta el camino. Interpreta, planifica, elige herramientas, observa resultados, corrige y decide cuándo terminó.

Entra el contrato → el agente interpreta el objetivo → planifica → lee → busca antecedentes → contrasta → redacta → revisa → decide si terminó → informe.

Cada flecha la decide el modelo. El rasgo definitorio no es la presencia de IA generativa, que ya existía en el híbrido, sino que el agente construye la ruta en tiempo de ejecución a partir de un objetivo, en lugar de seguir una ruta preestablecida⁴⁰.

Con los cien contratos, 100 dólares si cada uno recorre el camino del capítulo III. Pero ese camino no está fijado: un contrato simple puede resolverse en tres llamadas y uno complejo puede exigir doce, dos búsquedas y una relectura. Entre 100 y 600 dólares, y no se sabe cuál de los dos hasta que el agente termina.

Esa variabilidad es la que documentan las estimaciones del capítulo III: entre cinco y treinta veces más tokens por tarea que una interacción estándar según Gartner⁴¹. Lo relevante acá no es el multiplicador sino su dispersión. Dos ejecuciones con el mismo objetivo pueden requerir recorridos distintos y, por tanto, costos distintos.

A la variabilidad del costo se suma la del resultado. Los benchmarks consolidados de capacidad agéntica muestran, para los mejores sistemas comerciales, tasas de resolución de tareas reales que oscilan entre el 30 % y el 70 % según el dominio⁴². Un agente que resuelve la mitad de los casos puede generar valor. Pero la organización que lo despliega paga también la otra mitad: las ejecuciones que fallan, se repiten o producen un resultado que una persona debe rehacer.

Por eso, habilitar un agente es habilitar una capacidad de gasto variable. La organización necesita definir desde el diseño límites de pasos, presupuesto máximo, herramientas disponibles, condiciones de detención, puntos de intervención humana y trazabilidad suficiente para reconstruir cómo se produjo el consumo. La autonomía, en cualquier organización humana, es condicionada por defecto; la de un agente no puede concebirse de otro modo⁴³.

El mayor costo se justifica cuando la tarea exige lo que el determinista y el híbrido no pueden dar: exploración, adaptación a resultados intermedios, selección flexible de herramientas, razonamiento iterativo sobre información que cambia. Cuando la tarea no lo exige, el agente agrega gasto, variabilidad y superficie de riesgo sin un retorno proporcional.

Habilitar un agente es habilitar una capacidad de gasto variable.

EN POCAS PALABRAS

Nadie sabe cuántas veces va a leer el contrato hasta que lo lee. Cien contratos, entre 100 y 600 dólares, y se sabe después.

24. La regla del diseño escalonado

La recomendación general es partir de la arquitectura más simple capaz de resolver adecuadamente el problema y aumentar la complejidad sólo cuando exista una razón funcional, económica y de gobernanza que lo justifique.

En términos operativos, son tres preguntas en orden:

PRIMERA PREGUNTA

¿Alguien puede escribir la regla?

No hay necesidad de incorporar IA generativa.

SI ES SÍ

Flujo determinista

SEGUNDA PREGUNTA

¿Hay un paso que exige leer, interpretar, clasificar o redactar?

El modelo entra en ese paso y en ningún otro.

SI ES SÍ

Flujo híbrido

TERCERA PREGUNTA

¿El camino no se puede fijar de antemano porque depende de lo que se va encontrando?

Y con límites definidos antes de la primera ejecución.

SI ES SÍ

Agente

CADA ESCALÓN AGREGA COSTO, VARIABILIDAD Y SUPERFICIE DE RIESGO

El **criterio económico** es la relación entre el valor esperado de la tarea y el costo adicional que introduce cada escalón. Si una tarea se resuelve con un híbrido a un costo bajo y predecible, elegir un agente no mejora la gobernanza: aumenta el gasto, la variabilidad y la superficie de riesgo sin un retorno proporcional. Si el problema requiere exploración y adaptación que no pueden anticiparse, el agente puede estar justificado.

Como no tomamos un avión para ir a la esquina, las soluciones tecnológicas deben responder proporcionalmente a las necesidades de la organización que las adopta.⁴⁴

La pregunta no es "¿podemos usar un agente?" sino "¿cuál es la arquitectura mínima suficiente para resolver este problema con un nivel aceptable de costo, riesgo y control?".

Una escalera con historia. Conviene decirlo porque suele presentarse como una novedad de la era agéntica: la regla es la misma desde antes de la IA generativa. En 2018, al presentar Prometea, distinguimos tres matices de automatización: la automatización completa por reglas, la automatización con intervención humana reducida y la automatización más predicción⁴⁵. En 2021 propusimos ordenar la autonomía de los sistemas por escalas, como el estándar de los vehículos autónomos, en lugar de preguntar si un sistema "es o no es" autónomo⁴⁶. En 2025, con la IA generativa consolidada, la escala pasó a cinco niveles, desde la automatización sin IA hasta los agentes en sentido estricto.⁴⁷ Y en 2026 la simplificamos en los tres niveles que usa este libro, con la pregunta que los ordena: quién diseña el camino⁴⁸.

Lo que cambió en ocho años es el escalón más alto, que antes no existía. Lo que no cambió es la regla para subir: se sube cuando el escalón anterior no alcanza, no cuando el siguiente está disponible.

La facilidad de activar no debería confundirse con la simplicidad de gobernar⁴⁹.

25. Un segundo ejemplo: la bandeja de correo

La misma escalera se ve en una tarea que cualquier oficina reconoce: revisar la bandeja de entrada, clasificar los mensajes y preparar una respuesta cuando corresponde.

Con reglas, la clasificación se hace por remitente, asunto, palabras clave y etiquetas; ciertos mensajes se derivan, otros disparan una respuesta preestablecida. Sin tokens, con costo estable. Con un híbrido, las reglas ordenan la bandeja y el modelo interviene solo donde el contenido es ambiguo: clasifica esos mensajes, resume los extensos, prepara un borrador para revisión humana. El consumo queda acotado a esas funciones. Con un agente, el sistema recibe un objetivo más amplio, gestionar lo pendiente, y para cada mensaje decide qué información necesita, consulta documentos o sistemas internos, redacta y evalúa si tiene suficiente antes de cerrar. La secuencia varía con cada mensaje y con ella el costo.

La tarea es la misma. Lo que cambia es quién diseña el camino y, con eso, cuánto cuesta recorrerlo y cuánto se sabe de antemano.

26. Recapitulemos: la misma tarea, tres arquitecturas

DETERMINISTA	HÍBRIDO	AGENTE
CIEN CONTRATOS Desarrollo y mantenimiento de reglas	CIEN CONTRATOS 6,50 a 20 USD	CIEN CONTRATOS 100 a 600 USD
QUIÉN DISEÑA EL CAMINO La persona	QUIÉN DISEÑA EL CAMINO La persona; el modelo, un nodo	QUIÉN DISEÑA EL CAMINO El modelo
LLAMADAS AL MODELO Ninguna	LLAMADAS AL MODELO Fijas por diseño	LLAMADAS AL MODELO Variables por ejecución
CONSUMO DE TOKENS Cero	CONSUMO DE TOKENS Acotado	CONSUMO DE TOKENS Multiplicado
PREVISIBILIDAD DEL COSTO Alta	PREVISIBILIDAD DEL COSTO Alta	PREVISIBILIDAD DEL COSTO Baja
CUÁNDO SE CONOCE EL COSTO Al diseñar	CUÁNDO SE CONOCE EL COSTO Al diseñar	CUÁNDO SE CONOCE EL COSTO Al terminar

De cero a 600 dólares para la misma tarea. Ninguna de esas diferencias depende del modelo. Todas dependen de la arquitectura.⁵⁰

La arquitectura no es una decisión técnica que después tiene consecuencias económicas. Es la decisión económica.

Una vez elegida la arquitectura, quedan las decisiones que la hacen más o menos cara de operar: dónde corre el modelo, cuál se usa para cada paso, cuánto contexto carga y cómo se reduce. De eso se ocupa el capítulo siguiente.

CAPÍTULO VI

Estrategias de optimización y despliegue

Volvamos por última vez a los cien contratos. Elegimos la arquitectura: híbrida, a unos 20 dólares los cien, o agéntica, a unos 100. Ya sabemos que el caché y el lote bajan esa cuenta y que las herramientas y los recargos la suben. Queda un nivel más de decisiones, el que aparece cuando el uso escala: cómo organizar los recursos para que tareas de distinta complejidad, sensibilidad y valor no consuman indiscriminadamente lo mismo.

Son cuatro perillas. Dónde corre el modelo. Qué modelo se usa en cada paso. Cuánto contexto carga. Y cuánto razona antes de responder. Las cuatro son técnicas en su implementación y económicas en sus efectos: mueven el costo total, la dependencia de proveedores, el control sobre los datos y la capacidad de escalar.

Elegida la arquitectura, todavía queda decidir cuánto cuesta operarla.

27. Dónde corre el modelo

Hasta acá, el modelo corrió en la infraestructura del proveedor y cada llamada se pagó por token. Hay otra opción: correr un modelo de pesos abiertos en un servidor propio o alquilado. La organización deja de pagar por token y pasa a pagar por capacidad.

Es la misma lógica de la suscripción del capítulo III, llevada al nivel de la institución. La suscripción convierte el consumo variable de una persona en un precio fijo con techo en el tiempo. El servidor convierte el consumo variable de una organización en un precio fijo con techo en la capacidad. En los dos casos, el precio se conoce de antemano y el límite aparece cuando se usa mucho.

El servidor es la suscripción de la organización.

Pongámosle números. Cien contratos por día son 2.200 al mes. Por API, el flujo híbrido cuesta entre 143 dólares mensuales con Gemini 3.5 Flash y 440 con Claude Opus 4.8; el agéntico con Opus, unos 2.150. Del otro lado, una GPU capaz de correr un modelo abierto comparable se alquila, al 5 de septiembre de 2026, entre 1,73 y 3,32 dólares la hora según el proveedor: entre 1.250 y 2.400 dólares al mes las veinticuatro horas. Comprarla cuesta entre 30.000 y 40.000 dólares, más el servidor que la aloja⁵¹. A eso hay que sumar lo que ninguna tarifa muestra: la persona que instala, actualiza, monitorea y asegura el entorno.

La cuenta tiene dos lecturas. Para el flujo híbrido, el servidor no conviene: 2.200 contratos por mes cuestan por API menos de lo que cuesta tener la GPU encendida, y la GPU pasaría más del 90 % del tiempo ociosa. Para el flujo agéntico, la cuenta se acerca: 2.150 por API contra 1.250 a 2.400 por el servidor. Y si los contratos fueran mil por día, el servidor gana con claridad.

La regla que sale de ahí es contraintuitiva: cuanto más caro es el recorrido por API, antes conviene el servidor propio. Un uso agéntico intensivo llega al punto de equilibrio con mucho menos volumen que un uso híbrido acotado. Por eso la decisión no depende sólo de cuántas tareas hay, sino de cuántos tokens consume cada una.

Qué modelo correr cambia la cuenta. Qwen3.8-27B, publicado en agosto de 2026 con licencia Apache, corre en una sola GPU y alcanza para extraer y clasificar. DeepSeek V4-Pro, el modelo más barato de la tabla del capítulo III y también de pesos abiertos, necesita ocho GPUs de ese tipo para ejecutarse: unos 19.000 dólares al mes de alquiler. El modelo abierto más barato de usar por API es uno de los más caros de correr uno mismo⁵².

Hay además razones que no son de costo. La sensibilidad de los datos, los requisitos de residencia o soberanía y la intención de reducir la dependencia de un proveedor pueden justificar un despliegue propio aunque salga más caro. Pero conviene precisar qué se compra con eso. Correr la aplicación en servidores propios y llamar a la API de un proveedor externo no es soberanía: los datos del prompt salen igual.

La soberanía total exige que el modelo también corra adentro; hay un grado intermedio en el que el modelo corre adentro y solo las herramientas consultan afuera; y hay un grado contractual en el que el modelo corre en la nube del proveedor con garantías sobre retención y uso⁵³. Cada grado tiene un precio y un nivel de control distintos, y el recargo por residencia de datos del capítulo IV es la versión más barata del último.

Los modelos locales no son gratis. Cambian tokens por infraestructura, y la infraestructura se paga aunque no se use.

EN POCAS PALABRAS

Con 2.200 contratos al mes, la API es más barata para el híbrido y empata con el servidor para el agente. Con 20.000, el servidor gana. Y si el motivo es que los datos no salgan, el servidor se justifica aunque pierda.

28. Qué modelo en cada paso

No todas las tareas justifican el mismo modelo. Es la lección del abogado del capítulo IV: nadie contrata al socio para fotocopiar. El enrutamiento consiste en asignar cada tarea al modelo suficiente para resolverla, según su complejidad, su sensibilidad y el riesgo de equivocarse.

Con los cien contratos, supongamos que 80 son estándar y 20 tienen cláusulas inusuales. En el flujo agéntico, los 80 van a Gemini 3.5 Flash, a 31 centavos cada uno, y los 20 a Claude Opus 4.8, a 98 centavos. Total: 44 dólares en lugar de 98. En el híbrido, la misma división baja de 20 dólares a 9.

La mitad del costo, sin cambiar la arquitectura ni el resultado, con una sola regla: decidir antes de llamar al modelo cuál modelo llamar. El criterio puede ser una clasificación previa del contrato, la longitud del documento, la presencia de ciertos términos o, dentro de una misma ejecución, el tipo de paso: un modelo de frontera para planificar y contrastar, uno económico para leer y extraer.

El servidor propio de la sección anterior es un destino más del enrutamiento. Los 80 contratos estándar pueden ir al modelo abierto en la GPU propia, que ya está pagada, y los 20 complejos a la API. La organización obtiene el grado intermedio de soberanía para el grueso del trabajo y paga el modelo de frontera solo donde lo necesita.

El enrutamiento también admite criterios de gobernanza. Los datos sensibles se procesan en el entorno local; las operaciones de mayor impacto requieren el modelo más robusto, controles adicionales o revisión humana. La decisión no es usar siempre el modelo más potente ni siempre el más barato: es no pagar capacidad que la tarea no usa.

El modelo por defecto es la decisión de costo más barata de tomar y la más cara de no tomar.

EN POCAS PALABRAS

80 contratos simples al modelo económico y 20 difíciles al de frontera: 44 dólares en vez de 98. La regla se escribe una vez.

29. Cuánto contexto carga

Volvamos a las cinco llamadas del capítulo III. La primera envió 25.000 tokens; la quinta, 36.000. La diferencia es el historial: cada paso arrastró lo que produjeron los anteriores. El caché abarató las relecturas del contrato. Pero hay una pregunta anterior: ¿hace falta releerlo entero?

La compresión de contexto persigue un objetivo distinto del caché. El caché reduce el precio de reutilizar lo que se repite; la compresión reduce la cantidad de información que entra al modelo en cada paso. Puede hacerse resumiendo, eliminando duplicados, descartando lo irrelevante, recuperando solo los fragmentos necesarios o representando de forma compacta los resultados previos.

Supongamos que después de la primera lectura el agente conserva un resumen de 5.000 tokens del contrato en lugar de los 25.000 originales, y que en cada paso agrega solo sus hallazgos. Las cinco llamadas suman unos 65.000 tokens de entrada en lugar de 151.000. El informe baja de 1 dólar a unos 55 centavos, sin caché. Con caché sobre lo que sigue repitiéndose, menos.

El riesgo está en la palabra "resumen". Comprimir es decidir qué se conserva y qué se descarta, y un resumen mal hecho puede omitir la cláusula que importaba. Por eso la compresión exige criterios de validación, trazabilidad y, cuando el caso lo requiera, la posibilidad de volver al documento original. El objetivo no es cargar lo mínimo, sino lo mínimo suficiente.

El ecosistema ya trata esto como una capa propia. Los proveedores documentan mecanismos de compactación del historial en sus API y sus herramientas agénticas lo aplican de manera automática cuando el contexto se acerca al límite; proyectos de código abierto como Headroom ofrecen esa función como componente independiente. Las cifras de ahorro que reportan son de sus propios benchmarks, no porcentajes generalizables; la tendencia es lo relevante⁵⁴.

El contexto se paga en cada llamada. Lo que no entra, no cuesta.

EN POCAS PALABRAS

Si el agente guarda un resumen del contrato en vez del contrato entero, el informe baja de 1 dólar a 55 centavos. Si el resumen omite la cláusula clave, ese ahorro fue el más caro del proceso.

30. Cuánto razona antes de responder

Hay una cuarta perilla que suele pasar inadvertida porque no está en la tabla de precios: cuánto "piensa" el modelo antes de responder.

Los modelos actuales pueden generar tokens de razonamiento intermedio antes de producir la respuesta, y los proveedores permiten fijar ese nivel de esfuerzo, de mínimo a máximo, o dejar que el modelo lo ajuste solo. Esos tokens no aparecen en pantalla, pero se facturan como tokens de salida, que son los más caros.⁵⁵ Un modelo que razona en profundidad puede producir varias veces más tokens de salida que uno que responde directo, para la misma pregunta.

En los cien contratos, extraer las cláusulas no necesita razonamiento profundo: es una lectura con criterios. Contrastar el contrato con la documentación y detectar una inconsistencia sí puede necesitarlo. Fijar el esfuerzo bajo en el primer paso y alto en el tercero es una decisión de diseño con precio, y hoy se toma casi siempre por omisión, con el valor que el proveedor puso por defecto.

Este es el mismo principio del enrutamiento aplicado dentro de un modelo: no pagar capacidad que la tarea no usa. Y tiene el mismo límite: un esfuerzo demasiado bajo en un paso que lo requería produce un error que se paga después, en revisión o en reintento.

El razonamiento es salida, y la salida es lo que más cuesta.

EN POCAS PALABRAS

El modelo puede pensar poco o mucho antes de responder, y cada pensamiento se cobra. Para leer, poco. Para contrastar, más. Lo peor es no decidirlo.

31. El caso Coinbase: la respuesta contraria a Uber

El capítulo IV terminó con Uber agotando su presupuesto en cuatro meses y respondiendo con un tope de 1.500 dólares por empleado. Dos meses después, otra empresa publicó la respuesta opuesta al mismo problema.

El 27 de junio de 2026, el CEO de Coinbase, Brian Armstrong, explicó cómo la compañía había reducido su gasto en IA casi a la mitad mientras el consumo de tokens seguía creciendo de manera exponencial. Su punto de partida fue un dato: el 91 % de sus empleados nunca llegaba a los topes de uso. Bajar los topes habría generado fricción y alertas para resolver un problema que solo tenía el 9 %. En lugar de eso, cambiaron lo que ocurre por defecto⁵⁶.

Las palancas que describió son cinco, y son las de este capítulo.

- 1. Valores por defecto:** la puerta de enlace interna deriva las solicitudes a modelos de pesos abiertos más económicos, y quien necesita un modelo de frontera lo elige.
- 2. Enrutamiento:** los prompts se preprocesan y se envían al modelo adecuado según precio y probabilidad de acierto en caché; un modelo de frontera para planificar, no para ejecutar.
- 3. Caché:** la tasa de aciertos pasó del 5 % al 60 %.
- 4. Disciplina de contexto:** sesiones nuevas al cambiar de tarea y menos archivos cargados.
- 5. Visibilidad:** cada ingeniero elige el modelo que quiere, pero ve lo que cuesta.

La diferencia con Uber no es de resultado sino de mecanismo. Uber gobernó el gasto con un límite sobre las personas. Coinbase lo gobernó con la infraestructura: cambió el camino que las solicitudes recorren por defecto y dejó las excepciones en manos de quien las necesita. La primera respuesta es más rápida de implementar; la segunda, más difícil de construir y más barata de operar.

Conviene anotar lo que la publicación no dice. Los modelos abiertos que Coinbase adoptó por defecto son de origen chino, y su uso en industrias reguladas plantea consideraciones de licencia, residencia de datos y cumplimiento que cada organización debe evaluar. Tampoco informa si el enrutamiento produce degradación de calidad en los casos límite⁵⁷. El caso ilustra el mecanismo, no dispensa de las preguntas del capítulo V.

El costo se gobierna mejor en el camino por defecto que en el tope por persona.

32. Recapitulemos: cuatro perillas y un límite

Las cuatro decisiones de este capítulo, aplicadas a los cien contratos del recorrido agéntico con Claude Opus 4.8, que al precio de lista costaban 98 dólares.

DECISIÓN	QUÉ CAMBIA	CIEN CONTRATOS	RIESGO
1 Dónde corre el modelo	Tokens por capacidad	No conviene a este volumen; conviene desde unos 20.000 al mes	Capacidad ociosa; responsabilidad operativa
2 Qué modelo en cada paso	80 al económico, 20 al de frontera	98 → 44 USD	Degradación silenciosa en casos límite
3 Cuánto contexto carga	Resumen en lugar del documento entero	98 → 55 USD	Omitir lo que importaba
4 Cuánto razona	Esfuerzo bajo donde no hace falta	Menor salida; depende del paso	Error donde el razonamiento era necesario

< 20 USD

Las cuatro decisiones se combinan: enrutamiento más compresión más caché pueden llevar los cien contratos por debajo de 20 dólares.

Combinadas, las cuatro perillas pueden bajar el costo unitario a una quinta parte.⁵⁸ Y ahí aparece el último punto, que el capítulo IV dejó planteado con Uber.

Todas estas estrategias bajan el costo por tarea. Ninguna baja el gasto total por sí sola. Cuando el costo unitario cae, el consumo tiende a subir: se habilitan más casos de uso, más personas, más pasos por ejecución. Es la paradoja de Jevons, y es exactamente lo que Uber enfrenta después de haber bajado su costo por token.⁵⁹ Coinbase la evitó por una razón que su CEO enunció al pasar: mantuvo la visibilidad del costo y los topes existentes, y cambió los defaults por debajo de ellos.

Optimizar y presupuestar son dos cosas distintas. La primera baja el precio de cada recorrido. La segunda decide cuántos recorridos se financian. Una organización que hace la primera sin la segunda descubre, más tarde, que gasta más que antes con tokens más baratos.

| La optimización sin presupuesto es una invitación a consumir más.

Con esto cierra el recorrido del costo: cómo se forma, cómo se convierte en factura, qué arquitectura lo define y qué decisiones lo optimizan. Lo que sigue es lo que una organización hace con todo eso.

CAPÍTULO VII

El costo que la factura no muestra

El informe del capítulo III costó 1 dólar. El agente lo produjo en treinta segundos. Ahora sigamos la historia después de que el agente terminó.

Alguien tiene que leer el informe. Verificar que las cláusulas identificadas son las relevantes, que el contraste con la documentación es correcto, que no hay un dato inventado. Después, ajustarlo: el matiz del caso, el formato institucional, el tono. Y si quien lo revisa no entiende por qué el agente concluyó lo que concluyó, la próxima vez el ciclo empieza de cero.

Nuestro propio trabajo sobre sedentarismo cognitivo describe ese ciclo en cinco fases: instrucción, ejecución, auditoría, corrección e integración, y aprendizaje institucional. En el ejemplo de un dictamen, la ejecución dura treinta segundos, la auditoría cuarenta y cinco minutos y la corrección treinta más.⁶⁰ El dólar del capítulo III paga la segunda fase. Las otras cuatro no aparecen en ninguna factura de ningún proveedor.

Pongámosle precio. Si la hora de quien firma vale 60 dólares, una hora y cuarto de auditoría y corrección son 75 dólares. El costo humano del informe es 75 veces el costo del token. Todo lo que hicimos en el capítulo VI para bajar el dólar a 28 centavos actuó sobre el 1 % del costo real.

Los cuatro niveles de la introducción eran token, ejecución, tarea y valor. Los capítulos anteriores pusieron precio a los dos primeros. Este capítulo pone precio a lo que hay entre la ejecución y la tarea completada: el juicio humano que convierte una salida en un resultado utilizable. Y a lo que hay entre la tarea y el valor: la capacidad de la organización de seguir sabiendo lo que delegó.

El token se paga al proveedor. El juicio se paga adentro, y no aparece en ninguna factura.

33. El ciclo completo y la parte que no se acelera

En 1967 Gene Amdahl demostró que el rendimiento de un sistema tiene un techo fijado por la parte que no puede acelerarse. No importa cuánto se comprima el resto: el tiempo total sigue dependiendo de lo más lento⁶¹.

Aplicado al informe: la redacción que llevaba cuatro horas ahora lleva treinta segundos, y la revisión que llevaba una hora sigue llevando una hora. Antes, la revisión era el 20 % del ciclo. Ahora es el 99 %. La IA no eliminó el trabajo humano; lo concentró en la parte que exige más juicio y lo volvió, proporcionalmente, el único costo que queda.

Esto cambia la lectura de todos los números anteriores. Medir el ahorro por la fase de ejecución es medir la velocidad de un auto sin contar el tránsito ni el estacionamiento: técnicamente correcto, operativamente engañoso. Comparado con una persona que hubiera hecho el dictamen sin IA en dos horas, el ciclo completo de una hora y cuarto ahorra 45 minutos, no una hora y cincuenta y nueve.

El capítulo III dijo que un token más barato no garantiza una tarea más barata. Acá está la versión que faltaba.

| Una ejecución más rápida no garantiza un ciclo más corto.

EN POCAS PALABRAS

| El agente tarda treinta segundos; la persona que lo revisa, una hora y cuarto.
La factura muestra los treinta segundos.

34. Tres formas de delegar, tres precios

La auditoría cuesta lo que cuesta cuando quien revisa puede revisar. Delegar no es un fenómeno uniforme, y la diferencia entre sus formas tiene precio.

Hay tres niveles de delegación que conviene distinguir⁶²:

DESCARGA EJECUCIÓN JUICIO CONTROL Se delega la ejecución y se conservan el juicio y el control. El abogado que leyó el contrato le pide al agente el informe y puede evaluarlo: sabe qué cláusula falta, qué contraste está mal, qué dato no existe. COSTO DE LA AUDITORÍA 45 minutos Y ese es el costo correcto: es lo que vale que el resultado sea confiable.
TERCERIZACIÓN EJECUCIÓN JUICIO CONTROL Se delegan la ejecución y el juicio, no se puede evaluar el resultado. Quien recibe el informe no conoce el contrato ni el derecho aplicable; lee, le suena bien y lo usa. COSTO DE LA AUDITORÍA Cero Y ese cero es una ilusión: el costo aparece después, como error, como reintento o como decisión equivocada, y se paga con intereses.
RENDICIÓN EJECUCIÓN JUICIO CONTROL Se delega la ejecución, el juicio y el control. La persona adopta la respuesta con más confianza que en su propio razonamiento, y si más tarde aparece el error, lo recuerda como propio. El error viene envuelto en certeza. 92,7 % ADOPTÓ RESPUESTAS CORRECTAS 79,8 % ADOPTÓ RESPUESTAS INCORRECTAS Tres estudios preregistrados con 1.372 participantes; la confianza subió después de consultar a la IA en los dos casos.⁶³ COSTO DE LA AUDITORÍA Nadie audita Porque nadie siente que haga falta.

La adulación de los modelos acelera el pasaje del primer nivel al tercero. Los modelos tienden a validar al usuario aunque esté equivocado. Un estudio publicado en Science en 2026 mostró, en once modelos de frontera, que la IA afirma las acciones del usuario un 50 % más que las respuestas humanas, y que una sola interacción con un modelo adulator reduce la disposición a reconsiderar. Lo más costoso es el ciclo: los participantes calificaron esas respuestas como de mayor calidad, confiaron más y expresaron mayor intención de volver⁶⁴. Nuestro propio experimento de 2024 encontró que, cuando la adulación se hace visible, la confianza cae; el problema es que casi nunca se hace visible⁶⁵.

En términos de costo: la descarga paga la auditoría hoy. La tercerización la paga mañana, más cara. La rendición no la paga nunca, hasta el día en que paga todo junto.

El costo de auditar no desaparece cuando nadie audita. Se difiere.

EN POCAS PALABRAS

Si quien revisa el informe sabe de contratos, la revisión cuesta una hora y vale la pena. Si no sabe, cuesta cero y el error se cobra después, con intereses.

35. Las cuatro paradojas, en dinero

Las paradojas de la automatización se conocen desde Bainbridge, en 1983⁶⁶. El ecosistema agéntico las potencia. Y cada una tiene una traducción contable que el presupuesto de tokens no registra.

El AI slop. Cuando generar es barato, se genera de más: informes que nadie lee, presentaciones redundantes, mensajes que saturan. Una encuesta a 1.150 trabajadores encontró que el 40 % había recibido en el último mes contenido generado con IA que aparentaba ser trabajo hecho y no lo era, con un costo promedio de casi dos horas de retrabajo por incidente⁶⁷. El mecanismo es una transferencia: el productor ahorra tiempo y el receptor lo paga. En los cien contratos, cada informe que llega con una cláusula mal identificada le cuesta dos horas a quien lo recibe. Diez de cien son veinte horas: 1.200 dólares en horas, contra 100 en tokens.

La descualificación. Cuando un sistema asume una tarea, la persona deja de practicarla y la habilidad se degrada por desuso. Hay dos formas: la de quien sabía y dejó de saber, y la de quien nunca supo porque se formó con la herramienta ya integrada y aprendió a supervisar lo que no aprendió a producir⁶⁸. La segunda es la más cara, porque sin la base no hay con qué detectar el error: la auditoría del apartado anterior se vuelve imposible aunque se pague.

Y tiene una consecuencia económica directa. A un empleado de Priceline la renovación de su herramienta de programación le llegó cuatro o cinco veces más cara. Si nadie en el equipo sabe hacer el análisis sin la herramienta, el proveedor puede fijar el precio que quiera. La descualificación convierte un costo variable en una dependencia, y la dependencia no tiene precio de lista.

La complacencia. Se automatiza porque el sistema es más preciso, y la supervisión que debería respaldarlo se degrada precisamente porque el sistema es más preciso. Quien revisa un agente que acierta el 95 % de las veces reduce su escrutinio, acepta sin verificar y no detecta la anomalía cuando llega⁶⁹. El costo es una tasa de error que crece mientras la organización cree que baja. Y el remedio habitual, "poner un humano en el circuito", está sobrevalorado: la persona cree que controla al sistema cuando el sistema está reconfigurando su capacidad de controlar⁷⁰.

El cerebro frito. La automatización no elimina la carga cognitiva: la concentra en los momentos de mayor complejidad. Una encuesta a 1.488 trabajadores encontró que supervisar herramientas de IA es la forma de uso más exigente: 14 % más esfuerzo mental, 12 % más fatiga y 19 % más sobrecarga de información que quienes supervisan menos. Entre quienes lo padecen, la intención de dejar la organización sube del 25 % al 34 %⁷¹. La rotación tiene precio conocido en cualquier área de recursos humanos, y es varias veces el salario mensual de quien se va.

Ninguno de estos cuatro costos figura en la factura del proveedor. Los cuatro figuran en la organización.

El presupuesto de tokens mide lo que la máquina hace. No mide lo que le pasa a la gente que la supervisa.

EN POCAS PALABRAS

Generar barato produce de más, y cada producto de más le cuesta dos horas a alguien. Delegar todo atrofia, y lo atrofiado no puede auditar. Supervisar cansa más que hacer, y el cansado se va.

36. El cuello de botella y el colapso

Volvamos a los cien contratos, pero ahora son mil por día. El agente los procesa. El equipo que los revisa es de tres personas y puede revisar cincuenta. Los errores que hay en los otros novecientos cincuenta, si los hay, salen sin que nadie los vea.

Este es el cuello de botella cognitivo, y cambia de escala con la arquitectura. En la IA de las tareas, la persona revisa una salida por vez y el ciclo es corto. En la IA de los procesos, el sistema genera muchas salidas simultáneas y la persona elige cuáles controlar. En los sistemas multiagente autónomos, el agente actúa y la persona se entera después: el control es retrospectivo, y en muchos contextos legales e institucionales eso no alcanza⁷².

En dinero: el "ahorro" en los novecientos cincuenta contratos no revisados no es ahorro. Es riesgo transferido al futuro, con una probabilidad de materializarse que la organización no midió porque no revisó. El capítulo IV distinguió el costo de una ejecución del costo de una tarea completada satisfactoriamente. Acá está la diferencia: la ejecución costó 1 dólar; la tarea, si nadie la validó, no se sabe cuánto costó todavía.

Y cuando el volumen de salidas excede la capacidad de procesarlas, aparece lo que nuestro trabajo llama colapso cognitivo: la saturación aguda de atención y juicio frente a un flujo que los desborda, sobre capacidades que ya se atrofiaron por delegación sostenida. Sedentarismo y colapso operan juntos: la persona cuyas competencias se degradaron es la misma que enfrenta demandas de supervisión cada vez más veloces⁷³.

El capítulo IV terminó con Uber sin poder mostrar que sus tokens produjeran funciones útiles. La razón está acá. Cuando la ejecución se abarata, el consumo sube; cuando el consumo sube, las salidas se multiplican; y cada salida pide un auditor que no se multiplicó. La paradoja de Jevons tiene una versión cognitiva.

La ejecución barata multiplica las salidas. Cada salida pide un auditor. Los auditores no se multiplican.

EN POCAS PALABRAS

Mil informes por día y tres personas para revisarlos. Lo que no se revisa no es más barato: es un error que todavía no apareció.

37. Cómo se presupuesta

Todo esto tiene una consecuencia práctica para el presupuesto de cualquier iniciativa agéntica: no tiene una línea sino tres.

Primera: tokens. Es la que este libro desarrolló en los capítulos III a VI. Modelo, arquitectura, caché, lote, herramientas, enrutamiento, contexto, razonamiento.

Segunda: horas humanas de dirección, auditoría y corrección. Es la que este capítulo desarrolló, y en la mayoría de los casos es la mayor de las tres. Incluye la fase de instrucción, que puede llevar semanas de iteración antes del primer resultado utilizable, y que no aparece en ninguna métrica.

Tercera: gimnasio cognitivo. Lo que la organización invierte para que la capacidad de auditar no se atrofie: práctica deliberada sin la herramienta en las tareas que definen la competencia, rotación entre producir y supervisar, puntos de control diseñados para que la fricción cognitiva siga activa.⁷⁴ Es la línea que ninguna organización presupuesta y la única que garantiza que la segunda siga siendo posible.

	FASE DEL CICLO	QUIÉN LA PAGA Y DÓNDE APARECE	EN EL INFORME DE 1 DÓLAR
1	Instrucción	La organización En ninguna métrica	Semanas de diseño, una vez
2	Ejecución ● EN LA FACTURA	El proveedor factura	30 segundos 1 DÓLAR
3	Auditoría	La persona que firma En su agenda	45 minutos
4	Corrección e integración	La persona que firma En su agenda	30 minutos
5	Aprendizaje institucional	La organización En lo que sabe hacer sin la herramienta	O se hace, o el ciclo vuelve a cero

↶ LA QUINTA FASE VUELVE A LA PRIMERA

30 segundos

FACTURADOS

75 minutos

FUERA DE TODA FACTURA⁷⁵

La pregunta con la que abrió este libro fue cuánto cuesta alcanzar un resultado, bajo qué arquitectura y qué valor obtenemos a cambio. La primera parte de la respuesta se paga en tokens y se lee en la factura. La segunda se paga en horas de juicio y se lee en la agenda de quien firma. La tercera se paga en capacidad preservada y se lee, con años de retraso, en lo que la organización todavía sabe hacer.

Gobernar el costo de la IA agéntica es gobernar las tres líneas. Quien mira solo la primera está optimizando el 1 %.

CAPÍTULO VIII

Conclusiones

Este trabajo empezó con una persona que abre ChatGPT, Claude o Gemini, escribe una instrucción y recibe una respuesta. Y con otra escena: la misma persona le pide al sistema que investigue, compare, calcule, redacte y revise. Las dos respuestas podían ocupar las mismas dos páginas. El recorrido que las produjo, no.

Los capítulos siguientes recorrieron ese trayecto con un solo ejemplo, un contrato de cincuenta páginas y un informe de una página, y le fueron poniendo precio a cada tramo. Conviene recapitular lo que ese recorrido dejó.

Lo que el ejemplo mostró

Una llamada, 20 centavos. Un agente, 1 dólar. El mismo modelo, el mismo informe, cinco veces más caro, porque el agente lee el contrato cinco veces y arrastra el historial en cada paso. Un token más barato no garantiza una tarea más barata.

El dólar aparece en dos facturas. Por suscripción, el precio es fijo y el límite es el tiempo: varios informes y la herramienta pide esperar. Por API, el precio es variable y el límite es el presupuesto: cuarenta contratos son 40 dólares, doscientos son 200. Si la organización no pone el techo, no hay techo.

El precio de lista no lo paga nadie. El mismo informe cuesta 28 centavos en lote con caché, 1 dólar al precio publicado, 1,30 con herramientas y más de 2 dólares con contexto largo y recargo regional. Siete precios para la misma página, y ninguna de las diferencias está en la lista de precios.

La arquitectura es la decisión económica. Cien contratos cuestan cero tokens con reglas, entre 6,50 y 20 dólares con un flujo híbrido y entre 100 y 600 con un agente. El determinista y el híbrido saben el costo al diseñar; el agente lo sabe al terminar. Se sube de escalón cuando el anterior no alcanza, no cuando el siguiente está disponible.

Cuatro perillas bajan el costo unitario. Dónde corre el modelo, cuál se usa en cada paso, cuánto contexto carga y cuánto razona. Combinadas, llevan los cien contratos de 98 dólares a menos de 20. Pero ninguna baja el gasto total por sí sola: cuando el costo unitario cae, el consumo sube.

Y la factura muestra el 1 %. El agente ejecuta en treinta segundos; la persona que firma audita y corrige durante una hora y cuarto. El costo humano del informe es 75 veces el costo del token. Una ejecución más rápida no garantiza un ciclo más corto.

Lo que eso significa

Los tokens siguen siendo la unidad básica de consumo, pero ya no alcanzan para saber cuánto cuesta completar una tarea.

En una arquitectura agéntica el costo se distribuye entre llamadas, iteraciones, herramientas, contexto reprocesado y decisiones que el sistema toma sin que nadie las vea. Y después de que el sistema termina, se distribuye entre las personas que verifican, corrigen e integran lo que produjo.

Por eso el análisis tiene que desplazarse tres veces. Del costo de una llamada al costo de una ejecución. Del costo de una ejecución al costo de una tarea completada, que incluye los reintentos, las fallas y la revisión humana. Y del costo de una tarea al valor que produce, que es lo único que justifica cualquiera de los anteriores.

Un modelo barato por token puede ser caro por tarea si necesita tres intentos. Un modelo caro por token puede ser barato por tarea si resuelve con menos pasos y menos corrección. Y una tarea barata puede ser una mala decisión si produce menos valor o más riesgo. El modelo más caro no es el más inteligente; es el más caro. Cuál es el más inteligente para una tarea solo se sabe midiendo cuántos pasos necesita para completarla.

La arquitectura ocupa el centro porque fija todo lo demás: cuántas veces se llama al modelo, cuánto contexto procesa, qué herramientas invoca, cuánta variabilidad introduce y cuánta supervisión exige. No toda tarea que un agente puede resolver necesita un agente. La regla, la misma desde 2018, es la arquitectura mínima suficiente.

Y el costo tiene tres líneas, no una. Tokens. Horas humanas de dirección, auditoría y corrección. Y la inversión en que la capacidad de auditar no se atrofie. Las tres se pagan; solo la primera llega en una factura.

Lo que pasó cuando no se hizo

En 2026 esto dejó de ser una advertencia. Uber armó su presupuesto anual con lógica de licencia fija para herramientas que cobran por consumo, premió a sus ingenieros por consumir y descubrió en abril que el presupuesto del año estaba agotado. Un tope de 1.500 dólares por persona llegó dos meses después. Su COO no pudo decir qué funciones útiles había producido el gasto.

Coinbase enfrentó la misma aritmética y respondió con la infraestructura: cambió el modelo por defecto, enrutó por tarea, llevó la caché del 5 % al 60 % y mantuvo la visibilidad del costo. Redujo el gasto casi a la mitad mientras los tokens seguían creciendo.

La diferencia entre las dos no es de tecnología ni de escala. Es la diferencia entre descubrir el costo en la factura y decidirlo antes de que el agente ejecute.

Lo que queda

El costo de la IA agéntica no debería descubrirse en producción. Antes del despliegue, la organización necesita estimar el costo esperado por tarea en sus tres líneas, justificar la arquitectura elegida, fijar límites de ejecución y definir qué recursos y herramientas estarán disponibles. Durante la operación, medir el consumo real, atribuirlo por caso de uso, detectar desvíos y revisar si el gasto guarda relación con el valor.

Gobernar el costo no significa consumir menos. Significa hacer visible el recorrido que produce cada resultado, entender qué decisiones lo explican y asignar recursos en proporción al valor, la complejidad y el riesgo de cada caso. Y significa algo que los presupuestos de tokens no capturan: que la organización siga sabiendo hacer lo que delegó.

La pregunta con la que empezamos ya no es cuánto cuesta el token de un modelo. Es cuánto cuesta alcanzar un resultado, bajo qué arquitectura, con cuánto juicio humano, y qué valor obtenemos a cambio.

En el ecosistema agéntico, la sostenibilidad depende cada vez menos de encontrar el token más barato y cada vez más de diseñar la tarea, la arquitectura, los límites y las personas que van a supervisarla.

CAPÍTULO IX

Implicancias para la alta dirección y las áreas de tecnología

Todo lo anterior deja una consecuencia práctica: la gobernanza económica de la IA agéntica no es un tema de tecnología ni de finanzas por separado. Exige una conversación común entre quien decide, quien construye, quien paga, quien responde jurídicamente, quien protege los datos y quien va a trabajar dentro del flujo cuando el agente esté operando.

La alta dirección define para qué se habilitan agentes, en qué procesos, con qué objetivos, bajo qué riesgo aceptable y con qué presupuesto en sus tres líneas. Las áreas de tecnología traducen esas decisiones en arquitecturas concretas: flujos deterministas, híbridos o agénticos; modelos locales o externos; herramientas habilitadas; límites de ejecución; monitoreo; reglas de escalamiento. Y las áreas que reciben los resultados definen quién los revisa, cuánto tiempo tiene para hacerlo y cómo conserva la capacidad de detectar un error.

El punto de partida no es "qué modelo usamos". Es qué problema queremos resolver, cuál es la arquitectura mínima suficiente para resolverlo y quién va a firmar lo que el sistema produzca.

Que un agente pueda ejecutar una tarea no significa que sea la forma más eficiente, más segura ni más sostenible de resolverla.

Seis preguntas antes de desplegar

Toda iniciativa basada en IA agéntica debería poder responder seis preguntas. Las cinco primeras estaban en la versión de este trabajo de 2024, pero las adaptamos al contexto agéntico; la sexta es la que el ecosistema agéntico obliga a agregar.

01

VALOR

¿Cuál es el valor esperado?

Qué mejora concreta se busca: reducir tiempos, ampliar capacidad, mejorar calidad, disminuir errores, fortalecer un proceso crítico. Sin esto, la factura no se puede evaluar; solo se puede pagar.

02

ARQUITECTURA

¿Cuál es la arquitectura y por qué?

Si alguien puede escribir la regla, determinista. Si hay un paso que exige leer o redactar, híbrido. Solo si el camino no puede fijarse de antemano, agente. La elección de un escalón más alto se funda en una necesidad funcional, no en la disponibilidad de la tecnología.

03

COSTO POR TAREA

¿Cuánto cuesta completar la tarea?

No la llamada: la tarea. Tokens de entrada y salida, contexto, caché, herramientas, infraestructura, reintentos y fallas. Y en las dos facturas: qué cupo consume por suscripción y qué monto acumula por API.

04

LÍMITES

¿Qué límites tiene el sistema?

Pasos máximos, presupuesto por ejecución y por período, herramientas habilitadas, condiciones de detención, puntos de intervención humana. La autonomía es condicionada por defecto; un agente sin límites no es un agente ambicioso, es un agente mal diseñado.

05

ATRIBUCIÓN

¿Cómo se mide y se atribuye el consumo?

Por modelo, herramienta, usuario, área, proyecto y caso de uso, con topes y alertas antes de que el gasto ocurra. Sin atribución, la organización ve una factura. Con atribución, ve una inversión que puede evaluar.

06

SUPERVISIÓN

¿Quién revisa, cuánto tiempo tiene y cómo conserva la capacidad de revisar?

Cuántas horas humanas de auditoría y corrección exige el flujo por cada hora de ejecución automática. Qué pasa con las salidas que nadie revisa. Y qué hace la organización para que quien supervisa siga sabiendo hacer lo que supervisa. Esta es la línea del presupuesto que casi nadie escribe.

Para las áreas de tecnología

La gobernanza económica se incorpora desde el diseño, no después del primer sobregiro. El control de costos forma parte de las decisiones sobre arquitectura, enrutamiento, gestión del contexto, nivel de razonamiento, uso de herramientas, límites de ejecución y trazabilidad.

Los valores por defecto son la decisión más barata de tomar y la más cara de no tomar: el modelo, el esfuerzo de razonamiento y el contexto que cargan las solicitudes cuando nadie elige son el presupuesto real de la organización.

Y el diseño incluye a las personas. Un flujo que produce mil salidas por día para tres revisores no es un flujo eficiente; es un flujo que transfirió su costo a un lugar donde nadie lo mide. Los puntos de intervención humana se diseñan para que sean cognitivamente viables: en volumen, en tiempo y en competencia.

Para la alta dirección

Los presupuestos de tokens y de infraestructura de IA son decisiones financieras, organizacionales y de riesgo. Definen quién puede consumir, para qué, bajo qué condiciones, con qué límites y con qué rendición de cuentas.

Un presupuesto anual fijado con lógica de licencia para una herramienta que cobra por consumo es un presupuesto que ya no existe; el primer semestre de 2026 lo demostró en empresas con miles de ingenieros y equipos financieros sofisticados.

Y hay una posibilidad que suele quedar fuera de la discusión: en algunos casos, la decisión correcta será no usar IA generativa; en otros, usarla sin arquitectura agéntica. Si el costo total de completar la tarea, en sus tres líneas, supera de manera persistente el valor esperado; si una arquitectura más simple resuelve el problema; o si la organización no dispone de las capacidades para monitorear el sistema y de las personas para supervisarlos, corresponde rediseñar, postergar o descartar.

Gobernar económicamente los agentes de IA es determinar dónde su uso genera valor suficiente para justificar los recursos, los riesgos y las capacidades institucionales que exige integrarlos. Y, en el ecosistema agéntico, esas capacidades incluyen la de seguir pensando.

El presupuesto de un agente tiene tres líneas. La organización que gobierna solo la primera está optimizando el 1 % y financiando el resto sin saberlo.

Notas

1 Corvalán, Juan G. y Sánchez Caparrós, Mariana, ¿Cuál es el costo de trabajar con inteligencia artificial generativa?, IALAB, 2024, disponible en <https://www.web.ialab.com.ar/webia/wp-content/uploads/2024/03/%C2%BFCual-es-el-costo-de-trabajar-con-modelos-de-inteligencia-artificial-generativa.pdf> [acceso: 5 de septiembre de 2026].

2 Véanse, a modo de ejemplo, Anthropic, Claude Model Pricing — All Platforms, precios vigentes desde el 12 de mayo de 2026, <https://www-cdn.anthropic.com/files/4zrzovbb/website/5678bc2f5978e5bcd4f1fe7c14b2c72284dcf9f8.pdf>; OpenAI, Codex rate card, <https://help.openai.com/en/articles/20001106-codex-rate-card>; y GitHub, “GitHub Copilot is moving to usage-based billing”, <https://github.blog/news-insights/company-news/github-copilot-is-moving-to-usage-based-billing/> [acceso: 5 de septiembre de 2026].

3 FinOps Foundation, State of FinOps 2026, <https://data.finops.org>; “Token Economics: The Atomic Unit of AI Value”, <https://www.finops.org/insights/token-economics-the-atomic-unit-of-ai-value/>; y “FinOps and Tokenomics: How do they fit together?”, <https://www.finops.org/insights/finops-tokenomics-fit-together/> [acceso: 5 de septiembre de 2026].

4 Anthropic, “The future of AI at work: Introducing Cowork”, <https://www.anthropic.com/webinars/future-of-ai-at-work-introducing-cowork>; OpenAI, “Introducing workspace agents in ChatGPT”, <https://openai.com/index/introducing-workspace-agents-in-chatgpt/>; xAI, “Grok”, <https://x.ai/grok>; Google Developers Blog, “Build with Google Antigravity, our new agentic development platform”, <https://developers.googleblog.com/en/build-with-google-antigravity-our-new-agentic-development-platform/> [acceso 5 de septiembre de 2026].

5 OpenAI, Tokenizer, <https://platform.openai.com/tokenizer> [acceso junio de 2026].

6 OpenAI, Pricing, <https://openai.com/api/pricing> [acceso junio de 2026]; Anthropic, API Pricing, <https://www.anthropic.com/pricing> [acceso junio de 2026].

7 OpenAI, Tokenizer, <https://platform.openai.com/tokenizer> [acceso junio de 2026].

8 Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N. y Tsvetkov, Y. (2023). “Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models”. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9904–9923. <https://aclanthology.org/2023.emnlp-main.614/> [consultado el 3 de julio de 2026].

9 El concepto alude al mayor costo que enfrentan los hablantes de lenguas no anglófonas al utilizar modelos de lenguaje por decisiones de diseño del tokenizador, no a un impuesto en sentido fiscal.

10 Ovcharov, V. (2026). “The Tokenizer Tax Across 25 European Languages: Domain Invariance, Cross-Lingual Few-Shot Effects, and the Ukrainian Penalty” (preprint). <https://arxiv.org/pdf/2605.24718> [consultado el 3 de julio de 2026]. El español, si bien no está entre los casos más penalizados, tampoco queda exento: una medición controlada sobre 25 lenguas europeas y diez modelos fundacionales midió una fertilidad promedio de 1,46 tokens por palabra en español (rango 1,31–1,61 según el modelo), frente a 1,23 en inglés, es decir, entre un 6 % y un 31 % más tokens para transmitir el mismo contenido, según el modelo utilizado.

11 Anthropic, “Ventanas de contexto”, <https://platform.claude.com/docs/es/build-with-claude/context-windows> [acceso junio de 2026].

12 Además, una ventana más amplia no siempre tiene el mismo precio. Algunos proveedores aplican regímenes tarifarios diferenciados cuando la solicitud supera determinado umbral de tokens. En OpenAI, por ejemplo, las solicitudes de GPT-5.5 que exceden los 272.000 tokens de entrada ingresan en un esquema de long context pricing, en el que los tokens de entrada se facturan al doble y los tokens de salida a 1,5 veces el precio base para la sesión completa.

13 Corvalán, Juan G.; Sánchez Caparrós, Mariana; Martín, Carolina M., Gobernar la IA agéntica, La Ley, 2026.

14 Corvalán, Juan G.; Sánchez Caparrós, Mariana; Martín, Carolina M., *Gobernar la IA agéntica*, La Ley, 2026.

15 IALAB, *Costos IA*, presentación de trabajo, junio de 2026. El escenario considera un expediente de 50 páginas, equivalente a 25.000 tokens de entrada, y una respuesta de 3.000 tokens de salida. Tarifas relevadas al 24 de junio de 2026.

16 Anthropic, "How we built our multi-agent research system", 13 de junio de 2025, <https://www.anthropic.com/engineering/multi-agent-research-system> [consulta: junio de 2026].

17 Gartner, "Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90 Percent Less Than in 2025", 25 de marzo de 2026, <https://www.gartner.com/en/newsroom/press-releases/2026-03-25-gartner-predicts-that-by-2030-performing-inference-on-an-llm-with-1-trillion-parameters-will-cost-genai-providers-over-90-percent-less-than-in-2025> [consulta: junio de 2026].

18 Anthropic, *Claude Platform Pricing*, sección "Claude Managed Agents pricing", <https://platform.claude.com/docs/en/about-claude/pricing> [consulta: 5 de septiembre de 2026].

19 Véanse, a modo de ejemplo, OpenAI, *API Pricing*, <https://developers.openai.com/api/docs/pricing>; Anthropic, *Claude Platform Pricing*, <https://platform.claude.com/docs/en/about-claude/pricing>; Google, *Gemini Developer API Pricing*, <https://ai.google.dev/gemini-api/docs/pricing>; y xAI, *API Pricing*, <https://docs.x.ai/developers/pricing> [consulta: 5 de septiembre de 2026].

20 xAI, "Tool Usage Details", <https://docs.x.ai/developers/tools/tool-usage-details> [consulta: 5 de septiembre de 2026].

21 Nota sobre las cifras. Las tarifas por token corresponden al relevamiento de IALAB al 24 de junio de 2026. Las condiciones de los planes de suscripción y de los créditos de Anthropic y OpenAI fueron verificadas en la documentación oficial de ambos proveedores al 5 de septiembre de 2026. Los proveedores publican los cupos como rangos y los modifican con frecuencia; los valores deben leerse como órdenes de magnitud vigentes a esas fechas.

22 Fuente: documentación de precios de OpenAI y Anthropic, verificada al 5 de septiembre de 2026. Tarifas verificadas el 5 de septiembre de 2026 en fuentes que reproducen la documentación de precios de OpenAI (developers.openai.com) y Anthropic (platform.claude.com/docs/en/about-claude/pricing). Incluyen los recargos por contexto largo, caché, lote, procesamiento regional y búsqueda web citados en el capítulo. Los recálculos del informe usan las tarifas de Claude Opus 4.8 salvo indicación en contrario y son estimaciones redondeadas. GPT-5.6 Sol, sucesor de GPT-5.5, bajó a 4 y 20 dólares por millón el 22 de agosto de 2026 como precio promocional.

23 Umbral y tarifa de contexto largo de OpenAI según su documentación de precios al 5 de septiembre de 2026. Sobre el recargo por contexto largo en los modelos de Anthropic las fuentes consultadas no coinciden; el ejemplo del capítulo se mantiene por debajo de cualquier umbral, por lo que la diferencia no afecta los cálculos.

24 Estimaciones redondeadas con las tarifas de Claude Opus 4.8 (y de GPT-5.5 para el umbral de contexto largo) al 5 de septiembre de 2026.

25 Artificial Analysis, *Intelligence Index*, metodología de la métrica "cost to run": artificialanalysis.ai/methodology.

26 Janakiram MSV, "Why Your Engineers' Favorite AI Tools Are Wrecking Your 2026 Budget", *Forbes*, 26 de mayo de 2026.

27 Janakiram MSV, "Uber Burns Its 2026 AI Budget In Four Months On Claude Code", *Forbes*, 17 de mayo de 2026, a partir de la entrevista del CTO con *The Information*.

28 "Why Uber has Already Burned Through its AI Budget", *AI Magazine*, 23 de abril de 2026, con base en el reporte de *The Information*.

29 Janakiram MSV, "Uber Burns Its 2026 AI Budget In Four Months On Claude Code", *Forbes*, 17 de mayo de 2026, a partir de la entrevista del CTO con *The Information*.

30 "Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it", *Fortune*, 26 de mayo de 2026, a partir de la entrevista de Andrew Macdonald en el podcast *Rapid Response*.

31 Lucia Auerbach, "Uber Blew Through Its 2026 AI Budget in 4 Months. Now It's Capping Employee Use", *Inc.*, 4 de junio de 2026, con base en el reporte de *Bloomberg*.

32 Sasha Rogelberg, "After blowing through its entire 2026 AI budget in months, Uber CTO says, 'We're coming to the end of the so-called tokenmaxxing era'", Fortune, 7 de agosto de 2026.

33 "The token bill comes due: Inside the industry scramble to manage AI's runaway costs", TechCrunch, 5 de junio de 2026.

34 Paul Roetzer, citado en "Uber, Microsoft, and Others Burning Through AI Budgets. Now What?", SmarterX, 2 de junio de 2026.

35 Sasha Rogelberg, "After blowing through its entire 2026 AI budget in months, Uber CTO says, 'We're coming to the end of the so-called tokenmaxxing era'", Fortune, 7 de agosto de 2026.

36 Estudio de Mavvrik y Benchmarkit (2025) sobre 372 empresas, citado en Forbes, 26 de mayo de 2026.

37 Corvalán, Juan G.; Sánchez Caparrós, Mariana y Martín, Carolina M., *Gobernar la IA agéntica*, UBA IALAB, 2026, Introducción, "Tercero": las automatizaciones previas a la IA generativa "no desaparecen con la llegada de la IA agéntica: conviven con ella y también la complementan".

38 *Gobernar la IA agéntica*, cit., sección 2, "Automatización sin IA": la metáfora del reloj mecánico y los tres rasgos (ausencia de inferencia, cierre operativo, determinismo).

39 Corvalán, Juan G., *Inteligencia artificial para transformar organizaciones públicas*, Astrea / Editorial Universidad del Rosario / IMODEV / DPI, 1ª ed., 2019 (texto ampliado de la conferencia ante el Consejo Permanente de la OEA, 22 de agosto de 2018), caps. 5 y 6: la función predictiva, entrenada con más de 2.400 sentencias y 1.400 dictámenes previos y un 96 % de acierto promedio, y los árboles de decisión que automatizan la generación de documentos, como el del control de autosuficiencia. El componente predictivo combinaba modelado de tópicos, bolsa de palabras y expresiones regulares.

40 *Ibíd.*, cap. 4. Mediciones sobre seis horas de trabajo por día y veintidós días por mes.

41 *Gobernar la IA agéntica*, cit., sección 2, "Agentes de IA: la IA proactiva por diseño": "La diferencia puede condensarse en una pregunta operativa: ¿quién diseña el camino?".

42 Gartner, estimación de 2026 citada en el capítulo III, apartado 2.

43 *Gobernar la IA agéntica*, cit., Introducción y sección 7, "El agent reliability gap como cautela estructural", con referencia a los benchmarks GAIA, SWE-bench Verified, AgentBench y WebArena a abril de 2026.

44 Corvalán, Juan G.; Sánchez Caparrós, Mariana y Papini, Carina M., *Organizaciones centauro. Playbook 1: IA agéntica*, La Ley, 2025, sección 6: "no se puede concebir una autonomía ilimitada por diseño. La autonomía condicionada o limitada es la clave". El mismo trabajo registra un dato que anticipa el problema de este libro: el Agente de ChatGPT, en julio de 2025, descontaba un crédito por cada interacción del usuario y no por tarea completada.

45 Corvalán, Juan G. y Sánchez Caparrós, Mariana, *Agentes de inteligencia artificial y workflows agénticos: la nueva frontera de la automatización*, La Ley / UBA IALAB, 2025, cap. 2, "¿Cómo decido si un agente es la solución más adecuada para mi problema?". En el mismo sentido, Anthropic, "Building effective agents", 19 de diciembre de 2024.

46 Corvalán, *Inteligencia artificial para transformar organizaciones públicas*, cit., cap. 6, "Gobernanza de datos, agentes conversacionales y matices de automatización".

47 Corvalán, Juan G.; Sa Zeichen, Gustavo y Albertsen, Lihué M., "Actividad administrativa automatizada", en *Tratado de Inteligencia Artificial y Derecho*, tomo IV, La Ley, 2021, con referencia al estándar SAE J3016.

48 Corvalán y Sánchez Caparrós, *Agentes de inteligencia artificial y workflows agénticos*, cit., cap. 3, "Cómo clasificar los agentes basados en IA generativa según su grado de autonomía: una propuesta en cinco niveles".

49 *Gobernar la IA agéntica*, cit., sección 2, "Del sistema que ejecuta al sistema que decide: tres niveles de automatización".

50 Estimaciones con las tarifas de Gemini 3.5 Flash (1,50 y 9 dólares por millón de tokens de entrada y salida) y Claude Opus 4.8 (5 y 25), verificadas al 5 de septiembre de 2026 y coincidentes con el relevamiento de junio usado en el capítulo III. El rango del agente refleja recorridos de entre tres y quince llamadas por contrato. *Gobernar la IA agéntica*, cit., cierre de la sección 6.

51 Documentación de precios de Google (ai.google.dev) y Anthropic, según fuentes que la reproducen, verificadas el 5 de septiembre de 2026. Gemini 3.5 Flash se lanzó el 19 de mayo de 2026 a 1,50 y 9 dólares por millón y mantenía esa tarifa a fines de agosto; sus sucesores (Gemini 3.6 y 3.7 Flash) se ofrecen a 0,75 y 3,75 dólares con precio introductorio hasta el 31 de diciembre de 2026. Con la tarifa de 3.5 Flash, la llamada única del capítulo III cuesta 0,065 dólares y la ejecución agéntica completa 0,31, las mismas cifras del relevamiento de junio.

52 Precios de alquiler de GPU Nvidia H100 verificados el 5 de septiembre de 2026: mediana de 3,32 dólares por GPU-hora entre cuarenta proveedores y mínimo verificado de 1,73 (GetDeploying, índice diario); rango de 2,89 a 11 dólares según proveedor y mediana de 4,17 en nubes especializadas frente a 7,89 en los grandes proveedores de nube (Thunder Compute, septiembre de 2026). Precio de compra de una H100 de 80 GB entre 30.000 y 40.000 dólares (GMI Cloud, abril de 2026; CloudZero, agosto de 2026). Los costos mensuales del texto suponen operación continua de una GPU y no incluyen servidor, energía ni personal.

53 Qwen3.8-27B, publicado por Alibaba el 14 de agosto de 2026 bajo licencia Apache 2.0, cuantiza a 14–17 GB y corre en una GPU de 24 GB; DeepSeek V4-Pro (0813), licencia MIT, requiere aproximadamente ocho GPU H100/H200 de 80 GB para pesos en 4 bits (Pinggy, agosto de 2026; comparativas de modelos de pesos abiertos a septiembre de 2026). El costo mensual de ocho GPU se estima con la mediana de alquiler de la nota anterior.

54 Corvalán, Sánchez Caparrós y Martín, *Gobernar la IA agéntica*, cit., sección 4, "Capa 1: infraestructura de despliegue" y "Tres grados de soberanía: una taxonomía operativa": soberanía total, de procesamiento y contractual. La advertencia sobre el despliegue on-premise con API externa es textual de ese trabajo.

55 OpenAI documenta la compactación del historial como guía de la API (Compaction, developers.openai.com); las herramientas agénticas de Anthropic y OpenAI compactan automáticamente el contexto al acercarse al límite. Headroom es un proyecto de código abierto orientado a reducir y administrar el contexto en cargas agénticas; sus cifras de ahorro corresponden a sus propios benchmarks.

56 OpenAI y Anthropic permiten configurar el nivel de esfuerzo de razonamiento por solicitud (reasoning effort; adaptive thinking en Claude Opus 4.8) y los tokens de razonamiento se facturan como tokens de salida. Lo mismo ocurre en Gemini, donde los tokens de pensamiento se cobran a tarifa de salida (CloudZero, septiembre de 2026). Documentación de precios y de razonamiento de los tres proveedores al 5 de septiembre de 2026.

57 Brian Armstrong, publicación en X, 27 de junio de 2026: "How to keep AI spend flat while token usage grows exponentially: Not with friction and spend alerts. With better defaults, routing, and caching". Cobertura en Business Insider, The Decoder y Yahoo Finance, 27 a 30 de junio de 2026. Los modelos adoptados por defecto fueron GLM 5.2 (Z.ai) y Kimi K2.7 (Moonshot AI).

58 Estimaciones sobre las tarifas verificadas al 5 de septiembre de 2026.

59 Las consideraciones de licencia, residencia y cumplimiento de los modelos abiertos de origen chino en industrias reguladas y la ausencia de datos sobre degradación de calidad en casos límite fueron señaladas en la cobertura de Let's Data Science y TechTimes, 28 y 29 de junio de 2026.

60 Sasha Rogelberg, "After blowing through its entire 2026 AI budget in months, Uber CTO says, 'We're coming to the end of the so-called tokenmaxxing era'", *Fortune*, 7 de agosto de 2026, citada en el capítulo IV.

61 Corvalán, Juan G. (con la colaboración de Mariana Sánchez Caparrós y Alejandro Melamed), *Sedentarismo cognitivo. Productividad, agentes de IA y los riesgos de delegar el pensamiento*, La Ley, 2026, cap. 11, "La métrica decisiva: redefinir la productividad en organizaciones centauro", con el ejemplo de los dictámenes administrativos: ejecución de treinta segundos, auditoría de cuarenta y cinco minutos, corrección de treinta, frente a dos horas de una persona sin IA.

62 Amdahl, G. M. (1967), "Validity of the single processor approach to achieving large scale computing capabilities", *AFIPS Conference Proceedings*, 30, 483-485. Aplicación al trabajo intelectual en *Sedentarismo cognitivo*, cit., cap. 10.

63 *Sedentarismo cognitivo*, cit., cap. 8, "Tres niveles de delegación: descarga, tercerización y rendición cognitiva". La descarga cognitiva (cognitive offloading) sigue a Risko, E. F. y Gilbert, S. J. (2016), *Trends in Cognitive Sciences*, 20(9), 676-688.

64 Shaw, S. D. y Nave, G. (2025), *Thinking—Fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender*, The Wharton School, University of Pennsylvania, <https://dx.doi.org/10.2139/ssrn.6097646>.

65 Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D. y Jurafsky, D. (2026), "Sycophantic AI decreases prosocial intentions and promotes dependence", *Science*, 391, eaec8352.

66 Carro, M. V. (2024), *Flattering to deceive: The impact of sycophantic behavior on user trust in large language models*, UBA IALAB, arXiv:2412.02802, presentado en NeurIPS 2024.

67 Bainbridge, L. (1983), "Ironies of automation", *Automatica*, 19(6), 775-779. Reformulación en el contexto agéntico en *Sedentarismo cognitivo*, cit., cap. 6.

68 Niederhoffer, K.; Rosen Kellerman, G.; Lee, A.; Liebscher, A.; Rapuano, K. y Hancock, J. T., "AI-generated 'workslop' is destroying productivity", *Harvard Business Review*, 22 de septiembre de 2025 (BetterUp y Stanford Social Media Lab, encuesta a 1.150 trabajadores de tiempo completo en Estados Unidos).

69 *Sedentarismo cognitivo*, cit., cap. 6, "Segunda paradoja: la descualificación (deskilling)", con referencia a Braverman, H. (1974), *Labor and Monopoly Capital*, Monthly Review Press.

70 *Sedentarismo cognitivo*, cit., cap. 6, "Tercera paradoja: el monitoreo y la complacencia automatizada".

71 Woodruff, A.; Shelby, R.; Gage Kelley, P.; Rousso-Schindler, S.; Smith-Loud, J. y Wilcox, L. (2024), "How knowledge workers think generative AI will (not) transform their industries", *CHI '24*, ACM, <https://dl.acm.org/doi/10.1145/3613904.3642700>.

72 Bedard, J. et al. (Boston Consulting Group), *Harvard Business Review*, marzo de 2026, encuesta a 1.488 trabajadores de tiempo completo en Estados Unidos, citado en *Sedentarismo cognitivo*, cit., cap. 6, "Cuarta paradoja: el cerebro frito (AI brain fry)".

73 *Sedentarismo cognitivo*, cit., cap. 10, "El cuello de botella cognitivo: la ley de Amdahl aplicada al trabajo intelectual", con los tres escenarios (IA de las tareas, IA de los procesos, sistemas multiagente autónomos) y el ejemplo de la aseguradora.

74 Sobre los "gimnasios cognitivos", *Sedentarismo cognitivo*, cit., cap. 9; sobre andamiaje y muleta, *ibíd.* La "deuda psicológica" de la adopción no estructurada de IA, en sus seis formas (cognitiva, de autonomía, de competencia, de relación, de credibilidad y de identidad), en Champniss, G., "The psychological costs of adopting AI", *Harvard Business Review*, 1 de mayo de 2026.

75 Fases según Corvalán, *Sedentarismo cognitivo*, cap. 11; tiempos del ejemplo del dictamen administrativo de ese trabajo, aplicados al informe del capítulo III.



IALAB

AI academy
by doinGlobal

LA LEY

 **Thomson
Reuters™**