



NARSIL

INTELLIGENCE

ESTUDIOS NARSIL · 06

Cuánto cuesta de verdad la IA en una pyme

Suscripciones, pago por uso o un modelo propio: el coste a tres años para una empresa de 20 personas, con precios de octubre de 2026 y los supuestos a la vista

OCTUBRE DE 2026

Resumen

1. **Hay tres formas de pagar la IA:** una cuota por usuario, el pago por uso a través de API y un modelo propio en un equipo de la empresa.
2. **Las cuotas rondan los 20 dólares por persona y mes.** ChatGPT Business y Claude Team cuestan 20 \$ en pago anual, sin impuestos; Microsoft 365 Copilot, 15,60 € en oferta; Gemini viene incluido en Google Workspace desde el plan Standard.
3. **Para 20 personas y tres años,** nuestras estimaciones dan unos 14.400 \$ en cuotas, unos 5.400 \$ por API con uso moderado (más el coste de construir la aplicación) y unos 6.000 € en un equipo propio con su electricidad (más su puesta en marcha y mantenimiento).
4. **La API es la más barata con poco uso,** pero su coste crece con cada documento largo. **El equipo propio** cuesta lo mismo se use poco o mucho, y además resuelve la privacidad.
5. **Los precios se mueven:** el hardware ha subido con la crisis de la memoria, algunos precios de API suben en 2027 y parte de las cuotas son ofertas.
6. **La opción más sensata suele ser híbrida:** cuotas de empresa para el trabajo general, modelo local para la información crítica y API solo donde haga falta una aplicación propia.

01

Tres formas de pagar la IA

Cuando una pequeña empresa decide usar inteligencia artificial de forma seria, tiene tres caminos. El primero es **pagar una cuota por usuario** a un proveedor, como ChatGPT Business, Claude Team, Gemini en Google Workspace o Microsoft 365 Copilot: cada persona tiene su asistente y el proveedor se ocupa de todo. El segundo es el **pago por uso a través de API**: se paga por la cantidad de texto que procesa el modelo, y se construye encima una aplicación propia. El tercero es un **modelo propio**: un modelo abierto ejecutado en un equipo de la empresa, que se paga una vez y se mantiene.

Este estudio compara los tres con precios publicados a 2 de octubre de 2026, para una empresa española de 20 personas y un horizonte de tres años. Las cifras finales son estimaciones propias con los supuestos a la vista, para que cada organización pueda rehacerlas con los suyos.

02

Cuotas por usuario

Plan	Precio por usuario y mes	Observaciones
ChatGPT Business	20 \$ anual · 25 \$ mensual	Sin impuestos. Asiento Premium: 100 \$ y 125 \$.
Claude Team	20 \$ anual · 25 \$ mensual	Sin impuestos. Asiento Premium: 100 \$ y 125 \$. Enterprise: 20 \$ por asiento más el uso a precio de API.
Google Workspace con Gemini	Starter 6,80 € · Standard 13,60 € · Plus 21,10 €	Precio mensual flexible. Gemini completo desde Standard; en Starter, solo en Gmail y de forma básica.
Microsoft 365 Copilot	15,60 € (oferta, antes 18,20 €)	Sin IVA, pago anual, complemento de una suscripción de Microsoft 365. Copilot Chat viene sin coste adicional con los planes que cumplen los requisitos.

Páginas de precios oficiales consultadas el 2 de octubre de 2026. OpenAI y Anthropic publican en dólares; Google y Microsoft, en euros para España.

Dos detalles cambian mucho la cuenta. Si la empresa ya paga Google Workspace Standard o superior, Gemini no tiene coste adicional. Y si ya paga Microsoft 365, el complemento de Copilot se suma a esa cuota. Para una empresa de 20 personas sin ninguna de las dos, ChatGPT Business o Claude Team suponen 20 \$ × 20 usuarios × 36 meses: **unos 14.400 \$ en tres años**, sin impuestos.

La ventaja de las cuotas es la sencillez: precio previsible, sin técnicos, con aplicaciones pulidas y contrato de encargo de tratamiento incluido. El inconveniente es que la información sale de la empresa, con las condiciones que analizamos en el estudio 01, y que el coste crece con cada persona que se suma, la use mucho o poco.

03

Pago por uso a través de API

Las API cobran por millón de **tokens**, los fragmentos de texto con los que trabajan los modelos. Según Anthropic, en inglés un token equivale aproximadamente a cuatro caracteres o tres cuartos de palabra; en castellano suelen hacer falta algunos más. Se cobra por separado el texto que se envía (entrada) y el que genera el modelo (salida), que es más caro.

Proveedor y gama	Entrada	Salida
OpenAI, gama alta (gpt-6-astra)	10 \$	50 \$
OpenAI, gama media	2 \$	10 \$
OpenAI, gama baja (gpt-6-luna)	0,10 \$	0,50 \$
Anthropic, Claude Opus 5.5	4 \$	20 \$
Anthropic, Claude Sonnet 5.5	2 \$	10 \$
Anthropic, Claude Haiku 4.5	1 \$	5 \$
Google, Gemini 3.1 Pro (versión preliminar)	2 \$	12 \$
Google, Gemini Flash	0,75 \$ (1,50 \$ desde enero de 2027)	3,75 \$ (7,50 \$ desde enero de 2027)
Google, Gemini 2.5 Flash-Lite	0,10 \$	0,40 \$

Dólares por millón de tokens, sin impuestos, tarifa estándar. Páginas de precios oficiales consultadas el 2 de octubre de 2026.

Una estimación de uso

Supongamos que cada una de las 20 personas hace **40 consultas al día**, con 2.000 tokens de entrada (una petición con algo de contexto) y 500 de salida, durante 21 días laborables al mes. Son unos 1,7 millones de tokens de entrada y 0,4 millones de salida por persona y mes. Con un modelo de gama media a 2 \$ y 10 \$ por millón, salen **unos 7,6 \$ por persona y mes**: unos 150 \$ al mes para toda la empresa y **unos 5.400 \$ en tres años**. Con un modelo de gama alta como Claude Opus 5.5, el doble.

La cifra es atractiva, pero tiene dos condiciones. La primera: la API no es una aplicación, y alguien tiene que construir y mantener la herramienta que la usa, lo que cuesta bastante más que el consumo de tokens. La segunda: el coste depende mucho del contexto. Si cada consulta incluye documentos largos (un contrato, un expediente, el resultado de una búsqueda documental), la entrada puede multiplicarse por diez y la factura con ella.

04

Un modelo propio

La tercera opción es comprar un equipo y ejecutar en él un modelo abierto, como explicamos en el estudio 03. Un equipo con chip NVIDIA GB10 y 128 GB de memoria, capaz de mover modelos como gpt-oss-120b, cuesta en España unos 6.050 € con IVA, **unos 5.000 € sin IVA**. Repartido en tres años entre 20 personas, son unos 7 € por persona y mes.

A eso se suma la **electricidad**. Estos equipos usan una fuente de 240 W. Con un consumo medio de 150 W, encendido día y noche, son unos 1.300 kWh al año. Al precio medio de la electricidad para pequeños consumidores no domésticos en España, unos 0,24 € por kWh sin IVA según Eurostat para el segundo semestre de 2025, **unos 320 € al año**, menos de 1.000 € en tres años.

Y está el **trabajo**: la puesta en marcha, la interfaz para los usuarios, las actualizaciones, la seguridad y el mantenimiento. No lo cuantificamos porque depende de si se hace con personal propio o se contrata, pero es la partida que más varía y la que no hay que olvidar. Un equipo de este tipo sirve bien a un grupo de personas con un uso moderado; si muchas lo usan intensamente a la vez, hará falta más capacidad.

Su gran ventaja no es solo el precio: la información no sale de la empresa y el coste no crece con el uso. Su riesgo, en 2026, es el precio del hardware, que se ha disparado con la crisis de la memoria.

05

La comparación a tres años

Opción	Coste estimado en tres años	Qué incluye y qué no
Cuotas (ChatGPT Business o Claude Team)	≈ 14.400 \$	Todo el servicio. No incluye impuestos.
API, modelo de gama media, uso moderado	≈ 5.400 \$	Solo el consumo. No incluye construir ni mantener la aplicación.
Modelo propio (equipo GB10 de 128 GB)	≈ 6.000 €	Equipo y electricidad. No incluye puesta en marcha ni mantenimiento.

Estimaciones propias para 20 personas, sin impuestos, con los supuestos de los apartados 02 a 04. Cifras en la moneda en que publica cada proveedor.

Leída con cuidado, la tabla dice algo más útil que «cuál es más barata». Las cuotas son la opción más cara en dinero, pero la más barata en esfuerzo. La API es la más barata en consumo, pero exige desarrollo y su coste crece con el uso. El modelo propio tiene un coste casi fijo: compensa más cuanto más se usa, y además es la única opción en la que los datos no salen.

A gran escala, la ventaja del modelo propio crece. Un análisis de Lenovo de julio de 2026 estimaba que alojar los modelos en servidores propios resultaba unas seis veces más barato por millón de tokens que hacerlo en la nube, y hasta diecisiete veces más que las API de los modelos de frontera, aunque para servidores mucho mayores que los de una pyme y con el sesgo evidente de un fabricante de hardware.

06

Lo que no sale en la factura

El precio no es la única variable. La **privacidad** cambia de una opción a otra, como analizamos en el estudio 01: con cuotas y API, los datos salen de la empresa, aunque con contrato. La **dependencia** también: un proveedor puede cambiar precios, condiciones o modelos, y migrar cuesta. Y los **precios se mueven**: el DGX Spark de 128 GB cuesta hoy un 74 % más que en su lanzamiento, Google duplicará el precio por API de Gemini Flash en enero de 2027 y el precio de Copilot que publica Microsoft para España es una oferta. Cualquier cálculo a tres años tiene que prever esos cambios.

07

Nuestra propuesta: híbrido y por criticidad

Para la mayoría de las pymes, la respuesta no es elegir una sola opción. Lo sensato suele ser combinar: **cuotas de empresa** para el trabajo general de quienes más usan la IA, un **modelo local** para la información que no puede salir (datos de clientes, expedientes, secretos de la empresa) y la **API** solo donde una aplicación propia lo justifique, por ejemplo un asistente sobre la documentación interna.

Antes de decidir, conviene medir cuánto se usa hoy la IA y para qué, separar la información crítica de la general, calcular con el uso real y no con el máximo imaginable, e incluir siempre las horas de puesta en marcha y mantenimiento. Con esos cuatro datos, la decisión deja de ser una cuestión de opinión.

Anexo

Fuentes

Páginas consultadas el 2 de octubre de 2026, salvo las noticias, fechadas en su publicación.

1. OpenAI. *ChatGPT Business pricing*. openai.com/business/pricing
2. OpenAI. *API pricing*. developers.openai.com/api/docs/pricing
3. Anthropic. *Claude pricing*. claude.com/pricing; *API pricing*. platform.claude.com/docs/en/about-claude/pricing
4. Google. *Google Workspace pricing*. workspace.google.com/pricing; *Gemini API pricing*. ai.google.dev/gemini-api/docs/pricing
5. Microsoft. *Precios de Microsoft 365 Copilot*. microsoft.com/es-es/microsoft-365/copilot/pricing
6. Eurostat. *Electricity prices for non-household consumers (nrg_pc_205)*, España, 2025-S2, consumo inferior a 20 MWh, sin IVA.
7. NVIDIA. DGX Spark, especificaciones. nvidia.com; Alternate España, ASUS Ascent GX10 (consultado el 2-10-2026).
8. The Register. *Nvidia debuts \$4,999 DGX Spark with half the RAM and storage amid memory crunch* (2-10-2026).
9. Lenovo Press. *On-Premise vs Cloud: Generative AI Total Cost of Ownership (2026 Edition)* (julio de 2026). lenovopress.lenovo.com/lp2368
10. NARSIL Intelligence. *Estudios 01 y 03*. narsilintelligence.com/estudios



NARSIL INTELLIGENCE

Diseñamos, construimos e implantamos sistemas de inteligencia artificial en la infraestructura de cada organización, y formamos a los equipos que los usan.

NARSILINTELLIGENCE.COM