Usar Claude y ChatGPT de forma eficiente: una guía con base documental

Vigente a 18 de julio de 2026 · Helena Debain · luxellency.com

Texto de divulgación pensado para adaptarse a distintos formatos (presentación, ficha, artículo, material de aula) y dirigido tanto a alumnado como a profesorado. Todas las afirmaciones técnicas sobre Claude están contrastadas con documentación oficial de Anthropic, y las referidas a ChatGPT con documentación oficial de OpenAI, ambas vigentes a 18 de julio de 2026. Esta versión actualiza la revisión anterior, cerrada el 10 de julio de 2026, para incorporar dos bloques nuevos (una lectura equivalente de la eficiencia en ChatGPT y el coste medioambiental de la IA generativa, con el informe de la ONU del 3 de junio de 2026 como referencia central) y una nota verificada sobre la suspensión gubernamental de Claude Fable 5 en junio de 2026.[1]

Por qué este texto

Buena parte del contenido que circula sobre "cómo usar Claude de forma eficiente" mezcla productos distintos, repite cifras sin fuente y presenta como dato lo que son estimaciones o pruebas anecdóticas.

El objetivo de este material no es solo enseñar a usar Claude mejor, sino entrenar el criterio para distinguir información verificable de lo que no lo es.

Las funcionalidades cambian cada pocas semanas, pero el hábito de comprobar en fuentes oficiales es estable y transferible a cualquier herramienta.

A lo largo del texto se señala, cuando procede, la diferencia entre lo documentado oficialmente y lo que circula sin respaldo.

El texto está pensado también como ejemplo de lectura crítica: cualquier afirmación concreta debería poder rastrearse hasta docs.anthropic.com o support.claude.com.[2][3]

El coste medioambiental de la IA generativa

Este es el bloque que quedó pendiente en la entrega anterior de la serie. El punto de partida no es una cifra por consulta, sino la fotografía de conjunto que ofrece, por primera vez con esta profundidad, un informe de Naciones Unidas.

El informe de referencia: UNU-INWEH, 3 de junio de 2026

El Instituto de la Universidad de las Naciones Unidas para el Agua, el Medio Ambiente y la Salud (UNU-INWEH) publicó el 3 de junio de 2026 el informe "Environmental Cost of AI's Energy Use", el primero que cuantifica de forma conjunta las huellas de carbono, agua y suelo asociadas a la electricidad que mueve la inteligencia artificial.[26]

Sus cifras centrales: en 2025, los centros de datos consumieron 448 teravatios hora de electricidad, una cantidad que los situaría como el undécimo país del mundo por consumo eléctrico si fueran uno solo. Para 2030, la proyección sube a 945 teravatios hora, casi el triple del consumo combinado de Pakistán, Bangladés y Nigeria. La huella hídrica proyectada para 2030 es de 9,3 billones de litros anuales, equivalente a las necesidades básicas de agua de toda la población del África subsahariana, y la superficie de suelo ocupada por esta infraestructura superaría los 14.500 kilómetros cuadrados.[26][27]

> **La escala, en tres cifras**
>
> 448 teravatios hora consumidos por los centros de datos en 2025: el undécimo país del mundo por consumo eléctrico si fueran uno solo. 945 teravatios hora proyectados para 2030. 9,3 billones de litros de agua al año. Fuente: UNU-INWEH, 3 de junio de 2026.

El informe recomienda que las empresas de IA publiquen de forma estandarizada su huella de carbono, agua y suelo, precisamente porque hoy esa información depende de lo que cada compañía decide revelar de forma voluntaria.[27]

Lo que se sabe por consulta individual

La cifra agregada del informe de la ONU describe la infraestructura global, no lo que consume una consulta suelta. Para eso, las únicas cifras verificables son las que cada compañía ha publicado por su cuenta:

- OpenAI declara, en una entrada de su propio blog, un consumo medio de aproximadamente 0,34 vatios hora por consulta de ChatGPT. Es una cifra propia de la compañía, sin revisión por pares independiente.[28]
- Google, en su informe técnico y en el blog de Google Cloud, mide una mediana de 0,24 vatios hora, 0,03 gramos de CO2 equivalente y 0,26 mililitros de agua por interacción de texto con Gemini, con una reducción del consumo energético de 33 veces en doce meses gracias a mejoras de eficiencia.[29]
- Anthropic no ha publicado, a fecha de este documento, una cifra equivalente de energía o agua por consulta para Claude. La compañía declara trabajar con infraestructura de Google Cloud orientada a energía limpia, sin cifra propia de Scope 1, 2 o 3.[1]

Una investigación independiente encargada para verificar estos mismos datos, realizada el 18 de julio de 2026, confirma las tres cifras anteriores y su procedencia, y señala expresamente que la ausencia de cifra propia de Anthropic sigue vigente a esa fecha. La misma investigación advierte de que conviene distinguir siempre entre la declaración oficial de una compañía (sin revisión por pares independiente) y el informe de un organismo intergubernamental como la ONU, que sí pasa por un proceso de revisión institucional propio.[31]

Conviene etiquetar siempre el tipo de cada cifra: la de OpenAI y la de Google son declaraciones propias de empresa, sin revisión por pares independiente; la del informe de la ONU procede de un organismo intergubernamental con proceso de revisión institucional. Las tres circulan ampliamente en cobertura periodística secundaria, a veces sin que el artículo que las cita enlace de vuelta a la fuente primaria; por eso este documento remite siempre a la declaración o al informe original, no a la cobertura que los resume.[31]

Por qué esto no es un mensaje de alarma, sino de criterio

Cada consulta individual es minúscula frente a la escala de un país o de un sector industrial. El problema no está en la culpa de la consulta suelta, está en un mecanismo bien documentado: el efecto rebote.

Cuanto más barata y rápida se vuelve cada consulta gracias a las mejoras de eficiencia, más se usa la herramienta en conjunto, y el consumo total de la industria sigue subiendo aunque el coste por consulta

baje. Es la misma dinámica que ha hecho que los datos de Google mejoren individualmente mientras el informe de la ONU proyecta un aumento global hasta 2030.[26][29]

La consecuencia práctica para quien usa Claude o ChatGPT a diario no es dejar de usarlos, es aplicar el mismo criterio de tamaño de herramienta que se describe más adelante en este documento para el ahorro de cuota: reservar los modelos de mayor capacidad y los niveles de esfuerzo más altos para las tareas que de verdad los necesitan, y usar los modelos más ligeros para todo lo demás. Ahorrar iteraciones y elegir bien el modelo no es solo una cuestión de factura mensual, es también la palanca individual real frente a un consumo agregado que, de otro modo, queda completamente fuera del control de quien hace la pregunta.

La distinción que lo ordena todo: interfaz, API y Claude Code

El error más frecuente en el contenido divulgativo consiste en tratar la interfaz de claude.ai como si fuera la API.

Son productos distintos, con lógicas y controles distintos, y las técnicas de optimización de uno no se trasladan automáticamente al otro.[4][2]

Conviene fijar tres superficies desde el principio:

Interfaz de chat (claude.ai, app de escritorio y móvil). Es lo que usa la mayoría de las personas: se interactúa escribiendo mensajes, el usuario trabaja con límites de uso visibles como barras de progreso y no ve tokens.[3][2]

API de Claude. Es la vía técnica para integrar Claude en aplicaciones mediante código; aquí sí se manipulan parámetros como el caché de prompts, las cabeceras beta o los presupuestos de razonamiento, y se paga por token consumido.[4]

Claude Code. Es el sistema de programación agéntica de Anthropic que vive en el terminal y en extensiones como VS Code, con comandos propios, archivo de configuración y documentación específica.[5][6]

Esta distinción importa porque determina si una recomendación es relevante o inútil.

Cuando una guía explica “cómo ahorrar tokens con prompt caching manual” a usuarios de claude.ai, está enseñando a manejar un mando que esos usuarios no tienen: el caché manual de prompts pertenece a la API y a los SDK; lo mismo ocurre con muchos comandos “/” que aparecen en infografías y que en realidad son nativos de Claude Code, no del chat.[6][4]

Qué significa “eficiencia” para quien usa el chat

Quien usa la interfaz no paga directamente por token; dispone de límites de uso que controlan cuánto puede interactuar con Claude en una ventana de tiempo, visibles en Configuración > Uso en los planes de pago.[3]

El sistema mide internamente en tokens, pero la persona solo ve barras de progreso y avisos de límite. Por eso “ahorrar tokens”, expresión importada de la documentación de API, resulta engañosa en la interfaz.

Para quien usa el chat, lo que se ahorra o se malgasta no son tokens abstractos, sino cosas muy concretas:

- Iteraciones: mensajes de ida y vuelta para corregir, aclarar o rehacer instrucciones.
- Mensajes de clarificación: provocados por consultas ambiguas o mal acotadas.
- Archivos adjuntos: cada turno que reusa archivos grandes consume contexto y uso.
- Herramientas activas: búsqueda web, Research, conectores y ejecución de código consumen contexto aunque no se usen de forma explícita en cada mensaje.[2][3]

> **La reformulación que lo cambia todo**
>
> La eficiencia en claude.ai consiste en obtener un buen resultado con menos iteraciones, con el contexto bien configurado y sin arrastrar conversaciones innecesariamente largas. No se trata de perseguir una cifra de tokens que la interfaz ni muestra ni permite controlar directamente.[3]

Las palancas reales de la interfaz

Estas son las herramientas que la interfaz sí ofrece y que con frecuencia se pasan por alto en favor de conceptos más técnicos. Todas están documentadas por Anthropic en el centro de ayuda o en la documentación de producto.[2][3]

Preferencias de perfil

En Configuración > Perfil hay un campo de texto que actúa como instrucciones permanentes definidas por la persona usuaria. Se carga automáticamente en cada conversación nueva y está disponible en todos los planes, incluido el gratuito.[2]

Sirve para evitar repetir en cada sesión quién eres, cómo trabajas y qué formato prefieres.

Recomendación práctica: escribir instrucciones específicas, observables y breves, del estilo “respuestas de menos de 150 palabras salvo que pida lo contrario; empieza por la conclusión; evita frases de relleno”. Estas preferencias son un buen lugar para fijar el tono general, los idiomas y las limitaciones que conviene respetar de forma estable.

Estilos y Skills

Los estilos controlan el tono, el formato y la longitud de las respuestas. Anthropic está migrando este sistema hacia las Skills, que encapsulan patrones de tarea reutilizables.[7][4]

La distinción útil es esta: las preferencias de perfil definen quién eres y qué contexto traer; los estilos o Skills definen cómo responde Claude. Un estilo o Skill bien configurado evita repetir en cada mensaje instrucciones de formato.

A medida que Anthropic consolide las Skills, convendrá revisar periódicamente qué estilos siguen vigentes y cuáles es mejor sustituir por una Skill específica para la tarea (por ejemplo, “evaluar trabajos de 2.º de Bachillerato con esta rúbrica”).[7]

Proyectos

Los proyectos combinan instrucciones de proyecto, una base de conocimiento con archivos y el historial de chats de ese proyecto. El plan gratuito permite hasta cinco proyectos; los planes de pago, un número mucho mayor o ilimitado, según el caso.[3][2]

La documentación oficial destaca dos propiedades poco conocidas:

- El contenido subido a un proyecto se cachea y, cuando se reutiliza, solo las partes nuevas consumen contexto.
- El contexto no se comparte automáticamente entre chats de un mismo proyecto; lo que los conecta es la base de conocimiento común, no el historial de cada chat.[4][3]

Consecuencia práctica: subir todos los materiales de trabajo recurrentes (guías, rúbricas, temarios, plantillas, manuales internos) a la base de conocimiento del proyecto, usar las instrucciones del proyecto para el contexto estable y reservar las instrucciones concretas de cada tarea para el mensaje inicial de cada chat.

Es, en la práctica, la palanca de eficiencia más potente de la interfaz para entornos educativos y de consultoría.

Memoria

La memoria está disponible desde marzo de 2026 en todos los planes y genera una síntesis automática del historial de conversaciones para reutilizarla como contexto en sesiones nuevas.[3]

La memoria se segmenta por proyecto y admite indicaciones directas de “recuerda esto” y “olvida esto”, además de ofrecer un modo incógnito para chats que no se guardan.

En la práctica, la memoria sirve para conceptos de larga duración (quién eres, cómo trabajas, qué tipo de alumnado tienes, qué proyectos tienes activos) que no cambian de mensaje a mensaje. Conviene revisar periódicamente lo que la memoria ha almacenado para corregir sesgos o desajustes.

Control de esfuerzo

Desde Claude Opus 4.8 existe un control de esfuerzo en la propia interfaz, junto al selector de modelo. Permite elegir cuánto razonamiento dedica el modelo a cada respuesta: en niveles más bajos responde antes y consume uso más despacio; en niveles más altos piensa más a fondo, a costa de más uso y más latencia.[8][9][10]

En claude.ai, el menú muestra niveles con etiquetas como Low, Medium, High, Extra y Max según el modelo, mientras que en la API y en Claude Code se usan las formas low, medium, high, xhigh y max.[11][12]

El nivel xhigh, inicialmente disponible solo en Opus 4.8, se ha extendido también a Claude Sonnet 5 y Claude Fable 5, mientras que modelos de la generación anterior (Sonnet 4.6, Opus 4.6) se quedaron sin él. Claude Sonnet 5 usa High como nivel por defecto, pensado para razonamiento complejo y trabajo donde la calidad importa más que la velocidad o el coste.[11][19]

El artículo de ayuda resume su uso así: Low y Medium estiran más el uso para tareas rutinarias; High ofrece el mejor equilibrio general; Max se reserva para tareas difíciles y trabajo prolongado.[9]

Recomendación operativa para interfaz:

- Usar Low o Medium para tareas simples (resumir, reformatear, clasificar, correcciones menores).
- Dejar High como valor general para análisis, redacción cuidada y trabajo habitual de aula o consultoría.
- Reservar Extra/Max para razonamiento complejo, proyectos largos, planificación o problemas técnicos delicados, sabiendo que consumen más rápido el límite de uso.[13][8][9]

Pensamiento extendido y gestión de herramientas activas

El pensamiento extendido aparece en la interfaz como un interruptor asociado a “Thinking” o “Extended” dentro del menú de modelo y esfuerzo. Permite que Claude muestre un bloque colapsable con su razonamiento antes de la respuesta, lo que ayuda en tareas complejas, a costa de más uso.[9]

Las herramientas activas (búsqueda web, Research, conectores, ejecución de código) también gastan contexto. La documentación recomienda activarlas solo cuando la tarea las necesita y desactivarlas después, para evitar consumo innecesario de uso.[2][3]

Elección del modelo

Desde junio de 2026, la jerarquía de modelos tiene un peldaño más. Claude Fable 5, presentado el 9 de junio de 2026 y disponible de nuevo en todo el mundo desde el 1 de julio, tras una suspensión de casi tres semanas ordenada por el Gobierno de Estados Unidos (ver la nota siguiente), es el modelo insignia: supera en capacidad a cualquier modelo que Anthropic haya publicado antes, con un salto especialmente marcado en ingeniería de software, trabajo de conocimiento e investigación científica.[21][22]

Le sigue Claude Opus 4.8, recomendado para codificación agéntica compleja y trabajo de nivel empresarial.[23] Claude Sonnet 5, publicado el 30 de junio de 2026, sustituyó a Sonnet 4.6 como modelo por defecto en los planes gratuito y Pro: su rendimiento se acerca al de Opus 4.8, a un precio muy inferior.[19][20] Claude Haiku 4.5 sigue siendo la opción más rápida y económica para tareas sencillas.[23]

El precio también ordena la decisión: Fable 5 cuesta bastante más por token que Sonnet 5, así que reservarlo para las tareas donde su capacidad marca una diferencia real (no usarlo por defecto) es en sí mismo un criterio de eficiencia.[21]

En la interfaz, la elección del modelo y el nivel de esfuerzo funcionan como dos reguladores complementarios. Para tareas sencillas, Sonnet 5 o Haiku 4.5 a bajo esfuerzo suelen ser suficientes; para trabajos de evaluación compleja, análisis de documentación extensa o diseño de materiales didácticos avanzados, Fable 5 u Opus 4.8 con esfuerzo High o xhigh justifican el coste.

Nota sobre la suspensión de Fable 5 (12 al 30 de junio de 2026)

El caso de Fable 5 merece contarse completo, porque ilustra hasta qué punto la frase “el modelo más avanzado” puede caducar por motivos que nada tienen que ver con la técnica. El 12 de junio de 2026, tres días después del lanzamiento, Anthropic recibió a las 17:21 (hora del este de Estados Unidos) una directiva de control de exportaciones del Gobierno estadounidense que, invocando atribuciones de

seguridad nacional, ordenaba suspender el acceso a Fable 5 y a Mythos 5 para cualquier ciudadano extranjero, dentro o fuera del país, incluidos los propios empleados extranjeros de Anthropic. Al no poder verificar la nacionalidad de cada usuario en tiempo real, la compañía desactivó ambos modelos para todos los usuarios del mundo. Anthropic acató la orden y, en el mismo comunicado, discrepó públicamente de ella: a su juicio, el hallazgo de un jailbreak puntual no justificaba retirar un modelo comercial ya desplegado, y hacerlo con ese criterio paralizaría cualquier lanzamiento futuro de cualquier proveedor.[32]

El origen de la orden fue un informe de investigadores de Amazon que describía una técnica para sortear las salvaguardas de Fable 5 y lograr que el modelo identificara vulnerabilidades de software y, en un caso, generara una demostración de explotación. Las pruebas posteriores de Anthropic, revisadas junto al Gobierno, mostraron que varios modelos menos capaces, entre ellos Opus 4.8, GPT-5.5 y Kimi K2.7, reproducían los mismos resultados; es decir, la técnica no exponía capacidades exclusivas del nivel Mythos.[22]

La salida pactada consistió en un nuevo clasificador de seguridad, entrenado en colaboración con el propio Gobierno, que bloquea la técnica descrita en más del 99 % de los casos y redirige las peticiones marcadas a Opus 4.8, con aviso al usuario. Investigadores del CAISI, el centro de estándares e innovación en IA del Departamento de Comercio, verificaron las salvaguardas. La Administración estadounidense (la Administración Trump, según la atribución de la cobertura de prensa) levantó los controles de exportación el 30 de junio y Fable 5 volvió a estar disponible globalmente el 1 de julio de 2026.[22][33]

> **Por qué esta nota importa**
>
> Tres días después de su lanzamiento, el Gobierno de Estados Unidos ordenó apagar el modelo más capaz del mercado, y estuvo casi tres semanas fuera de servicio. Hasta una afirmación de tipo “disponible desde” necesita fecha y matiz, porque la disponibilidad de un modelo puede interrumpirse de un día para otro por decisión regulatoria, no solo técnica. La secuencia completa solo puede reconstruirse acudiendo a los comunicados oficiales de la compañía y contrastándolos con la cobertura de prensa que aporta la atribución política.[32][22][33]

De la ingeniería de prompts a la ingeniería de contexto

En 2026, Anthropic ya no presenta el "prompt engineering" (entendido como el arte de encontrar la frase mágica) como el centro del trabajo con modelos de lenguaje. El planteamiento oficial, recogido en el artículo de ingeniería "Effective context engineering for AI agents", desplaza el foco hacia la configuración del contexto.[14][15]

La idea central es que el contexto es un recurso finito con rendimientos decrecientes. Todos los modelos tienen una ventana de contexto limitada y un "presupuesto de atención" que no conviene llenar con información de baja señal.[15]

A medida que crece el número de tokens en el contexto, disminuye la probabilidad de que el modelo recupere con precisión una pieza concreta de información, fenómeno conocido como context rot.

La meta ya no es "dar todo el contexto posible", sino construir el conjunto mínimo de información de alta señal que maximiza la probabilidad de obtener el comportamiento deseado.[15]

Para quien usa la interfaz, esto se traduce en cinco hábitos prácticos, alineados con las buenas prácticas que Anthropic publica para construir agentes y workflows con Claude.[7][4]

1. Planificar antes de escribir

Antes de iniciar el chat, conviene decidir qué se necesita exactamente, si varias preguntas relacionadas pueden agruparse en un solo mensaje y qué contexto de fondo merece la pena aportar de entrada. Un primer mensaje bien construido ahorra mensajes de aclaración y reduce el desgaste de la cuota.[15][7]

2. Ser específico

Las consultas vagas obligan a Claude a pedir aclaraciones o a hacer supuestos. Esa ambigüedad se traduce en más iteraciones y más consumo de uso. Ser específico en objetivo, formato, extensión deseada y restricciones reduce el número de vueltas necesarias.[15][3]

3. Agrupar tareas relacionadas en un único mensaje

Si hay varias preguntas o tareas conectadas, es más eficiente enviarlas juntas, claramente numeradas, que fragmentarlas en mensajes sucesivos. Cada nuevo turno reprocesa el historial, de modo que la fragmentación excesiva tiene un coste acumulativo.[15][3]

4. Usar la base de conocimiento de Proyectos para material recurrente

Cuando la documentación de fondo (manuales, temarios, dossieres de cliente, normativa) se reutiliza con frecuencia, conviene alojarla en la base de conocimiento del proyecto. Así, el sistema puede cachear parte del contenido y solo las secciones nuevas o modificadas cuentan contra el límite.[4][3]

En la práctica, la base de conocimiento de Proyectos es el equivalente en la interfaz del caché de prompts de la API, pero automatizado y sin necesidad de tocar parámetros técnicos.[4]

5. Empezar una conversación nueva al cambiar de tarea

Cada turno de una conversación reprocesa todo el historial previo. Un mensaje enviado en el turno 25 de un hilo largo consume mucha más cuota que el mismo mensaje en el turno 3 de un hilo nuevo, porque la ventana de contexto tiene que incluir mucha más información.[3]

Arrastrar una conversación larga y desordenada no conserva la calidad, la degrada. Lo recomendable es cerrar un hilo productivo con un breve resumen de traspaso y, cuando la tarea cambie, abrir un chat nuevo con ese resumen como punto de partida.

Cómo reconocer contenido poco fiable sobre Claude

Esta sección sirve para entrenar la lectura crítica. Los ejemplos se desmontan con una verificación rápida en la documentación oficial.

El dato de los “346 tokens”

Circula una cifra concreta, 346 tokens, como coste fijo del system prompt cuando se usan herramientas. Esa cifra no aparece en la documentación de Anthropic. Las tablas oficiales de coste muestran valores distintos según modelo y configuración; además, se actualizan con frecuencia a medida que cambian modos y parámetros.[16][4]

Lección. *Un número sin enlace directo a documentación oficial no es un dato, es un rumor con formato de dato.*

Los “89 comandos de Claude”

Abundan infografías que prometen decenas de comandos “/” en el chat de claude.ai. En realidad, la mayoría de esos comandos pertenecen a Claude Code, que vive en el terminal y en el IDE; en la interfaz de chat, lo que existe con “/” son básicamente invocaciones de estilos o Skills.[5][6]

Lección. *Antes de enseñar una función, conviene comprobar en qué producto vive y a qué tipo de usuario se dirige.*

Las cifras de mensajes por plan

Es habitual leer frases del tipo “tienes unos 45 mensajes en Pro” o “900 en Max 20x”. Anthropic no publica un número fijo de mensajes porque el uso depende de la longitud de las conversaciones, del modelo, del nivel de esfuerzo, de las herramientas activas y de la demanda del sistema.[17][3]

Lección. *Cualquier cifra de mensajes por plan debe presentarse como estimación basada en pruebas de terceros, nunca como dato oficial.*

Las cifras de rendimiento en español

También circulan porcentajes concretos (98,1%, 98,2%) citados como rendimiento de la generación actual en español. Esos valores corresponden, según las tablas oficiales de soporte multilingüe publicadas en distintas generaciones, a modelos anteriores; las tablas se actualizan conforme salen nuevas versiones, por lo que la combinación “porcentaje sin modelo ni fecha” es incompleta.[2][4]

Lección. *Una cifra de rendimiento sin el modelo y la fecha al lado es una cifra a medias.*

El modelo “tope”

Otra fuente frecuente de confusión es la afirmación de que un modelo concreto es “el más avanzado” sin fecha.

Este mismo documento sirve de ejemplo. La versión anterior, cerrada el 31 de mayo de 2026, señalaba a Claude Opus 4.8 como el modelo más capaz disponible, lanzado apenas tres días antes. Cinco semanas después, Claude Sonnet 5 sustituyó a Sonnet 4.6 como modelo por defecto y Claude Fable 5 se convirtió en el nuevo modelo insignia, con capacidades superiores a las de Opus 4.8.[19][21][22] La afirmación caducó antes de que nadie tuviera ocasión de señalarla como desactualizada.

Lección. *Toda afirmación de tipo “modelo tope” debería ir fechada y enlazada al anuncio oficial, y quien la lea debería asumir, por defecto, que en unas semanas dejará de ser exacta.*

> **El criterio, en una frase**
>
> Contenido que habla de límites, precios o capacidades sin enlazar ni a docs.anthropic.com ni a support.claude.com no es material de referencia, sino opinión.

Más allá del chat: Claude Code, Cowork y Design

La familia de productos de Claude incluye varias superficies que comparten la misma cuota de uso pero tienen lógicas de trabajo distintas a la del chat. Conviene conocerlas, aunque muchas personas solo usen claude.ai.[5][3]

Claude Code

Claude Code es el sistema de programación agéntica de Anthropic. Se ejecuta en el terminal y se integra con editores como VS Code, opera a nivel de proyecto y puede leer el código, proponer cambios, ejecutar pruebas y manejar flujos de Git de forma semiautónoma.[6][5]

Su documentación específica detalla comandos, archivo de configuración, conexión con servicios externos mediante MCP y gestión de esfuerzo y contexto para código.[6][4] Es también el lugar donde viven muchos de los comandos "/" que a veces se atribuyen por error a la interfaz.

Claude Cowork

Claude Cowork es la superficie orientada a flujos de automatización para perfiles no técnicos. Permite gestionar archivos y carpetas locales, automatizar tareas repetitivas y configurar Skills personalizados que luego pueden invocarse tanto en Cowork como en claude.ai.[18][2]

Comparte modelo y cuota de uso con el resto de superficies. El control de esfuerzo y funciones como Dynamic Workflows en Opus 4.8 hacen de Cowork un punto de entrada accesible a flujos agénticos sin necesidad de escribir código.[10][18][8]

Claude Design

Claude Design, lanzado como producto de Anthropic Labs en fase de research preview el 17 de abril de 2026, es la superficie de creación visual. Permite generar prototipos interactivos, slides, one-pagers y maquetas desde lenguaje natural, y se integra como sección dentro de claude.ai.[8][2]

Está disponible en planes Pro, Max, Team y Enterprise, con acceso desde la barra lateral. A fecha de este documento sigue en research preview, sin fecha anunciada de disponibilidad general: Anthropic ha señalado que dejará que el uso real y el feedback de las personas usuarias determinen cuándo el producto está listo.

Para docencia y consultoría, permite ir de una especificación en lenguaje natural a un primer prototipo visual sin pasar por herramientas de diseño tradicionales.

ChatGPT: qué cambia y qué no

Buena parte de lo dicho sobre Claude tiene un equivalente directo en ChatGPT, con una salvedad importante: la estructura de planes de OpenAI cambia con más frecuencia y las fuentes fiables al respecto son la documentación oficial y la ayuda de OpenAI, no las comparativas de terceros que proliferan en redes.[24][25]

La estructura de planes, a fecha de hoy

A 18 de julio de 2026, ChatGPT mantiene un plan gratuito, un plan intermedio y planes de pago superiores. OpenAI introdujo a comienzos de 2026 un nivel llamado ChatGPT Go, pensado para cubrir el hueco entre no pagar nada y la suscripción estándar; ofrece un límite de mensajes más amplio que el plan gratuito, aunque en algunas regiones incluye publicidad y sigue restringido a las variantes más ligeras de cada modelo.[25]

OpenAI también retira modelos de ChatGPT por calendario público: GPT-4.5 salió de la interfaz el 27 de junio de 2026 tras un periodo de retiro de 30 días, y OpenAI o3 saldrá el 26 de agosto de 2026 tras un periodo de 90 días. En los planes Enterprise y Edu, varios modelos quedaron retirados de ChatGPT desde el 13 de febrero de 2026, sin que el acceso por API se viera afectado. Estos cambios confirman que la oferta de ChatGPT cambia con más frecuencia que la de Claude, y que cualquier guía de uso debe fecharse y contrastarse contra el centro de ayuda oficial antes de darse por buena.[24][30]

El rumor de “pronto todo será de pago”

Circula la idea de que ChatGPT, y la IA generativa en general, dejará de tener versión gratuita en un futuro cercano. No hay ningún anuncio oficial de OpenAI, Anthropic ni Google que confirme el cierre de sus niveles gratuitos.

Lo que sí es verificable es una tendencia distinta y más matizada: los planes gratuitos incorporan cada vez más publicidad y restricciones, y los modelos más recientes llegan primero, o exclusivamente, a quien paga. La presión hacia el pago es real; el cierre total del acceso gratuito, a día de hoy, es una proyección sin confirmar, no un hecho.

Lección. *Antes de tomar una decisión de presupuesto basada en “esto va a desaparecer”, conviene comprobar si existe un anuncio oficial con fecha o si es una lectura de tendencia de un periodista o de un artículo de opinión.*

Ahorro de contexto en ChatGPT frente a Claude

El principio de fondo (menos iteraciones, mensajes bien planteados, reutilización de contexto en vez de repetirlo) es el mismo en ambas herramientas, porque responde a cómo funciona cualquier modelo de lenguaje, no a una particularidad de marca.

Lo que cambia es dónde vive cada palanca. En ChatGPT, el equivalente a las preferencias de perfil y a la base de conocimiento de Proyectos de Claude son la personalización de memoria y los GPT personalizados; el principio de "sube una vez el material recurrente y reutilízalo" aplica igual, aunque la implementación técnica sea distinta.[24]

Antes de comparar cifras concretas de rendimiento o de coste entre Claude y ChatGPT, hay que verificar que se compara el mismo tipo de plan (gratuito, intermedio o superior) y la misma variante de modelo, porque ambas compañías mantienen varias versiones activas a la vez.

Una idea para llevarse

La eficiencia con Claude, con ChatGPT o con cualquier IA generativa en la interfaz no consiste en memorizar trucos ni en obsesionarse con cifras de tokens.

Consiste en configurar bien el contexto reutilizable (preferencias, estilos, proyectos, memoria), en escribir mensajes claros y específicos, en abrir conversaciones nuevas cuando cambia la tarea y en ajustar modelo, esfuerzo y herramientas al trabajo concreto de cada momento.[15][2][3]

Y, por encima de todo, consiste en verificar lo que se aprende contra la documentación oficial de Anthropic y de OpenAI, y contra fuentes primarias como el informe de la ONU citado en este documento cuando se trata de impacto ambiental.

> **Para llevarse**
>
> En un terreno que cambia cada pocas semanas, el hábito de comprobar vale más que cualquier dato aislado o atajo puntual. Y en un terreno con coste ambiental real, ese mismo hábito de pedir solo lo necesario deja de ser solo una cuestión de cuota mensual.[2][4][26]

Fuentes

[1] Anthropic. Documentación oficial de Claude. https://docs.anthropic.com

[2] Anthropic. Introducción a Claude. https://platform.claude.com/docs/en/intro

[3] Anthropic. Límites de uso y longitud. https://support.claude.com/en/articles/11647753

[4] Anthropic. Documentación de la API de Claude. https://platform.claude.com/docs/en/home

[5] Anthropic. Claude Code (producto). https://www.anthropic.com/product/claude-code

[6] Anthropic. Claude Code (repositorio). https://github.com/anthropics/claude-code

[7] Anthropic. Building Effective AI Agents. https://resources.anthropic.com/building-effective-ai-agents

[8] Anthropic. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8

[9] Anthropic. Ajustes de modelo, esfuerzo y pensamiento. https://support.claude.com/en/articles/8664678

[10] Pronetic / Geeknetic. Anthropic lanza Claude Opus 4.8. https://pronetic.geeknetic.es/Noticia/38774

[11] Anthropic. Effort. Claude API Docs. https://platform.claude.com/docs/en/build-with-claude/effort

[12] LiteLLM. Anthropic Effort Parameter. https://docs.litellm.ai/docs/providers/anthropic_effort

[13] FindSkill. Claude Opus 4.8: niveles de esfuerzo. https://findskill.ai/es/blog/claude-opus-4-8-niveles-de-esfuerzo/

[14] Anthropic. Engineering blog. https://www.anthropic.com/engineering

[15] Anthropic. Effective context engineering for AI agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

[16] Anthropic. What's new in Claude Opus 4.8. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-8

[17] Anthropic. Manage usage credits. https://support.claude.com/en/articles/12429409

[18] Pisapapeles. Anthropic presenta Claude Opus 4.8. https://pisapapeles.net/anthropic-presenta-claude-opus-4-8-con-mejoras-en-codigo-agentes-y-uso-de-herramientas/

[19] Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5

[20] Anthropic. What's new in Claude Sonnet 5. Claude Platform Docs. https://platform.claude.com/docs/en/about-claude/models/whats-new-sonnet-5

[21] Anthropic. Claude Fable 5 and Claude Mythos 5. https://www.anthropic.com/news/claude-fable-5-mythos-5

[22] Anthropic. Redeploying Claude Fable 5. https://www.anthropic.com/news/redeploying-fable-5

[23] Anthropic. Models overview. Claude Platform Docs. https://platform.claude.com/docs/en/about-claude/models/overview

[24] OpenAI. Notas de lanzamiento de modelos. OpenAI Help Center. https://help.openai.com/es-419/articles/9624314-notas-de-lanzamiento-del-modelo

[25] OpenAI. Planes y precios de ChatGPT (documentación y ayuda oficial de producto). https://openai.com/chatgpt/pricing

[26] UNU-INWEH (Instituto de la Universidad de las Naciones Unidas para el Agua, el Medio Ambiente y la Salud). Environmental Cost of AI's Energy Use, 3 de junio de 2026. https://inweh.unu.edu

[27] Naciones Unidas. La ONU pidió a las empresas de inteligencia artificial revelar su impacto ambiental. Noticias ONU, 3-4 de junio de 2026. https://news.un.org/es/story/2026/06/1541523

[28] OpenAI. Sam Altman, blog personal. Estimación de consumo energético e hídrico por consulta media de ChatGPT. https://blog.samaltman.com

[29] Google. Informe técnico sobre el coste energético de la inferencia en Gemini. Google Cloud Blog. https://cloud.google.com/blog

[30] OpenAI. ChatGPT Enterprise and Edu: modelos y límites. OpenAI Help Center. https://help.openai.com/es-es/articles/11165333-chatgpt-enterprise-y-edu-modelos-y-l%C3%ADmites

[31] Investigación independiente de verificación cruzada sobre coste ambiental y eficiencia de contexto en IA generativa, realizada el 18 de julio de 2026 a partir de fuentes primarias (UNU-INWEH, OpenAI, Google) y de este mismo documento como referencia de partida.

[32] Anthropic. Statement on the US government directive to suspend access to Fable 5 and Mythos 5, 12 de junio de 2026. https://www.anthropic.com/news/fable-mythos-access

[33] CNBC. Anthropic says Trump admin has lifted export controls on Claude Fable 5 and Mythos 5, 30 de junio de 2026. https://www.cnbc.com/2026/06/30/anthropic-says-trump-admin-has-lifted-export-controls-on-claude-fable-5-and-mythos-5.html

Criterio de fiabilidad

Para esta investigación conviene priorizar documentación oficial, blog de producto e ingeniería de Anthropic y de OpenAI, además de informes de organismos internacionales como Naciones Unidas cuando se trata de impacto ambiental, porque son las fuentes que describen el funcionamiento real del sistema, sus cambios recientes y su huella verificada. Las fuentes comunitarias pueden servir para detectar síntomas o casos de uso, pero no deberían ser la base principal de una investigación técnica o docente.

Cuando el objetivo es explicar coste, rendimiento y calidad lingüística, la combinación más sólida es: documentación técnica oficial, notas de ingeniería y una lectura crítica del contexto de producto. Esa combinación permite separar mejor qué pertenece al modelo, qué pertenece al prompting y qué pertenece a la interfaz o al entorno de ejecución.