

La mejor IA para tus tareas no existe (y por eso ganas tú)

Te voy a ser sincera: no te cases con ninguna IA. Sí, hoy Claude va por delante en las tareas largas que se ejecutan solas. Pero la palabra clave es hoy. Cada pocas semanas cambia quién manda, y cada modelo gana en cosas distintas. Por eso lo inteligente no es elegir una IA, sino montar un sistema que mande cada tarea a la que mejor la resuelve, y que puedas cambiarla sin rehacerlo todo. Esto es lo que no se queda atrás.

Por qué el líder cambia cada pocas semanas

Esto no es una exageración para vender. Es literal. Solo en lo que va de 2026 ha cambiado el número uno varias veces, y los modelos se actualizan a un ritmo que no habías visto en ninguna otra herramienta de trabajo.

A finales de junio de 2026, Claude Opus 4.7 lidera SWE-bench Verified (la prueba estándar de programación) con un 91%, por delante de GPT-5.5 y Gemini 3.1 Pro. Pero baja un escalón a la prueba difícil, SWE-bench Pro, y el orden se mueve: ahí el modelo activo más fuerte es Claude Opus 4.8 con 69,2%, mientras GPT-5.4 lidera la lista pública de Scale. Tres semanas antes, esa foto era distinta. Tres semanas después, lo será otra vez.

El error que te cuesta caro

Casarte con una IA es apostar a que el líder de hoy seguirá siéndolo. Nunca ha sido verdad ni una sola vez. Cuando montas todo tu trabajo encima de un solo modelo, cada vez que otro lo adelanta te quedas atrás, y rehacerlo todo cuesta más que haberlo montado bien desde el principio.

Cada modelo gana en cosas distintas

Aquí está lo que casi nadie te dice: no hay un "mejor modelo". Hay un mejor modelo para cada cosa. Y las diferencias son reales, no de matiz.

Para tareas largas que se ejecutan solas, manda Claude. Cuando necesitas que la IA encadene muchos pasos sin perder el hilo —revisar un proyecto entero, trabajar turno tras turno hasta terminar— Claude Opus 4.8 es hoy el más fiable, y por eso es la base de herramientas como Claude Code. Aguenta los procesos largos sin descarrilar.

Para volumen, Gemini te sale mucho más barato. Cuando vas a procesar miles de elementos, el precio por uso decide. Gemini 3 Pro cuesta alrededor de \$2 de entrada y \$12 de salida por millón de tokens (las "palabras" que consume la IA), frente a los \$5 / \$25 de Claude Opus y los \$5 / \$30 de GPT-5.5. Y sus versiones Flash son aún más baratas. En una tarea puntual no lo notas; en una

que corre mil veces al mes, es la diferencia entre \$30 y \$300.

Para leer un documento enorme de una sola vez, Gemini otra vez. Su ventana de contexto llega a entre 1 y 2 millones de tokens: unas 1.500 páginas en una sola pasada. Si le tienes que dar un contrato gigante, un año de informes o un manual completo sin trocearlo, ahí gana.

Para agentes que viven en la terminal, GPT. GPT-5.5 es hoy el más fuerte en ese terreno concreto (82,7% en Terminal-Bench 2.0). No es el mejor en todo, pero en lo suyo va delante.

Para datos en tiempo real, Grok. Es el único modelo de primera línea conectado en directo a lo que se publica en X, unos 68 millones de posts en inglés al día. Si tu tarea necesita lo que está pasando ahora mismo —vigilar tu marca, seguir una tendencia, reaccionar a una noticia— ahí no tiene rival. Encima va rapidísimo y barato: su versión Fast cuesta unos \$0,20 / \$0,50 por millón de tokens, lo más económico del mercado frontera. Y en las pruebas de razonamiento más duras también va en cabeza.

Un detalle que confunde a todo el mundo

Los precios "por token" no se comparan directos entre marcas, porque cada una cuenta los tokens distinto. El nuevo Claude Opus 4.7, por ejemplo, puede generar hasta un 35% más de tokens que su versión anterior para el mismo texto. Mira el coste real de tu tarea, no el número de la tabla.

Lo inteligente: un sistema que enruta, no una IA que eliges

Si cada modelo gana en algo distinto y el líder cambia cada pocas semanas, elegir uno solo es la jugada equivocada por diseño. La jugada buena es montar un sistema que reciba la tarea y la mande a la IA correcta:

- 1. Clasifica la tarea.** ¿Es un proceso largo, un volumen enorme, un documento gigante, datos de hoy, una respuesta rápida? Cada tipo tiene su modelo.
- 2. La manda al modelo que mejor la resuelve.** El proceso largo va a Claude, el volumen barato a Gemini Flash, el documento de 800 páginas a Gemini Pro, lo que necesita datos de hoy a Grok. Tú no eliges a mano cada vez: el sistema lo hace.
- 3. Te deja cambiar el modelo sin rehacer nada.** Esta es la pieza clave. Si mañana sale uno mejor o más barato, cambias una línea y todo lo demás sigue igual.

Eso último es lo que de verdad importa. No estás eligiendo la mejor IA de junio de 2026: estás montando algo que usa la mejor IA sea cual sea el mes.

La forma fácil de pensarlo

No contrates a un solo empleado y reza por que siga siendo el mejor del mercado para siempre. Monta un equipo y un jefe que reparte cada tarea a quien la hace mejor. Cuando aparece alguien superior, lo metes al equipo sin despedir a todos los demás.

Cómo se monta sin ser técnico

No necesitas programar. Las herramientas de automatización con IA (como las que conectan modelos por debajo) te dejan definir esa lógica de enrutado una vez. La idea práctica:

Empieza identificando tus tres o cuatro tareas repetitivas reales: "responder correos de soporte", "resumir documentos largos", "etiquetar entradas de una hoja de cálculo". Para cada una, prueba dos modelos y quédate con el que dé mejor resultado al menor coste. Luego conéctalos en un solo flujo donde cada tarea ya sale dirigida a su modelo. Lo construyes una vez; a partir de ahí solo cambias piezas cuando el mercado cambia.

Cuándo NO necesitas montar esto

Voy a ser honesta para que no pierdas el tiempo. Esto no es para todo el mundo todavía.

Si usas la IA tres veces al día, no montes infraestructura. Abre el chat del modelo que prefieras y ya. El enrutado justifica su trabajo cuando tienes tareas que se repiten en volumen o procesos que corren solos.

Si nunca has medido el coste, empieza por ahí. No optimices algo que no estás midiendo. Mira primero qué tareas te consumen tiempo o dinero de verdad, y monta el sistema solo para esas.

Un sistema mal montado es peor que un solo modelo. Si el enrutado manda la tarea al modelo equivocado, multiplicas errores en lugar de ahorrar. La lógica de "qué tarea va a qué modelo" es lo que tienes que pensar bien, no la tecnología.

Empieza esta semana

Elige UNA tarea repetitiva que ya hagas con IA. Pásala por dos modelos distintos y compara resultado y coste. Solo con eso ya vas a ver que ninguno gana en todo, y vas a entender por qué el futuro no es elegir una IA, sino saber repartir.