Talutil

Claude Fable 5: el modelo que Anthropic tenía bajo llave y acaba de soltar al público

10 Jun 2026 · 8 min de lectura

Anthropic llevaba meses prometiendo que Mythos era demasiado peligroso para el público general. Hoy, 9 de junio de 2026, lo ha lanzado de todas formas, con un nombre diferente y con los frenos puestos. Claude Fable 5 es el primer modelo de clase Mythos disponible para cualquier persona con una suscripción o acceso a la API, y las primeras horas de datos confirman que el salto de capacidad es real. Si usas IA para programar, analizar documentos complejos o construir flujos de trabajo automatizados, este lanzamiento cambia el punto de referencia que tenías. No es una actualización incremental del modelo anterior: es una categoría nueva que Anthropic coloca por encima de Opus, con todo lo que eso implica en rendimiento, precio y restricciones operativas que conviene entender antes de integrarlo en cualquier proyecto.

Qué es Claude Fable 5 y de dónde viene

Claude Fable 5 no nació en el ciclo de desarrollo ordinario de Anthropic. Es, en esencia, la versión pública del modelo Mythos, una arquitectura que la empresa presentó en abril de 2026 con acceso restringido a un grupo muy reducido de organizaciones bajo el marco de Project Glasswing, una iniciativa en colaboración con el gobierno de Estados Unidos centrada en ciberseguridad defensiva.

Por qué Mythos estuvo cerrado hasta hoy

Mythos se presentó inicialmente con acceso restringido por un motivo concreto: se identificó que el modelo tenía la capacidad de detectar y explotar fallas de ciberseguridad en segundos. Eso no es una afirmación de marketing, sino el resultado de evaluaciones internas que llevaron a Anthropic a no lanzarlo de forma abierta durante meses. Según documentación técnica de abril de 2026, el modelo demostró capacidades para descubrir y explotar vulnerabilidades zero-day de forma autónoma, ejecutar escapes de sandbox en entornos controlados e identificar miles de fallos críticos en software sin supervisión humana.

La solución que encontró Anthropic para hacer el lanzamiento viable no fue esperar a que el modelo fuera menos capaz, sino añadirle un sistema de clasificadores que intercepta las consultas de riesgo antes de que el modelo responda.

La relación entre Fable 5 y Mythos 5

Claude Mythos 5 ofrece el mismo modelo principal que Fable 5, pero con menos restricciones para investigadores aprobados. Dicho de otra forma: ambos modelos comparten la misma arquitectura base. La diferencia no está en las capacidades, sino en los filtros que se aplican sobre ellas. Claude Mythos 5 no está disponible de forma general; se ofrece con disponibilidad limitada a clientes aprobados en Project Glasswing. Para el resto del mundo, existe Fable 5, que llega con los clasificadores de seguridad activos por defecto.

Qué puede hacer que sus predecesores no podían

El argumento habitual en cada lanzamiento de modelo es que los benchmarks mejoran. Lo que distingue a Fable 5 es la magnitud del salto y el tipo de tareas donde ese salto ocurre.

Los números en benchmarks de programación

Claude Fable 5 obtuvo un 95,0% en SWE-bench Verified, 80,0% en SWE-bench Pro y lideró FrontierCode tanto en los subconjuntos Diamond como Main. Para contextualizar esos números: en SWE-Bench Pro, el benchmark estándar para evaluar modelos en tareas reales de ingeniería de software, Fable 5 alcanzó un 80,3% frente al 69,2% que registra su predecesor Opus 4.8. Es además el primer modelo en superar el 90% en el benchmark analítico de Hex, compuesto por tareas analíticas largas y complejas.

Estos números tienen peso en contextos reales. Stripe reportó que Fable 5 comprimió meses de trabajo de ingeniería en días, completando la migración de una gran base de código de Ruby que habría llevado a un equipo más de dos meses.

Razonamiento científico autónomo

Lo más llamativo de las demostraciones publicadas por Anthropic no es el rendimiento en código: es lo que el modelo Mythos 5 hizo en investigación científica antes del lanzamiento público. Mythos 5 llevó a cabo investigaciones genómicas durante más de una semana de trabajo en gran medida autónomo, ensambló datos de células individuales de millones de células de 138 especies animales, y diseñó y entrenó un modelo de aprendizaje automático que con únicamente orientación humana de alto nivel superó a un modelo reciente publicado en la revista Science, siendo 100 veces más pequeño.

Anthropic ha indicado que planea publicar esos resultados en los próximos meses. También se informa que Mythos 5 logró una aceleración de 10x en partes del proceso de diseño de medicamentos.

La ventaja crece con la complejidad

La ventaja del modelo sobre sus modelos anteriores aumenta con la longitud y la complejidad de las tareas. Esto es relevante para quienes trabajan con documentos extensos, flujos de trabajo multi-paso o proyectos de código que requieren coherencia a lo largo de muchas iteraciones. En tareas cortas y simples, la diferencia con Opus 4.8 puede no justificar el coste adicional. En proyectos complejos de largo plazo, los primeros reportes sugieren que sí lo hace.

Los clasificadores de seguridad: cómo funcionan y qué significa en la práctica

Este es el punto que más confusión ha generado en las primeras horas de uso. Fable 5 no es un modelo sin restricciones que llega con advertencias de seguridad en los términos de servicio. Es un modelo con un sistema activo de filtrado que puede interrumpir conversaciones de forma automática.

Qué bloquea y a dónde va la solicitud

Fable 5 lleva integrados clasificadores de seguridad que vigilan tres terrenos concretos: técnicas ofensivas de ciberseguridad como crear exploits, malware y herramientas de ataque; contenido de biología y ciencias de la vida que incluye métodos de laboratorio y mecanismos moleculares; y la extracción del pensamiento resumido del modelo.

Cuando uno de esos clasificadores detecta una solicitud en esas áreas, la conversación no se interrumpe: se redirige. El cambio automático de modelo está habilitado por defecto la primera vez que seleccionas Claude Fable 5. Con el cambio automático de modelo desactivado, una solicitud bloqueada pausa la conversación en lugar de cambiar modelos.

La tasa de falsos positivos

Anthropic reconoce que estas medidas de seguridad en ocasiones detectarán solicitudes inofensivas, aunque se activan en promedio en menos del 5% de las sesiones. Para la mayoría de los casos de uso profesional eso es manejable. Para entornos de producción con alto volumen, conviene configurar el comportamiento de fallback desde el primer día.

Disponibilidad, precios y qué plan lo incluye ahora mismo

La disponibilidad de Fable 5 es uno de los pocos aspectos del lanzamiento sin ambigüedad: el modelo está disponible hoy.

Dónde está disponible

Claude Fable 5 está disponible de forma general en la Claude API, Claude Platform en AWS, Amazon Bedrock, Vertex AI y Microsoft Foundry. En la API el identificador del modelo es `claude-fable-5`. Para quienes usan Claude directamente desde claude.ai, el acceso está incluido en los planes Pro, Max, Team y Enterprise hasta el 22 de junio de 2026. A partir del 23 de junio, podrían requerirse créditos de uso adicionales hasta que se aumente la capacidad.

Precio en la API de Claude Fable 5

Claude Fable 5 y Claude Mythos 5 tienen un precio de 10 dólares por millón de tokens de entrada y 50 dólares por millón de tokens de salida. El precio por lotes (batch) es de 5 dólares por millón de tokens de entrada y 25 por millón de tokens de salida.

Aunque el precio por token es más alto que el de modelos anteriores, los primeros reportes de clientes indican que Fable completa las tareas usando menos tokens en total, lo que reduce el costo real por proyecto. Eso no es una garantía, pero es un patrón que vale la pena medir con casos de uso propios antes de descartar el modelo por precio.

Una restricción técnica importante para equipos enterprise

Fable 5 y Mythos 5 son Covered Models: tienen retención de datos de 30 días y no están disponibles bajo retención cero de datos (zero data retention). Si la política de empresa exige ZDR, esto afecta directamente la viabilidad de uso. Es un detalle que conviene revisar con el equipo legal o de seguridad antes de integrarlo en flujos de trabajo con datos sensibles.

Qué significa esto para quienes ya usaban Claude Opus 4.8

Talutil publicó hace poco el análisis completo de Claude Opus 4.8, el modelo que hasta hoy era el techo de capacidad pública de Anthropic. Con el lanzamiento de Fable 5, Opus 4.8 no desaparece, pero su posición cambia.

Cuándo seguir usando Opus 4.8

Para tareas con volumen alto de solicitudes cortas, análisis de documentos estándar o cualquier flujo de trabajo donde los costes de API sean un factor determinante, Opus 4.8 sigue siendo una opción sólida a menor precio. También es el modelo al que Fable 5 redirige automáticamente las solicitudes bloqueadas, lo que en la práctica lo convierte en el segundo nivel del nuevo sistema de Anthropic.

Cuándo tiene sentido pasarse a Fable 5

Si el trabajo implica migraciones de código complejas, análisis de documentos muy largos, investigación técnica avanzada o flujos de trabajo agénticos de varias horas, los datos iniciales apuntan a que Fable 5 justifica el coste adicional. La ventana hasta el 22 de junio con acceso incluido en los planes de suscripción es el momento natural para

hacer esa evaluación con datos reales.