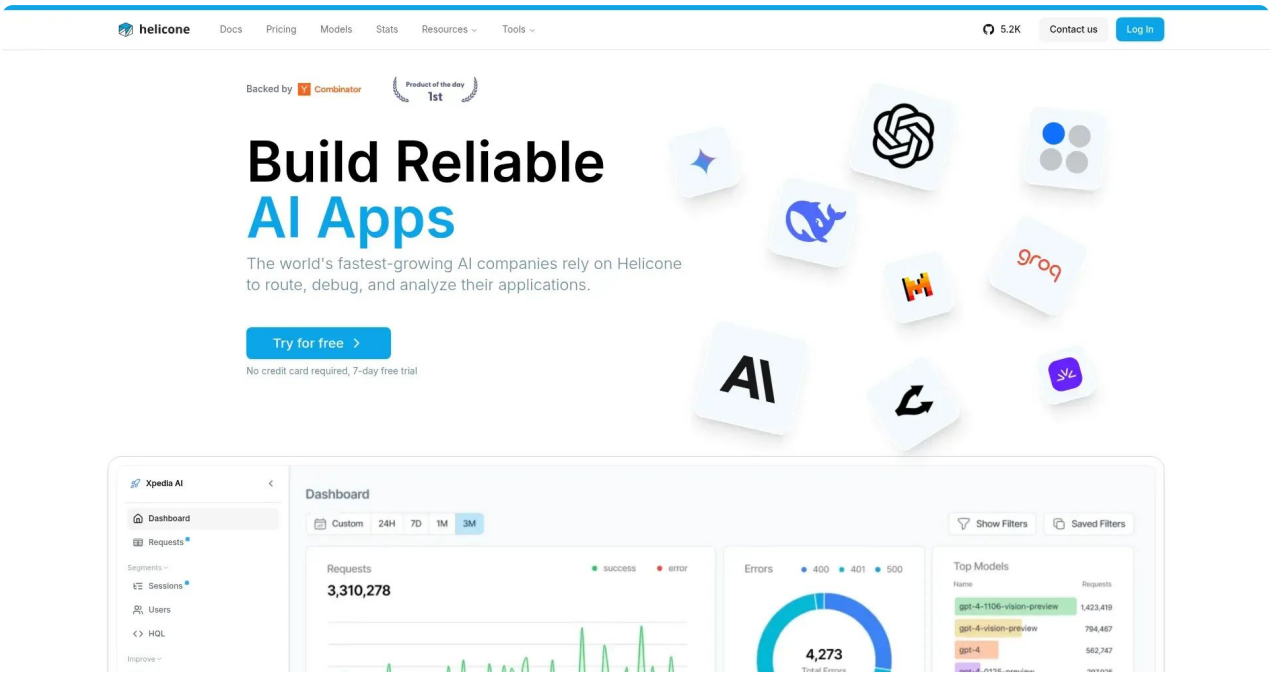


# Helicone

*Helicone es una plataforma de observabilidad y gateway de código abierto diseñada para desarrolladores y empresas que construyen aplicaciones con modelos de lenguaje extenso (LLM). Actúa como middleware entre la aplicación y proveedores como OpenAI o Anthropic, permitiendo monitorizar, depurar y optimizar interacciones en tiempo real. Es ideal para equipos de ingeniería que necesitan visibilidad total sobre costes, latencia y rendimiento, ofreciendo funciones de caching, reintentos y gestión de prompts.*

Visitar Sitio Oficial Preguntar a ChatGPT Preguntar a Claude Preguntar a Grok

## Contenido del Dossier

## INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es
Helicone es una plataforma de observabilidad y pasarela (gateway) de código abierto diseñada específicamente para desarrolladores y empresas que construyen aplicaciones basadas en modelos de lenguaje extenso (LLM). Actúa como un middleware entre la aplicación y los proveedores de IA (como OpenAI, Anthropic o Google), permitiendo monitorizar, depurar y optimizar cada interacción en tiempo real. Está dirigido a equipos de ingeniería, arquitectos de soluciones de IA y departamentos de producto que buscan visibilidad total sobre sus costes y el rendimiento de sus modelos.

Principal ventaja profesional
La capacidad de centralizar la observabilidad de múltiples proveedores de IA mediante una única línea de código, permitiendo implementar funciones críticas como el almacenamiento en caché (caching), reintentos automáticos (retries) y límites de cuota (rate limiting) sin modificar la lógica interna de la aplicación, lo que reduce drásticamente los costes operativos y el tiempo de depuración.

Para quién no es
No es una herramienta para usuarios finales sin conocimientos de programación ni para empresas que solo realizan experimentos aislados y manuales con ChatGPT. Profesionales de marketing o gestión de contenidos que no intervengan en el flujo técnico de la API encontrarán la herramienta excesivamente técnica y fuera de su ámbito de operación.

funcionalidades clave
- Gateway unificado: Acceso a más de 100 modelos (OpenAI, Anthropic, Llama, etc.) a través de una API compatible.
- Observabilidad en tiempo real: Rastreo completo de solicitudes, respuestas, latencia y uso de tokens.
- Caching inteligente: Almacenamiento de respuestas para ahorrar costes y reducir la latencia en consultas repetitivas.
- Gestión de Prompts: Versionado y pruebas de prompts directamente desde la plataforma (Playground).
- Análisis de costes: Desglose detallado por usuario, modelo o propiedades personalizadas para control presupuestario.
- Resiliencia: Configuración de reintentos automáticos y fallbacks (cambio automático a otro modelo si uno falla).
- Exportación y Datasets: Capacidad de crear conjuntos de datos a partir de los logs para procesos de fine-tuning.

Precios
- Versión gratuita (Hobby): Gratuita. Incluye hasta 10.000 solicitudes al mes, 1 GB de almacenamiento de logs y retención de datos por 7 días.
- Pro: 79$ al mes. Incluye usuarios ilimitados, alertas, informes detallados y lenguaje de consulta HQL (Helicone Query Language). Aplica tarificación adicional por uso excedente.
- Team: 799$ al mes. Añade 5 organizaciones, cumplimiento SOC-2 y HIPAA, y soporte dedicado por Slack.
- Enterprise: Precio bajo presupuesto. Incluye despliegue on-premise (en servidores propios), Single Sign-On (SAML) y contratos personalizados (MSA).

Perfil del usuario
- Empresas tecnológicas (SaaS) que integran funciones de IA en sus productos.
- Departamentos de ingeniería y DevOps que gestionan infraestructuras de IA a escala.
- Equipos de Producto que necesitan validar la calidad de las respuestas de los modelos.
- Agencias de desarrollo de software que gestionan múltiples clientes y necesitan separar costes y accesos.

Nivel técnico requerido
- Nivel de uso: Medio. Requiere familiaridad con el manejo de APIs y paneles de control técnicos.
- Nivel de instalación: Medio. Se integra sustituyendo la URL base del proveedor (ej. OpenAI) por la de Helicone o instalando su SDK de una línea.
- Soporte necesario: Requiere colaboración mínima de ingenieros de backend para la configuración inicial de las variables de entorno.
- Competencias necesarias: Conocimientos de desarrollo en Python, Node.js o cURL, y comprensión básica de la arquitectura de APIs REST.

Ejemplos de uso profesional
- Control de costes por cliente: Un SaaS puede etiquetar cada solicitud con un ID de cliente para refacturar el consumo de IA de forma exacta.
- Auditoría y cumplimiento: Registro y revisión de todas las interacciones de los empleados con LLMs para asegurar que no se comparte información sensible.
- Optimización de latencia: Uso de la caché de Helicone para servir respuestas instantáneas a preguntas frecuentes de un chatbot, ahorrando el coste del modelo.
- Pruebas A/B de modelos: Comparar el rendimiento y coste entre GPT-4 y Claude 3.5 Sonnet para una tarea específica antes de pasar a producción.

Uso y distribución
- Versión web: Panel de control completo para visualización de datos y configuración.
- SDKs: Disponibles para TypeScript/JavaScript y Python.
- CLI: Herramientas de línea de comando para integración en flujos de trabajo.
- Docker: Opción de auto-alojamiento (self-hosting) al ser código abierto.

Open source
Helicone es de código abierto bajo la licencia Apache 2.0. El código fuente está disponible para auditoría, modificación y contribución por parte de la comunidad, lo que evita el bloqueo con un solo proveedor (vendor lock-in).

Integraciones
- Facilidad de integración: Nivel "Low-code" para la instalación básica (cambio de URL) y "Full-code" para funciones avanzadas como propiedades personalizadas.
- API propia: Dispone de API para extraer datos de logs y consumirlos en herramientas externas de BI.
- Integraciones nativas: Compatible con OpenAI, Anthropic, Azure, Google Vertex AI, Groq, OpenRouter y marcos de trabajo como LangChain o LiteLLM.
- Conectividad: Capacidad de enviar logs a través de webhooks para automatizar flujos de trabajo en otras plataformas.

Notas finales
información legal, licencias , contratos
Helicone cumple con normativas de seguridad empresarial exigentes en sus planes superiores, incluyendo certificaciones SOC-2 Type II y cumplimiento de la normativa HIPAA para datos de salud. Recientemente ha pasado a formar parte de la suite de herramientas de Mintlify.

Para más información:
- Sitio web oficial: https://www.helicone.ai
- Precios: https://www.helicone.ai/pricing
- Documentación técnica: https://docs.helicone.ai
- Github: https://github.com/Helicone/helicone
- Discord: https://discord.gg/helicone

## CONSEJOS DE IMPLANTACIÓN

Aplicación profesional
Helicone está diseñado para empresas tecnológicas (SaaS), departamentos de ingeniería y equipos de producto que operan infraestructuras basadas en IA. Es especialmente relevante para organizaciones que manejan volúmenes significativos de peticiones a LLMs y requieren una gestión unificada de costes, rendimiento y fiabilidad. El presupuesto varía desde una versión gratuita para prototipado hasta planes empresariales superiores a los 799$ mensuales para despliegues con requisitos de cumplimiento legal (SOC-2, HIPAA).

Madurez digital requerida
- Usuarios y equipo: Requiere un equipo de desarrollo familiarizado con el consumo de APIs REST, manejo de SDKs en Python o Node.js y flujos de trabajo de backend. El equipo de producto debe tener capacidad de análisis para interpretar métricas de latencia y tasas de error.
- Empresa y departamentos: La organización debe haber superado la fase de experimentación manual con IA y estar en proceso de integración productiva (producción) de modelos de lenguaje, necesitando mecanismos formales de observabilidad y control de costes.

Plan orientativo de implantación
Pasos necesarios y estimaciones
- Tiempos de despliegue: Entre 1 hora para una integración inicial mediante pasarela (Gateway) y 1-2 semanas para una configuración avanzada con propiedades personalizadas y despliegue on-premise.
- Evaluación inicial: Auditoría del volumen de tokens consumidos mensualmente, identificación de cuellos de botella en la latencia y selección de los proveedores (OpenAI, Anthropic, etc.) a monitorizar.
- Implantación inicial: Configuración de la clave API de Helicone y sustitución de la URL base del proveedor en las variables de entorno de la aplicación.
- Configuración avanzada: Implementación de reglas de almacenamiento en caché para optimizar costes y activación de reintentos automáticos para mejorar la resiliencia del sistema.
- Seguimiento y feedback: Revisión semanal de los informes de costes por usuario y análisis de los logs en el panel de control para detectar prompts ineficientes o errores recurrentes.

Necesidades de formación del equipo
Es fundamental capacitar al equipo técnico en el uso del Helicone Query Language (HQL) para extracciones de datos complejas y en la gestión del versionado de prompts a través del Playground integrado. El personal operativo debe aprender a interpretar los KPIs de rentabilidad por modelo.

Perfiles necesarios
- Perfiles técnicos: Desarrolladores Backend o Ingenieros de Software para la integración del SDK y configuración de middleware.
- Personal externo: Consultores en optimización de costes de IA o expertos en seguridad si se requiere despliegue en infraestructuras críticas.
- Otros: Ingenieros de Prompt para validar la eficacia de los cambios de versión directamente en la plataforma.

Retorno de la inversión (ROI)
- Tiempos: El impacto en la reducción de latencia (gracias al caching) y detección de errores es inmediato tras la activación. La consolidación de ahorro de costes suele ser visible a partir del primer ciclo de facturación mensual.
- Medición y KPIs: Reducción del coste por cada 1.000 tokens, disminución del porcentaje de errores de API (rate limit errors), mejora en el tiempo medio de respuesta y ahorro económico directo derivado de las respuestas servidas desde caché.

Otros
Al ser una herramienta de código abierto (Apache 2.0), permite el auto-alojamiento mediante Docker, lo que resulta crítico para empresas con políticas de privacidad estrictas que no pueden enviar logs a servidores externos. Su compatibilidad con marcos de trabajo como LangChain y LiteLLM facilita la migración hacia arquitecturas de nube múltiple (multi-cloud) sin fricciones.

## PREGUNTAS FRECUENTES

### ¿Qué es Helicone y cuál es su función principal en un stack tecnológico?

Helicone es una plataforma de observabilidad y gateway de código abierto diseñada para aplicaciones que utilizan modelos de lenguaje extenso (LLM). Funciona como un middleware que se sitúa entre la aplicación del desarrollador y los proveedores de IA, permitiendo registrar, analizar y optimizar todas las solicitudes y respuestas de la API en tiempo real.

### ¿Es Helicone una solución de código abierto?

Sí, Helicone es open source bajo la licencia Apache 2.0. Su código fuente es público y está disponible en GitHub, lo que facilita la auditoría de seguridad, la contribución de la comunidad y la posibilidad de realizar un despliegue autogestionado (self-hosting) mediante Docker.

### ¿Cómo garantiza la privacidad y seguridad de los datos?

La plataforma implementa medidas de seguridad de nivel empresarial, contando con certificaciones SOC-2 Type II y cumplimiento de la normativa HIPAA para el manejo de datos sensibles de salud. Al ser de código abierto, permite a las organizaciones auditar cómo se procesan las comunicaciones antes de su implementación.

### ¿Cumple con la normativa española y europea de protección de datos?

Helicone proporciona las herramientas necesarias para el cumplimiento normativo, como el registro y auditoría de interacciones para evitar la fuga de información sensible. Su cumplimiento de estándares internacionales como SOC-2 suele alinearse con los requisitos de debida diligencia técnica exigidos en entornos profesionales bajo el marco del RGPD.

### ¿Cuáles son los costes asociados y existe una versión gratuita?

Dispone de un plan gratuito (Hobby) limitado a 10.000 solicitudes mensuales y 7 días de retención de datos. Los planes profesionales comienzan en 79$ mensuales (Pro) y escalan hasta los 799$ (Team) para funciones avanzadas de cumplimiento y soporte. Para despliegues a medida o en servidores propios, existe una modalidad Enterprise bajo presupuesto.

### ¿Qué nivel de dificultad técnica requiere su implementación?

El nivel técnico es medio. La integración básica es sencilla, ya que solo requiere sustituir la URL base del proveedor de IA por la de Helicon o utilizar su SDK de una línea de código. Sin embargo, requiere conocimientos de desarrollo en lenguajes como Python o Node.js y manejo de variables de entorno para una configuración completa.

### ¿Con qué proveedores de IA es compatible?

Es compatible con más de 100 modelos y proveedores, incluyendo OpenAI, Anthropic, Google Vertex AI, Azure, Groq y OpenRouter. También se integra con frameworks de desarrollo populares como LangChain y LiteLLM.

### ¿Qué beneficios aporta la función de almacenamiento en caché (caching)?

El almacenamiento en caché permite guardar las respuestas de consultas previas. Si la aplicación realiza una solicitud idéntica, Helicone sirve la respuesta almacenada en lugar de llamar de nuevo al proveedor de IA, lo que reduce significativamente los costes de tokens y disminuye la latencia para el usuario final.

### ¿Cómo ayuda Helicone en la depuración y resiliencia de aplicaciones de IA?

Ofrece un sistema de rastreo completo para identificar fallos en las solicitudes, así como herramientas de resiliencia que incluyen reintentos automáticos (retries) y sistemas de fallback, que cambian automáticamente a un modelo secundario si el principal no responde.

### ¿Permite la exportación de datos para procesos de mejora de modelos?

Sí, la plataforma permite crear y exportar conjuntos de datos (datasets) a partir de los logs de interacciones reales. Estos datos son fundamentales para procesos de ajuste fino (fine-tuning) que buscan mejorar la precisión de los modelos en tareas específicas.

## CONTRATOS Y CONDICIONES

Informe técnico descriptivo

Principales recomendaciones
- Priorizar el uso de la región de datos de la Unión Europea (EU datacenter) disponible en su configuración para garantizar que la persistencia de los logs se mantenga dentro del EEE.
- Implementar la función de "Omit Logs" (omisión de registros) para evitar el almacenamiento del cuerpo de las solicitudes y respuestas en los servidores de Helicone si se maneja información altamente sensible o datos personales no anonimizados.
- Utilizar el modo "Asynchronous Logging" si se desea capturar únicamente metadatos (tokens, latencia) sin que el contenido real de la interacción pase por el proxy de Helicone, reduciendo el riesgo de exposición.
- En sectores regulados (salud, banca), evaluar la opción de "Self-hosting" (autohospedaje) mediante Docker o Kubernetes para mantener la soberanía total de los datos dentro de la infraestructura de la empresa.
- Configurar el "Vault" de Helicone para gestionar llaves de API mediante "Proxy Keys", evitando distribuir las credenciales originales de los proveedores de IA (OpenAI, Anthropic, etc.) entre los desarrolladores.

Ley de Inteligencia Artificial (AI Act)
- Helicone actúa como una herramienta de "observabilidad y gobernanza", lo cual facilita el cumplimiento de las obligaciones de registro de logs (logueo) y transparencia que exige la Ley de IA para sistemas de alto riesgo y modelos de propósito general.
- La funcionalidad de "Datasets" permite recoger datos para procesos de reentrenamiento y evaluación, lo cual es clave para la documentación técnica y las pruebas de robustez exigidas por la normativa.

Privacidad y protección de datos
- Responsabilidades: La empresa española actúa como Responsable del Tratamiento, mientras que Helicone (Helicone, Inc.) actúa como Encargado del Tratamiento.
- Ubicación de los datos: Permite elegir almacenamiento en la región UE. Por defecto, Helicone es una empresa con sede en EE. UU.
- Transferencia internacional: Si se utiliza la versión Cloud estándar fuera de la región UE, existe una transferencia internacional de datos a EE. UU. Se requiere verificar la firma de un Acuerdo de Procesamiento de Datos (DPA) que incluya Cláusulas Contractuales Tipo.
- Derechos ARCO: La plataforma permite la eliminación completa de datos del usuario y logs bajo solicitud a privacy@helicone.ai, cumpliendo con el derecho de supresión y acceso.

Propiedad intelectual
- Propiedad de datos: El usuario conserva todos los derechos sobre el contenido (prompts y respuestas) enviado a través del servicio.
- Propiedad del resultado: Helicone no reclama propiedad sobre las salidas de los modelos monitorizados ni sobre el código desarrollado por la empresa cliente para integrar su SDK.
- Licencia de software: La versión de código abierto utiliza la licencia Apache 2.0, lo que permite su uso comercial, modificación y distribución sin regalías.

Usos y prohibiciones
- Usos prohibidos: Prohibido el uso de la plataforma para actividades ilegales, infringir derechos de propiedad intelectual de terceros, o intentar descompilar/atacar la infraestructura de red de Helicone.
- Usos admitidos: Monitorización, almacenamiento en caché, gestión de costes y optimización de flujos de trabajo con modelos de lenguaje de gran escala (LLM) en entornos profesionales.

Seguridad y certificaciones
- Seguridad: Cifrado AES-256 en reposo y TLS 1.3 en tránsito. Dispone de gestión de llaves AEAD en su funcionalidad de Vault.
- Certificaciones: Certificación SOC 2 Type II disponible bajo petición para clientes Enterprise. Cumplimiento declarado de HIPAA para datos de salud en planes superiores.

Otros
- Integración con Mintlify: Helicone ha sido recientemente adquirida o integrada en la suite de Mintlify, lo que puede implicar cambios en la estructura de soporte y gestión de cuentas, aunque mantiene su naturaleza Open Source.

Fuentes consultada:
- Privacidad
- Seguridad y Cumplimiento
- Términos de Servicio
- Documentación de Cumplimiento
- Código Fuente y Licencia

**Para más información y herramientas:**

Explora look4.tools para descubrir las mejores soluciones tecnológicas del mercado.

Inicio Todas las herramientas Categorías