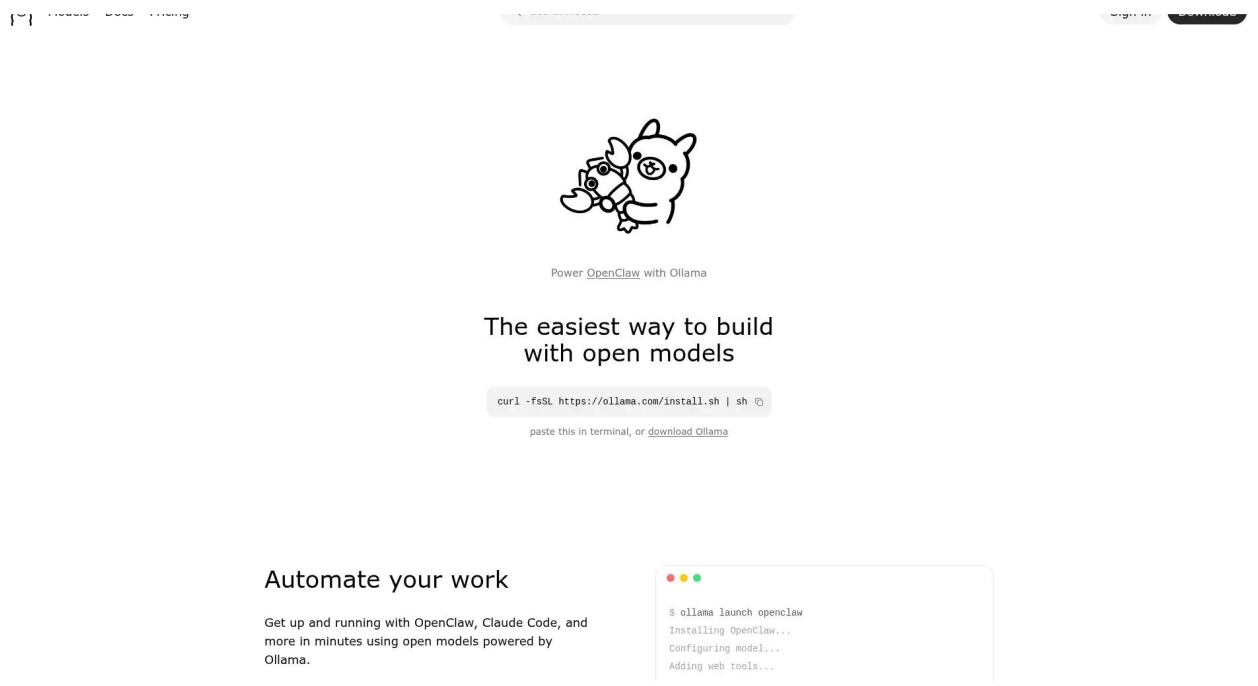

Ollama

Ollama es una herramienta de código abierto diseñada para desarrolladores, ingenieros de software y científicos de datos que necesitan ejecutar modelos de lenguaje de gran tamaño (LLM) de forma local. Permite desplegar infraestructuras de IA privadas y soberanas, eliminando la dependencia de servicios en la nube y garantizando la privacidad de datos sensibles. Es ideal para integrar modelos como Llama 3 o Mistral en flujos de trabajo profesionales mediante una API compatible con OpenAI.

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

Contenido del Dossier

INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es
Ollama es una herramienta diseñada para ejecutar modelos de lenguaje de gran tamaño (LLM) de forma local en el hardware propio del usuario. Está dirigida a desarrolladores, ingenieros de software, científicos de datos y empresas que buscan integrar capacidades de inteligencia artificial (como Llama 3, Mistral o DeepSeek) sin depender de servicios en la nube de terceros, asegurando así la privacidad de los datos y reduciendo costes de latencia o tokens por uso.

Principal ventaja profesional
Permite el despliegue de infraestructuras de IA privadas y soberanas, eliminando la necesidad de enviar información sensible a servidores externos y proporcionando una interfaz compatible con estándares de la industria como la API de OpenAI para una integración inmediata en flujos de trabajo existentes.

Para quién no es
No es adecuado para profesionales sin acceso a hardware con capacidad de cómputo suficiente (especialmente GPUs), ni para aquellos que prefieren delegar la gestión de infraestructura y mantenimiento de modelos en proveedores Managed-SaaS (como OpenAI o Anthropic) por comodidad o falta de personal técnico especializado.

funcionalidades clave
- Orquestación local de modelos: Descarga y ejecución de modelos populares (Llama, Gemma, Mistral, Qwen) con un solo comando.
- API compatible con OpenAI: Facilita la migración de aplicaciones existentes que ya consumen servicios de IA comerciales.
- Gestión de modelos (Modelfiles): Permite crear variantes personalizadas de modelos especificando prompts de sistema y parámetros técnicos.
- Soporte multimodal: Capacidad para procesar imágenes y visión mediante modelos compatibles directamente desde la terminal o API.
- Biblioteca de modelos optimizada: Acceso a un registro con más de 40.000 integraciones de la comunidad y modelos ya cuantizados para mejorar el rendimiento en hardware doméstico o profesional.

Precios
- Versión gratuita: Open Source y gratuita para ejecución en hardware propio (local). Incluye CLI, API y modelos públicos ilimitados.
- Rango de precios: Desde 0€ (autohospedado) hasta 100$/mes por usuario para capacidades Cloud avanzadas.
- Versión Pro (20$/mes): Orientada a profesionales que necesitan ejecutar modelos en la nube de Ollama (Cloud), permitiendo ejecutar hasta 3 modelos simultáneos y compartir modelos privados.
- Versión Max (100$/mes): Para uso intensivo, permite hasta 10 modelos simultáneos en la infraestructura Cloud de Ollama.

Perfil del usuario
- Empresas que gestionan datos sensibles (Sector legal, médico, financiero) que requieren IA privada.
- Departamentos de IT y DevOps que buscan estandarizar el despliegue de IA en servidores internos o estaciones de trabajo.
- Desarrolladores de aplicaciones que necesitan entornos de prueba locales para prototipar agentes o asistentes.
Nivel técnico requerido
- Nivel de uso: Medio. Manejo básico de terminal/CLI.
- Nivel de instalación: Medio-Alto. Requiere configuración de drivers de GPU (NVIDIA CUDA, AMD ROCm o Metal en Mac) y gestión de Docker si se opta por contenedores.
- Conocimientos necesarios: Familiaridad con peticiones REST/API, manejo de terminal y conceptos básicos de modelos de IA (cuantización, contexto).

Ejemplos de uso profesional
- Asistente de codificación local: Integración con VS Code o editores para autocompletado de código sin enviar el código fuente de la empresa a la nube.
- Análisis de documentos privados: Procesamiento de contratos o informes financieros confidenciales en

servidores locales.
- Automatización de soporte técnico: Implementación de chatbots internos entrenados con manuales de producto específicos de la compañía.
- Extracción de entidades y resumen: Procesamiento masivo de logs o correos electrónicos para generar reportes automáticos en local.

Uso y distribución
- Versión web: A través de Ollama Cloud (versiones de pago).
- Versión escritorio: Aplicaciones nativas para Windows (Vista previa), macOS y Linux.
- CLI: Interfaz de línea de comandos robusta para gestión y ejecución.
- Docker: Imagen oficial disponible para despliegues en contenedores.

Open source
Ollama es un proyecto de código abierto distribuido bajo licencia MIT, lo que permite su modificación y uso comercial sin restricciones significativas.

Integraciones
- Facilidad de integración: No code-full code (desde aplicaciones con interfaz hasta desarrollo puro).
- API propia: Dispone de una API REST propia y una capa de compatibilidad con la API de OpenAI.
- Integraciones nativas: Compatible con LangChain, LlamaIndex, Continue.dev, y frameworks de agentes como CrewAI o Autogen.
- Bibliotecas oficiales: Dispone de librerías oficiales para Python y JavaScript/TypeScript.

Notas finales
información legal, licencias , contratos
Ollama utiliza licencias permissivas, pero el uso de los modelos específicos (como Llama 3) está sujeto a las licencias de sus creadores respectivos (Meta, Google, etc.). En las versiones Cloud, los datos de los prompts no se utilizan para entrenamiento y no se registran logs.

Para más información:
- Sitio web oficial: https://ollama.com
- Precios: https://ollama.com/pricing
- Github: https://github.com/ollama/ollama
- Documentación: https://ollama.com/docs
- Discord: https://discord.com/invite/ollama

CONSEJOS DE IMPLANTACIÓN

Informe técnico descriptivo para la implantación de **Ollama** en entornos profesionales.

Aplicación profesional
- **Soberanía de datos:** Ideal para sectores con regulaciones estrictas (legal, salud, defensa) que no pueden enviar datos a nubes de terceros.
- **Optimización de costes:** Eliminación de costes por token en flujos de trabajo de alto volumen (análisis masivo de logs, clasificación de documentos).
- **Desarrollo offline:** Permite prototipar y testear agentes de IA en entornos de desarrollo sin latencia de red ni dependencia de internet.
- **Presupuesto:** El software es Open Source (0€). La inversión se desplaza al hardware (GPU) o infraestructura local.

Madurez digital requerida
- **Usuarios:** Capacidad para interactuar con interfaces de chat o herramientas integradas (extensiones de VS Code, Obsidian, etc.).
- **Equipo Técnico:** Dominio de la línea de comandos (CLI), conocimiento básico de contenedores (Docker) y gestión de drivers de vídeo (NVIDIA CUDA/AMD ROCm).
- **Empresa:** Infraestructura propia capaz de alojar servidores con aceleración gráfica o disposición para renovar el parque de estaciones de trabajo (Apple Silicon o GPUs dedicadas).

Plan orientativo de implantación
Pasos necesarios y estimaciones
- **Evaluación (1-2 días):** Inventario de hardware existente. Determinación de modelos según la tarea (ej. DeepSeek para código, Llama 3 para texto general).
- **Configuración inicial (1 día):** Instalación de Ollama y configuración de variables de entorno críticas como OLLAMA_HOST para acceso en red local y OLLAMA_MAX_LOADED_MODELS para concurrencia.
- **Prueba de concepto (1 semana):** Despliegue de una interfaz amigable (ej. **Open WebUI**) para que usuarios no técnicos validen la calidad de las respuestas.
- **Integración (2 semanas):** Conexión con flujos existentes mediante su API compatible con OpenAI. Configuración de ModelFiles para estandarizar prompts de sistema corporativos.
- **Escalado (Variable):** Paso de ejecución en estaciones de trabajo individuales a servidores centralizados con Docker y orquestación.

Necesidades de formación del equipo
- **Promoting avanzado:** Formación en cómo instruir a los modelos locales (que pueden tener capacidades de razonamiento distintas a GPT-4).
- **Gestión de modelos:** Instrucción sobre la selección de cuantizaciones (Q4_K_M es el estándar recomendado para equilibrio velocidad/calidad).
- **Privacidad:** Protocolos de manejo de información sensible dentro de la red local.

Perfiles necesarios
- **Ingeniero de DevOps/Sistemas:** Para el despliegue de servidores y optimización de recursos GPU.
- **Desarrollador de IA/Software:** Para integrar la API de Ollama en las aplicaciones internas de la empresa.
- **Administrador de IT:** Para la gestión de drivers y mantenimiento del hardware físico.

Retorno de la inversión (ROI)
- **Tiempos:** Reducción inmediata de la latencia en tareas automatizadas.
- **KPIs:** Ahorro mensual en suscripciones SaaS, volumen de tokens procesados localmente vs coste equivalente en nube, y reducción de riesgos por brechas de seguridad de datos.

Otros
- **Optimización de hardware:** En Mac (Apple Silicon), Ollama utiliza memoria unificada, lo que permite ejecutar modelos de gran tamaño (70B) que requerirían GPUs profesionales muy costosas en PC.
- **Concurrencia:** Por defecto, Ollama procesa peticiones de forma secuencial. En entornos con múltiples usuarios, es vital configurar OLLAMA_NUM_PARALLEL para habilitar el procesamiento simultáneo según la VRAM disponible.
- **Modelos recomendados (2025-2026):**

- **Llama 3.3 (70B):** Para razonamiento complejo y alta calidad.
- **Mistral Nemo (12B):** Excelente equilibrio para asistentes de propósito general.
- **DeepSeek-Coder-V2:** Especializado en programación y tareas lógicas.
- **Phi-4 (14B):** Alta eficiencia en hardware con menos recursos.

PREGUNTAS FRECUENTES

¿Qué es Ollama y para qué sirve en un entorno profesional?

Ollama es una herramienta de código abierto diseñada para ejecutar modelos de lenguaje de gran tamaño (LLM) de forma local en el hardware del usuario. Su función principal es permitir la implementación de capacidades de inteligencia artificial sin depender de servicios en la nube, facilitando la privacidad de los datos, la reducción de latencia y el ahorro en costes de tokens.

¿Qué coste tiene el uso de esta tecnología?

Ollama es gratuito y de código abierto para su ejecución en hardware propio (local). No obstante, ofrece planes de pago para su infraestructura en la nube (Ollama Cloud): la versión Pro por 20$/mes para profesionales que requieren modelos compartidos y ejecución simultánea, y la versión Max por 100$/mes para uso intensivo empresarial.

¿Es Ollama un proyecto Open Source?

Sí, el proyecto se distribuye bajo la licencia MIT, lo que permite su modificación, integración y uso comercial con restricciones mínimas. Es posible acceder a su código fuente y descargarlo directamente desde su repositorio oficial en GitHub.

¿Cómo aborda Ollama la privacidad y la seguridad de los datos?

La principal ventaja de Ollama es la soberanía de los datos. Al ejecutar los modelos localmente, la información sensible no se envía a servidores externos. En sus versiones Cloud, la empresa garantiza que los prompts no se utilizan para el entrenamiento de modelos y no se registran logs de actividad.

¿Cumple con la normativa española de protección de datos (RGPD)?

Al permitir el procesamiento local (On-Premise), Ollama facilita que las empresas cumplan con el RGPD y otras regulaciones estrictas, ya que los datos personales o confidenciales nunca salen de la infraestructura controlada por la organización.

¿Qué requisitos de hardware son necesarios para un uso profesional?

Se requiere hardware con capacidad de cómputo suficiente, preferiblemente GPUs con soporte para NVIDIA CUDA, AMD ROCm o la arquitectura Metal en dispositivos Apple. No se recomienda para profesionales que carezcan de estaciones de trabajo con aceleración gráfica o servidores dedicados.

¿Es compatible con otras herramientas de inteligencia artificial?

Sí, Ollama ofrece una API compatible con el estándar de OpenAI, lo que facilita la migración de aplicaciones existentes. Además, se integra de forma nativa con frameworks populares como LangChain, LlamaIndex, Continue.dev, CrewAI y bibliotecas oficiales para Python y JavaScript.

¿Qué modelos de lenguaje se pueden ejecutar?

Permite orquestar una amplia biblioteca de modelos optimizados, incluyendo Llama 3, Mistral, Gemma, Qwen y DeepSeek, además de modelos multimodales con capacidades de visión. El usuario también puede crear variantes personalizadas mediante archivos de configuración (Modelfiles).

¿Cuál es el nivel técnico necesario para su implementación?

El nivel técnico es medio-alto. Aunque el uso básico se realiza mediante comandos sencillos (CLI), la instalación profesional requiere conocimientos en configuración de drivers de GPU, gestión de contenedores Docker y manejo de peticiones REST/API.

¿Existen restricciones legales sobre los modelos utilizados?

Aunque el software Ollama es MIT, cada modelo específico (como Llama 3 o Gemma) está sujeto a la licencia de su creador (Meta, Google, etc.). Los profesionales deben revisar los términos de uso de cada modelo individual para asegurar el cumplimiento legal según el caso de uso.

CONTRATOS Y CONDICIONES

Principales recomendaciones
- Diferenciar claramente el flujo de datos: Al usar Ollama en modo local, la información no sale de su infraestructura (servidor o puesto de trabajo). Si se activan funciones "Cloud" o modelos con sufijo -cloud, los datos se enviarán a servidores de Ollama Inc.
- Revisar la licencia de cada modelo descargado: Ollama es un mediador. Aunque la herramienta es open source (MIT), cada modelo (Llama 3 de Meta, Gemma de Google, etc.) tiene términos de uso específicos que pueden restringir usos comerciales o requerir atribución.
- Implementar control de accesos: La API de Ollama por defecto no incluye autenticación robusta si se expone en red local. Se recomienda usar un proxy inverso o VPN para proteger el punto de acceso (endpoint) de la API.
- Desactivar funciones de nube si el cumplimiento es crítico: Para garantizar el aislamiento total, configure la variable de entorno OLLAMA_NO_CLOUD=1.

Ley de Inteligencia Artificial (AI Act)
- Clasificación de riesgo: El software Ollama en sí se considera un facilitador de modelos de propósito general (GPAI). El cumplimiento recae en la empresa española que implementa el modelo específico para un uso concreto (ej. selección de personal, análisis de solvencia).
- Transparencia: Al integrar modelos mediante Ollama en aplicaciones de cara al cliente, la empresa debe informar explícitamente de que se está interactuando con una IA, conforme a las obligaciones de transparencia del AI Act.
- Documentación técnica: Las empresas que utilicen Ollama para desarrollar sistemas de IA propios deben mantener documentación técnica actualizada sobre las capacidades y limitaciones de los modelos desplegados (especialmente si son de "alto riesgo").

Privacidad y protección de datos (RGPD)
- Responsabilidades: En el uso local (on-premise), la empresa española es el único Responsable del Tratamiento. Ollama Inc. no actúa como encargado ya que no tiene acceso a la ejecución. En el uso de Ollama Cloud, Ollama Inc. actúa como Encargado del Tratamiento.
- Ubicación de los datos: En modo local, los datos permanecen en España/UE. En modo Cloud, Ollama declara que los datos se procesan principalmente en Estados Unidos, con posible enrutamiento a Europa o Singapur.
- Transferencia internacional: El uso de Ollama Cloud implica una transferencia internacional de datos a EE.UU. Es necesario verificar que el tratamiento esté amparado por el Marco de Privacidad de Datos (Data Privacy Framework) o Cláusulas Contractuales Tipo.
- Derechos ARCO: La empresa debe garantizar que puede atender solicitudes de acceso o supresión de datos que hayan sido procesados por los modelos locales, asegurando que no queden rastros en logs de depuración si contienen datos personales.

Propiedad intelectual
- Propiedad de los datos: Usted mantiene la propiedad total de los datos de entrada (inputs) que procesa a través de la herramienta.
- Propiedad del resultado: Ollama no reclama derechos sobre el contenido generado (outputs). Sin embargo, la protección por derecho de autor de obras generadas íntegramente por IA es limitada bajo la legislación española actual.
- Licencia de software: Ollama utiliza la licencia MIT, una de las más permisivas, que permite el uso comercial, modificación y distribución siempre que se incluya el aviso de copyright original.

Usos y prohibiciones
- Usos prohibidos: No se debe utilizar para actividades ilegales, vulnerar propiedad intelectual de terceros o realizar ingeniería inversa de modelos protegidos. Se prohíbe el uso de la versión Cloud para desarrollar productos que compitan directamente con Ollama.
- Usos admitidos: Desarrollo de aplicaciones internas, automatización de procesos empresariales, análisis de documentos confidenciales (en modo local) y prototipado rápido de agentes de IA.

Seguridad y certificaciones
- Seguridad: Al funcionar de forma local, la responsabilidad de la seguridad perimetral, parches de sistema y cifrado de discos recae íntegramente en el departamento de IT de la empresa española.

- Vulnerabilidades conocidas: Se han reportado riesgos en el flujo de autenticación del registro de modelos (exfiltración de tokens) que requieren mantener la herramienta siempre actualizada a la última versión oficial.
- Certificaciones: Ollama Inc. no muestra certificaciones ISO 27001 o SOC2 públicas vinculadas específicamente a su software local; su seguridad depende de la infraestructura donde se aloje.

Otros
- Limitación de responsabilidad: Los términos de Ollama limitan su responsabilidad legal a un máximo de 100$ o lo pagado en los últimos 12 meses. Para una empresa española, esto implica que cualquier daño derivado de un mal funcionamiento de la IA debe estar cubierto por un seguro de responsabilidad civil propio.

Fuentes consultadas:
- Condiciones de servicio (Terms of Service)
- Política de privacidad (Privacy Policy)
- Precios y límites de uso
- Licencia MIT (Github)
- Documentación técnica de Cloud