DeepSeek

El LLM Open-Source que planta cara a ChatGPT

Autor: Leire Ahedo

Contenido

Introducción al Curso

DeepSeek se presenta como una nueva generación de modelos de lenguaje de código abierto, diseñado para competir de tú a tú con soluciones propietarias como GPT-4 o Claude. Este curso ofrece una inmersión completa y guiada en su potencial, con un enfoque eminentemente práctico y estratégico, especialmente orientado al mundo empresarial.

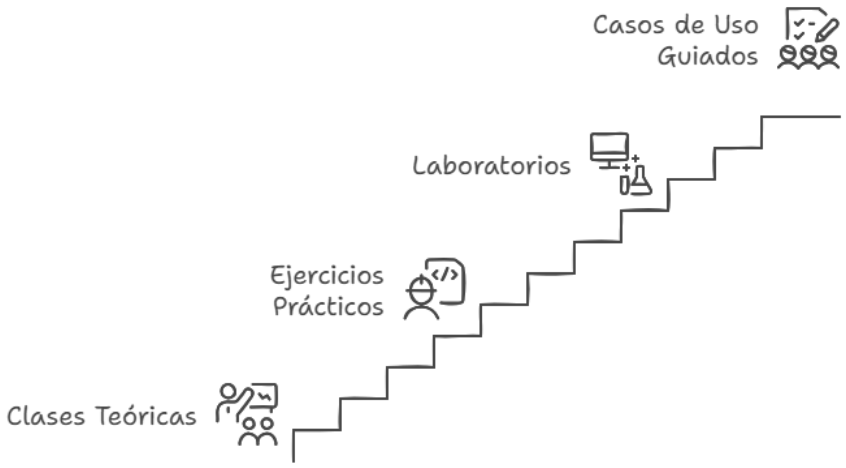
El objetivo principal es capacitar a los participantes para que no solo entiendan cómo funciona DeepSeek, sino que sean capaces de **transformar una necesidad o reto de negocio en una solución concreta basada en IA generativa**. Para ello, se estructura el aprendizaje en torno a casos de uso reales, herramientas accesibles, ejercicios guiados y una hoja de ruta clara.

Metodología de Formación

El curso combina teoría y práctica de manera equilibrada. Comienza con una sólida base conceptual sobre modelos de lenguaje, razonamiento y arquitecturas open-source, y avanza hacia el desarrollo de soluciones concretas.

Se trabaja con:

- **Clases teóricas estructuradas**, orientadas a entender cómo funciona DeepSeek desde el punto de vista técnico y estratégico.
- **Ejercicios prácticos**, donde se pondrán en marcha entornos locales, APIs y flujos de trabajo reales.
- **Laboratorios**, que permiten desplegar y probar aplicaciones empresariales directamente con los modelos.
- **Casos de uso guiados**, donde se verá cómo resolver retos empresariales con DeepSeek, paso a paso.



Qué es DeepSeek

DeepSeek es una familia de modelos de lenguaje desarrollada por High-Flyer, un fondo de inversión chino que decidió apostar por una solución de alto rendimiento, abierta y escalable.

A diferencia de los grandes modelos cerrados, DeepSeek ha sido diseñado para:

- Ser **eficiente en costes y consumo de recursos**.
- Tener **alta capacidad de razonamiento**.
- Ser **fácilmente ejecutable en local**, sin necesidad de depender de la nube o servicios de terceros.
- Contar con **licencia MIT**, lo que facilita su adopción incluso en entornos sensibles.

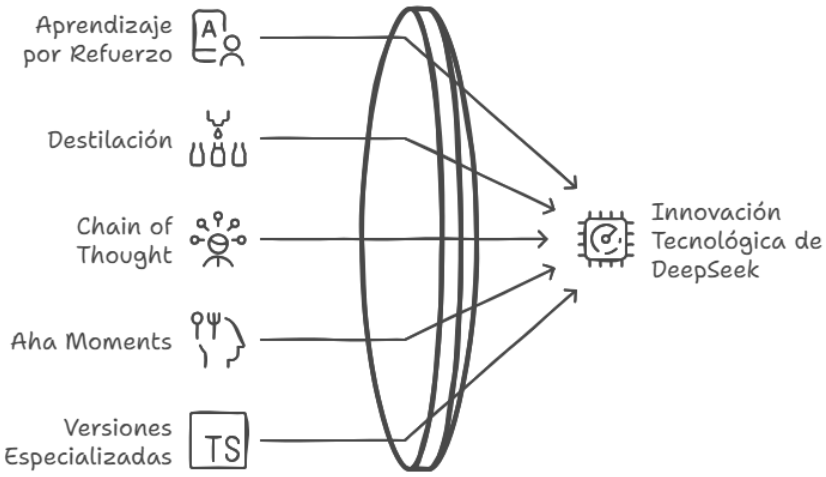
Es una propuesta estratégica que pone al alcance de empresas, centros educativos y desarrolladores capacidades de IA que antes solo estaban disponibles a través de grandes plataformas comerciales.

Enfoque Técnico

DeepSeek destaca por integrar tecnologías avanzadas que marcan la diferencia en términos de precisión y rendimiento:

- **Aprendizaje por Refuerzo (RL):** permite que el modelo aprenda de forma más natural, sin depender exclusivamente de datos etiquetados. Esto favorece la mejora del razonamiento.
- **Destilación:** transfiere capacidades de modelos grandes a modelos pequeños, lo que permite desplegarlos con bajo coste.
- **Chain of Thought:** técnica de razonamiento paso a paso que mejora significativamente la capacidad del modelo para resolver problemas complejos de forma lógica y transparente.
- **Aha Moments:** momentos de autocorrección del modelo, donde puede detectar errores en su propio razonamiento y corregirse de forma autónoma.

Además, se incluyen múltiples versiones especializadas: DeepSeek-R1, DeepSeek Coder, DeepSeek Math, DeepSeek VL (visión-lenguaje), adaptadas a tareas concretas.



Impacto en la Industria

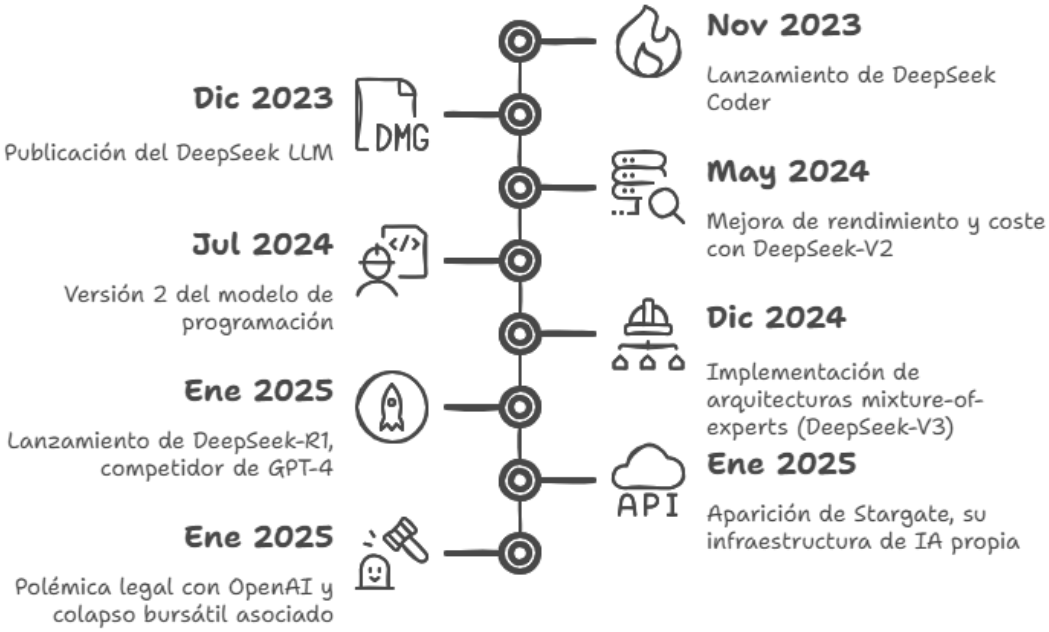
El crecimiento de DeepSeek ha tenido efectos profundos en el sector:

- Está desafiando la hegemonía de empresas como OpenAI o Anthropic.
- Su modelo de desarrollo prueba que **es posible innovar con menos recursos**, algo clave para países y organizaciones sin acceso a grandes infraestructuras.
- Ha provocado **reacciones legales**, como la demanda de OpenAI, preocupada por el uso de sus datasets.
- Su éxito pone de relieve el potencial tecnológico de China, incluso bajo restricciones de acceso a chips.

En paralelo, está influyendo en la evolución del mercado de hardware, especialmente en el dominio que Nvidia tenía hasta ahora con sus GPU.

Línea de Tiempo y Evolución del Proyecto

- **Nov 2023**: lanzamiento de DeepSeek Coder
- **Dic 2023**: publicación del primer modelo generalista (DeepSeek LLM)
- **May 2024**: mejora de rendimiento y coste con DeepSeek-V2
- **Jul 2024**: versión 2 del modelo para programación
- **Dic 2024**: implementación de arquitecturas mixture-of-experts (DeepSeek-V3)
- **Ene 2025**: lanzamiento de DeepSeek-R1, competidor directo de GPT-4
- **Ene 2025**: aparición de Stargate, su infraestructura de IA propia
- **Ene 2025**: polémica legal con OpenAI y colapso bursátil asociado



Aplicaciones Empresariales

DeepSeek puede usarse en infinidad de contextos empresariales, como por ejemplo:

- **Asistentes virtuales** con memoria, multicanal y multilingües.
- **Clasificación y análisis de documentos**, contratos o correos.
- **Resumen de grandes volúmenes de texto** o bases de datos internas.
- **Automatización de procesos**, como creación de informes, extracción de datos o detección de duplicados.
- **Análisis de sentimiento**, útil en encuestas o redes sociales.
- **Agentes conectados a APIs**, capaces de actuar, buscar información, y tomar decisiones.

Estas aplicaciones pueden ejecutarse tanto en la nube como **en servidores locales**, garantizando control y privacidad.

Integración Tecnológica

1. Alojarlo en tu ordenador: Instalas y ejecutas el modelo en tu PC con una GPU potente, permitiéndote procesar consultas localmente sin depender de terceros. Es ideal para pruebas, desarrollo personal y proyectos que requieren privacidad, pero limitado por el hardware disponible.

2. Alojarlo en un servidor propio (On-Premises): Montas un servidor con varias GPUs dedicadas para alojar y operar el modelo dentro de tu infraestructura, asegurando privacidad y personalización total. Es costoso y requiere mantenimiento técnico, pero permite un uso intensivo sin depender de proveedores externos.

3. Desplegarlo en la nube (Cloud): Utilizas servicios como AWS, GCP o Azure para correr el modelo en instancias con GPUs sin necesidad de comprar hardware. Es altamente escalable y flexible, pero implica costos recurrentes y dependencia del proveedor.

4. Consumirlo desde una API de terceros: Accedes al modelo a través de una API como DeepSeek Cloud o OpenAI, pagando por cada consulta sin preocuparte por infraestructura. Es la opción más sencilla y rápida, pero menos flexible y más costosa a largo plazo.

DeepSeek puede integrarse mediante:

- **Ollama o LMStudio** para ejecutarlo en local, sin necesidad de GPUs de alto coste.
- **Vía API REST**, desde plataformas como Hugging Face, OctoAI o directamente desde DeepSeek.
- **Frameworks como LangChain, Flowise o LlamaIndex**, que permiten crear flujos conversacionales, agentes, herramientas de recuperación de información y más.

Esto permite adaptar DeepSeek al nivel técnico de cada empresa, desde soluciones no-code hasta entornos avanzados de programación.

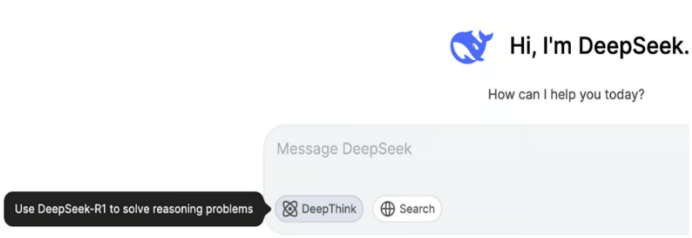
Acceso a través de API

DeepSeek puede utilizarse fácilmente a través de una **API REST**, que permite integrarlo en flujos y aplicaciones propias sin necesidad de infraestructura compleja.

DeepSeek-R1 está disponible a través de dos métodos: plataforma web y API.

Plataforma web:
Accede mediante **DeepSeek Chat**, registrándote y activando el modo **Deep Think** para razonamiento paso a paso

API:
Integra **DeepSeek-R1** en aplicaciones con su **API compatible con OpenAI**, obteniendo una clave de acceso tras el registro en la plataforma.



MODEL	CONTEXT LENGTH	MAX COT TOKENS	MAX OUTPUT TOKENS	1M TOKENS INPUT PRICE (CACHE HIT)	1M TOKENS INPUT PRICE (CACHE MISS)	1M TOKENS OUTPUT PRICE
deepseek-chat	64K	-	8K	$0.07 $0.014	$0.27 $0.14	$1.10 $0.28
deepseek-reasoner	64K	32K	8K	$0.14	$0.55	$2.19

Esto es especialmente útil para:

- Sistemas de atención al cliente
- Interfaces conversacionales personalizadas
- Automatización de tareas con herramientas low-code
- Conexión con dashboards y gestores documentales

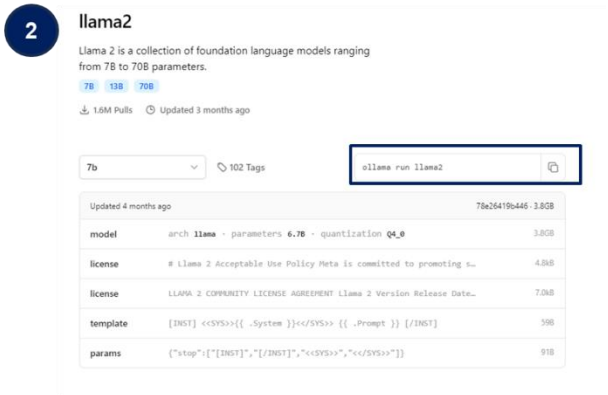
Actualmente, la API está disponible de forma directa a través del equipo de DeepSeek, o vía plataformas como **Replicate, OctoAI o Hugging Face**, lo que abre la puerta a una implementación rápida y escalable sin necesidad de gestión de infraestructura.

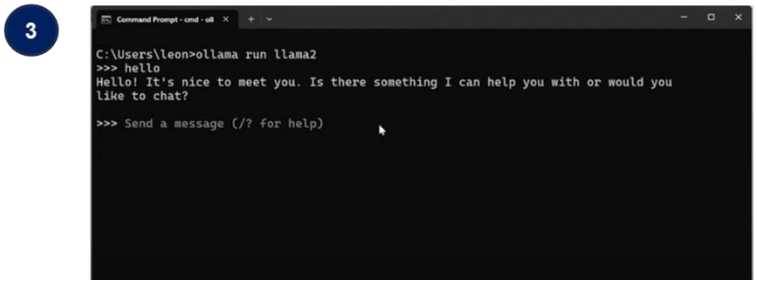
Además, su compatibilidad con JSON y estándar HTTP hace que sea integrable con prácticamente cualquier lenguaje o stack tecnológico.

DeepSeek privado y en local con Ollama

Una de las fortalezas más valoradas de DeepSeek es su capacidad de ejecutarse **en local**, sin necesidad de depender de servicios en la nube. Esto se puede hacer gracias a herramientas como **Ollama**, que permite cargar modelos LLM de forma rápida, segura y con soporte para variantes optimizadas.

Instalación de Ollama





Con Ollama, se puede:

- Desplegar DeepSeek-R1 y sus variantes directamente en el portátil o servidor de la empresa.
- Integrarlo con workflows propios sin exponer datos al exterior.
- Habilitar entornos de desarrollo o producción offline.
- Experimentar con seguridad jurídica y control total del entorno.

Esto es especialmente útil en sectores legales, sanitarios o industriales, donde la privacidad es una prioridad absoluta.

Desarrollo de soluciones empresariales con DeepSeek

DeepSeek se adapta con facilidad a desarrollos empresariales personalizados. A través de frameworks como LangChain, Flowise o directamente con scripts en Python, es posible construir:

- **Chatbots especializados**, conectados a bases de datos internas o documentos de la empresa.
- **Clasificadores de entrada de tickets**, correos o formularios.

- **Sistemas de decisión semántica**, que ayudan a interpretar, comparar y argumentar decisiones.
- **Agentes autónomos**, capaces de interactuar con APIs, tomar decisiones o ejecutar tareas específicas de negocio.

Además, la comunidad ha desarrollado un amplio conjunto de wrappers, integraciones y herramientas que permiten extender DeepSeek según las necesidades de cada empresa.

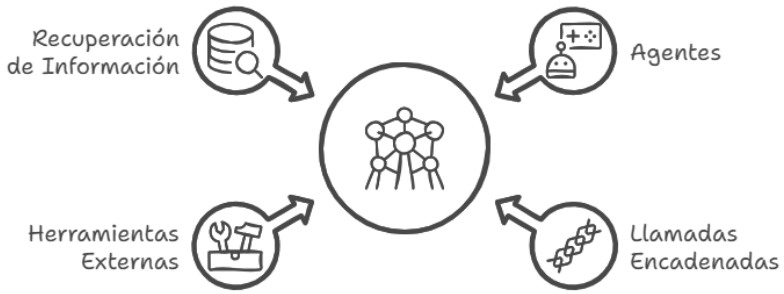
Orquestación de Soluciones con LangChain y Flowise

Aunque DeepSeek puede utilizarse directamente vía API o localmente con Ollama, su verdadero potencial se multiplica al integrarlo dentro de **entornos orquestados y componibles** como LangChain y Flowise. Estas herramientas permiten crear **sistemas conversacionales y agentes inteligentes** con lógica, memoria, razonamiento y acceso a múltiples fuentes externas.

LangChain: La herramienta base para flujos conversacionales complejos

LangChain es un framework de código abierto pensado para construir aplicaciones impulsadas por modelos de lenguaje, combinando elementos como:

- Agentes que toman decisiones
- Llamadas encadenadas con memoria de contexto
- Acceso a herramientas externas (búsqueda, bases de datos, APIs…)
- Recuperación de información (RAG) con embeddings y bases vectoriales



Con LangChain y DeepSeek puedes construir:

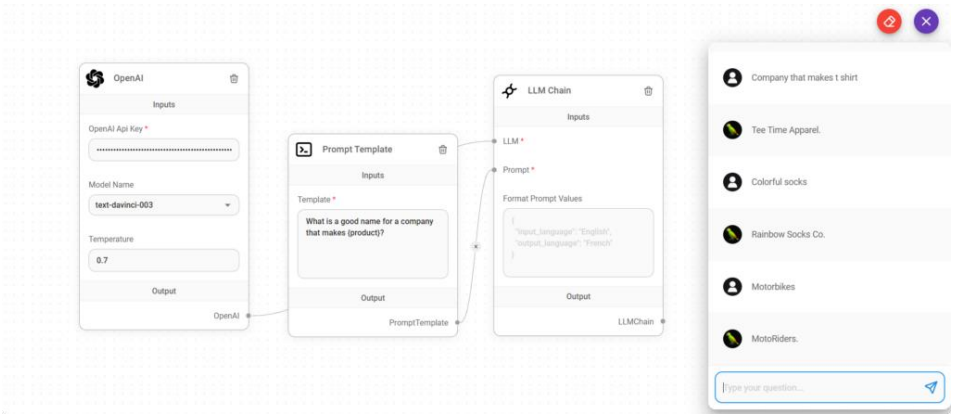
- Sistemas de atención jurídica con recuperación de legislación interna
- Agentes de soporte que interactúan con APIs de ticketing (como Jira o Zendesk)
- Flujos conversacionales adaptados al contexto del usuario
- Procesos complejos que implican validación, consultas externas, y respuestas condicionadas

LangChain es especialmente potente cuando se combina con **LangGraph**, que permite implementar flujos asincrónicos, decisiones condicionales, y loops inteligentes, ideales para el desarrollo de *agentic RAGs*.

Flowise: Orquestación visual sin código

Flowise es una alternativa visual y accesible a LangChain, que permite construir soluciones conversacionales o de RAG a través de una interfaz gráfica basada en nodos. Es ideal para equipos que:

- No tienen perfiles técnicos avanzados
- Necesitan prototipar rápidamente
- Quieren experimentar con prompts, modelos y embeddings sin programar



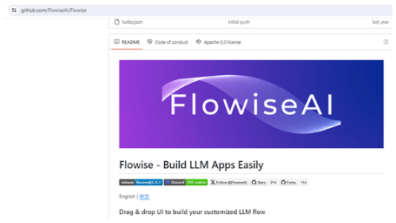
Con Flowise puedes:

- Crear asistentes personalizados en minutos
- Conectar DeepSeek con bases vectoriales, herramientas de scraping o sistemas internos
- Ajustar interacciones conversacionales de forma visual
- Exportar proyectos para integrarlos en apps, webs o CRMs

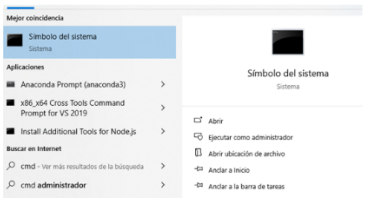
Flowise funciona directamente con modelos locales (como los cargados con Ollama), pero también permite integrar modelos desplegados en servidores o en la nube.

Instalación de Flowise

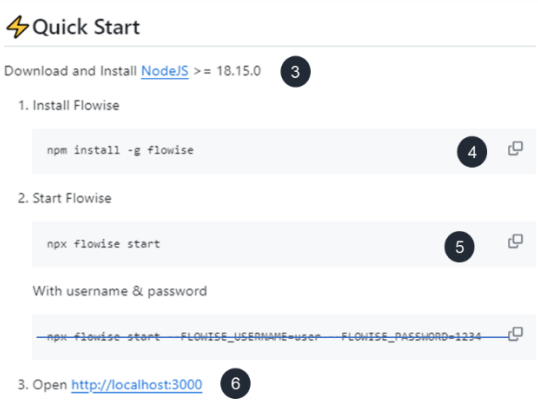
1. Acceder al **GitHub**: https://github.com/FlowiseAI/Flowise



2. Acceder a **CMD**



3. Seguir el **Quick Start**



¿LangChain o Flowise?

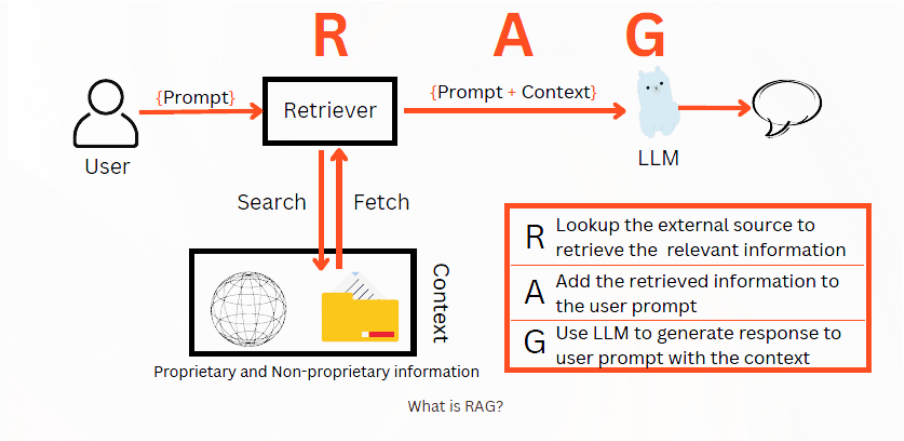
- Usa **LangChain** si tienes un equipo técnico y necesitas flexibilidad total, integración compleja o lógica personalizada.
- Usa **Flowise** si buscas velocidad, facilidad o acceso no técnico a flujos de trabajo conversacionales.

Ambas herramientas son **complementarias a DeepSeek** y forman parte del ecosistema habitual de desarrollo de aplicaciones inteligentes con LLMs.

DeepSeek con acceso a información empresarial

Uno de los escenarios más comunes en entornos corporativos es la necesidad de usar un modelo LLM sobre información interna, privada o estructurada. DeepSeek permite esto gracias a su compatibilidad con sistemas de recuperación semántica como:

- **RAG (Retrieval-Augmented Generation)**
- **Bases vectoriales** como Chroma, Weaviate, FAISS o Qdrant
- **Document loaders** para PDFs, Word, bases SQL o APIs internas



Esto significa que se puede construir un asistente o sistema de análisis que:

- Acceda a documentos internos (contratos, normativas, actas…)
- Interactúe con bases de datos empresariales
- Busque información relevante y responda con precisión
- Genere respuestas personalizadas con contexto empresarial

Este tipo de solución es ideal para departamentos legales, recursos humanos, atención al cliente o soporte técnico.

Componentes del RAG en Flowise

En Flowise, un flujo RAG se construye a partir de distintos bloques funcionales que se conectan entre sí. A continuación, se describen los componentes que aparecen en el flujo del esquema mostrado:

📄 **Text File (Extraer datos)**

Permite **cargar documentos** que serán utilizados como fuente de conocimiento. Es el punto de entrada del contenido, que se puede subir directamente desde un archivo local.

Character Text Splitter (Splitter)

Divide el documento en fragmentos (chunks) para facilitar la indexación y recuperación posterior.

- Parámetros ajustables como el **tamaño del chunk** y el **solapamiento** permiten controlar cómo se segmenta el texto.
- Esto mejora la precisión en la fase de recuperación de contexto.

OpenAI Embeddings (Embeddings)

Transforma cada fragmento de texto en un **vector numérico** mediante un modelo de embeddings (en este caso, de OpenAI).

- Este paso es esencial para que el contenido pueda buscarse de forma semántica.
- Aunque aquí se utiliza OpenAI, puede sustituirse por cualquier motor de embeddings compatible con Flowise (incluido uno basado en DeepSeek si está disponible).

In-Memory Vector Store (Vector Store)

Almacena los vectores generados y permite realizar búsquedas semánticas.

- El nodo actúa como una **base de datos temporal** que se consultará más adelante durante la conversación.
- Puede reemplazarse por soluciones más escalables como FAISS, Qdrant, Weaviate, etc.

ChatDeepSeek (LLM)

Este nodo configura el modelo de lenguaje principal que generará las respuestas.

- En este caso, se utiliza **DeepSeek Chat**, especificando la credencial y la temperatura.
- Este modelo tomará la información recuperada y generará la respuesta final.

Conversational Retrieval QA Chain (Cadena)

Es el componente que **conecta todo**:

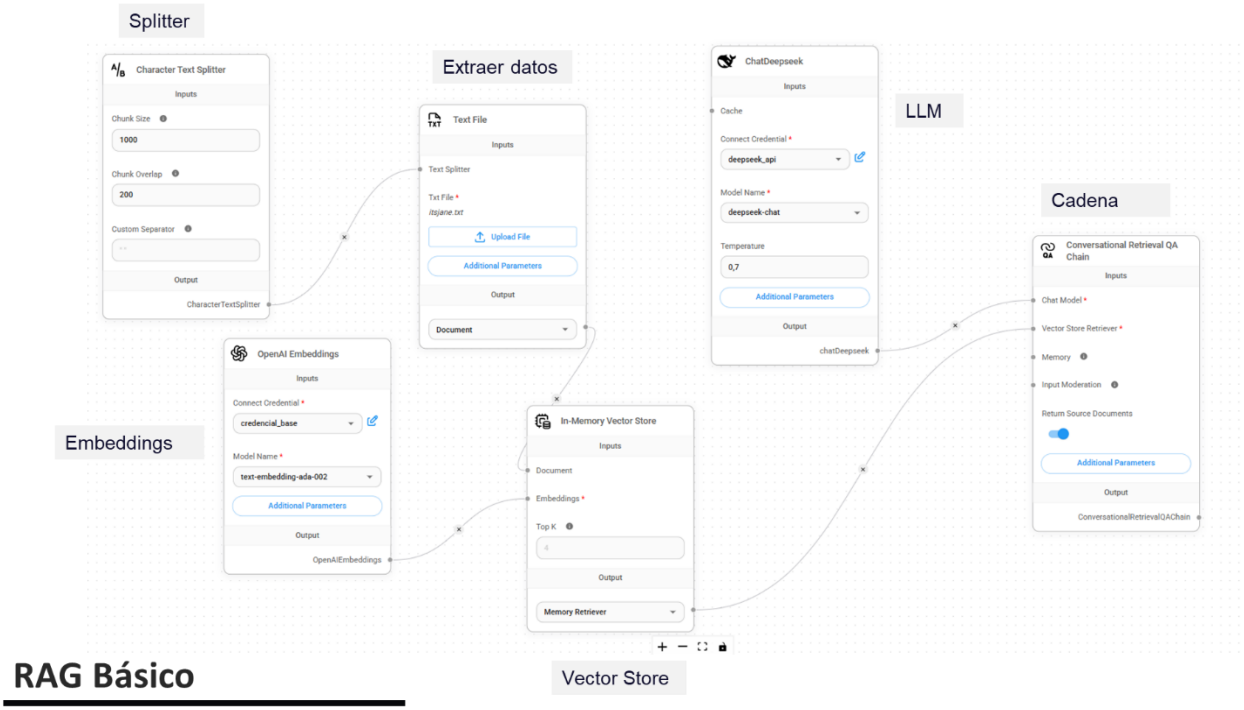
- Combina el modelo DeepSeek, la base vectorial y la memoria conversacional.
- Procesa la pregunta del usuario, realiza la búsqueda en los documentos y genera una respuesta fundamentada.
- También puede devolver los documentos utilizados como fuente.

Este conjunto de componentes conforma la base del flujo de trabajo de un RAG en Flowise, permitiendo construir asistentes conversacionales que integran recuperación semántica con generación de lenguaje natural, de forma modular, visual y sin necesidad de escribir código.

Función del flujo

Este RAG básico permite que un usuario haga preguntas sobre un documento que ha cargado previamente. El sistema:

1. Divide el documento en partes.
2. Calcula los embeddings de cada parte.
3. Almacena los vectores en memoria.
4. Usa DeepSeek para responder consultas.
5. Recupera los fragmentos relevantes para apoyar la respuesta generada.



Seguridad, Riesgos y Consideraciones Éticas

Aunque potente, DeepSeek no está exento de riesgos:

- Su uso de datos de entrenamiento podría tener **implicaciones legales**, especialmente si se reutilizan fragmentos protegidos.
- El modelo puede responder con **censura ideológica** dependiendo del idioma y tema.
- Existen riesgos técnicos como **prompt injection**, loops infinitos, o **jailbreaks**.
- Es importante evitar el uso de trust_remote_code salvo en entornos 100% controlados.

Estos factores deben considerarse especialmente en sectores regulados o donde se maneje información sensible.

Casos Prácticos del Curso

Durante la formación se desarrollan casos como:

- **Desarrollo de un asistente jurídico privado**, utilizando RAG, embeddings y base vectorial.
- **Pipeline de recuperación semántica con documentos empresariales**.
- **Moderación de contenidos mediante agentes especializados**.
- **Evaluación de precisión con datasets legales como CaseHOLD**.

Conclusión

DeepSeek no es solo una alternativa técnica. Es una herramienta estratégica para cualquier empresa que quiera:

- Tener control total sobre su modelo de IA.
- Reducir costes operativos.
- Evitar dependencias con terceros.
- Adaptar el comportamiento del modelo a sus valores y necesidades reales.

Gracias a su enfoque abierto, su compatibilidad con entornos locales y su rendimiento destacado en tareas de razonamiento, **DeepSeek representa una opción de alto valor para sectores legales, educativos, industriales y de innovación**.

Con esta formación, cualquier organización podrá pasar del *problema al prototipo funcional de IA generativa* con criterio, visión y autonomía.