

Arquitectura del Agente de IA Conversacional

Diseño modular y escalable para interacciones inteligentes



Visión General de la Arquitectura

El Agente de IA Conversacional está diseñado con una arquitectura modular de tres capas que separa claramente las responsabilidades.



Modularidad

Componentes independientes que pueden actualizarse sin afectar al resto.

Escalabilidad

Diseño que permite crecer para manejar mayor carga de usuarios.

Mantenibilidad

Clara separación de responsabilidades para facilitar el desarrollo.

Capa 1: Interfaz de Usuario (Frontend)

La capa de interfaz de usuario es la cara visible del agente, proporcionando una experiencia intuitiva y responsiva para los usuarios.



HTML5






CSS3



JavaScript

Componentes Principales

-  Chat UI: Interfaz principal para consultas y respuestas
-  Dashboard: Gestión de base de conocimiento y análisis
-  Configuración: Personalización del agente

Agente de IA Conversacional

Chat
Base de
Conocimiento
Análisis
Configuración

Escribe tu pregunta aquí...

Capa 2: Lógica de Negocio (Backend)

El backend actúa como el cerebro del agente, orquestando las interacciones entre la interfaz de usuario y los servicios de IA/datos.



Python 3.11



FastAPI



Uvicorn

Componentes Principales

- ↔ API RESTful: Endpoints para todas las funcionalidades
- ☑ Manejadores de Solicitudes: Procesamiento y validación
- 🔌 Coordinación: Orquestación entre servicios

API RESTful

POST	/api/ask	Procesa consultas de usuario
GET	/api/knowledge	Obtiene información de la base de conocimiento
POST	/api/knowledge	Añade información a la base de conocimiento
GET	/api/analytics	Obtiene estadísticas de uso
GET	/health	Verifica el estado del sistema

Capa 3: Servicios de IA y Datos

Esta capa contiene los módulos inteligentes y la persistencia de datos que potencian las capacidades del agente. Es el corazón de la inteligencia del sistema.



Servicio GenAI

Interactúa con modelos de lenguaje grandes (LLMs) para generar respuestas coherentes y contextualizadas a las consultas de los usuarios.

Tecnología: OpenAI API, modo simulado para demo



Servicio de Embeddings

Transforma el texto en representaciones numéricas (embeddings) que permiten la búsqueda semántica y la comparación de similitudes.

Tecnología: sentence-transformers, all-MiniLM-L6-v2



Base de Conocimiento

Gestiona el almacenamiento y la recuperación de información estructurada, permitiendo una recuperación eficiente de información relevante.

Tecnología: SQLite, vectores de embeddings



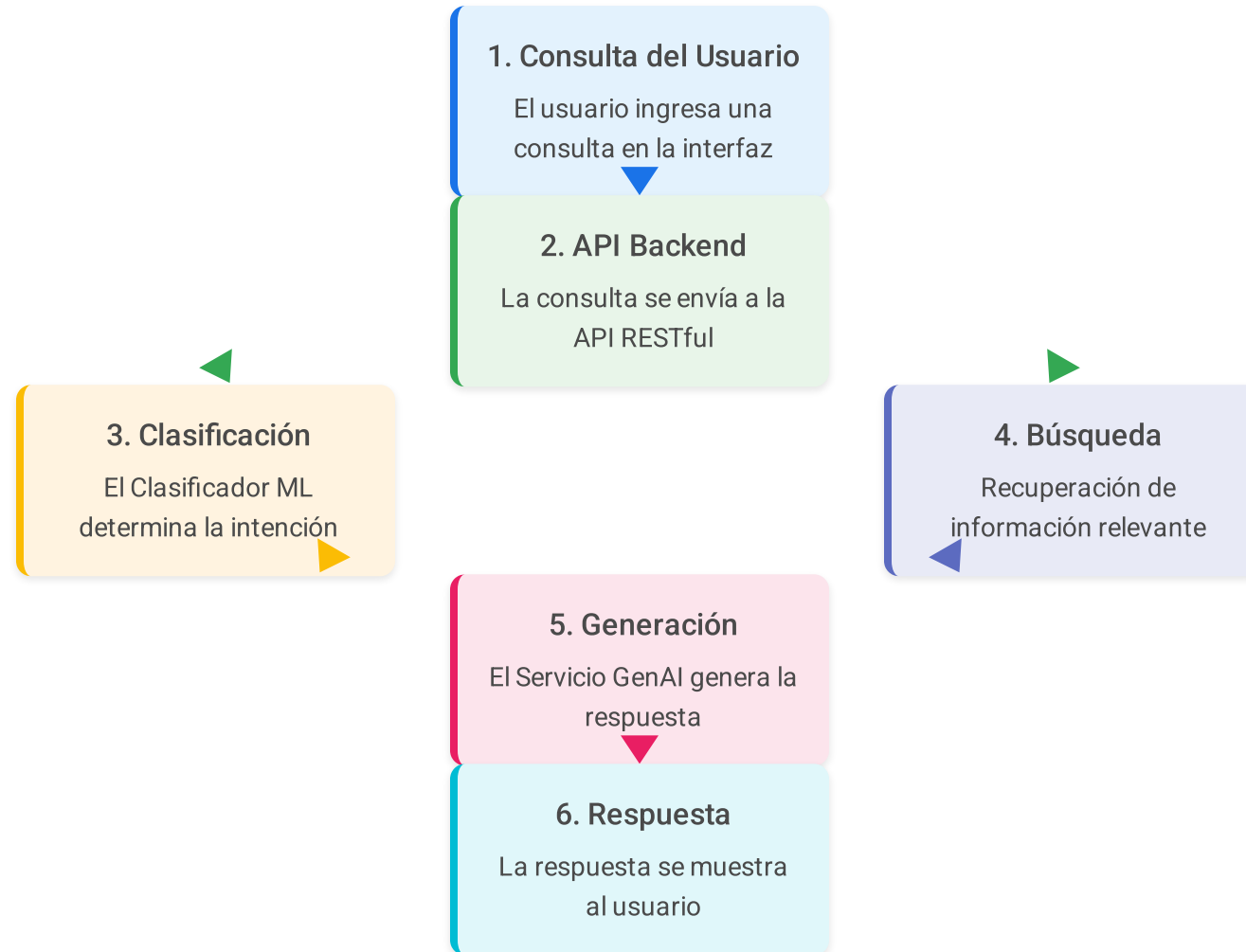
Clasificador de ML

Modelo entrenado para clasificar las consultas de los usuarios en categorías predefinidas, ayudando a dirigir la consulta al módulo correcto.

Tecnología: scikit-learn, RandomForestClassifier

Flujo de Interacción

El flujo de interacción muestra cómo las diferentes capas y componentes del sistema trabajan juntos para procesar las consultas de los usuarios y generar respuestas contextualizadas.



Conclusiones



Arquitectura Modular

Componentes independientes que facilitan el mantenimiento y la evolución



Escalabilidad

Diseño que permite crecer para atender mayor demanda



Robustez

Sistema resiliente con separación clara de responsabilidades



Rendimiento

Optimizado para respuestas rápidas y eficientes

La arquitectura de tres capas del Agente de IA Conversacional proporciona una base sólida para crear un sistema inteligente, adaptable y fácil de mantener que puede evolucionar con las necesidades de la organización.

Próximos Pasos



Expansión de la base de conocimiento



Integración con sistemas existentes



Mejora continua basada en análisis