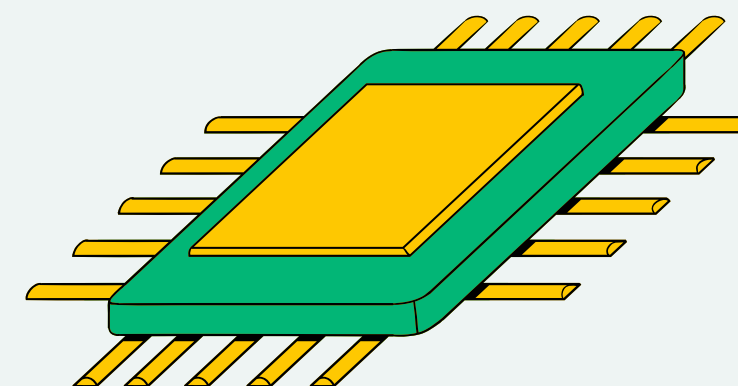


PROYECTO DE PRÁCTICAS: CHATBOT UGR PRESENTACIÓN

HECHA POR:

JAVIER TRUJILLO





ÍNDICE

- Contexto y Objetivo
- Arquitectura general
- Chatbot y Flujo de datos
- RAG
- QLora y Fine-Tuning
- Podman
- Conclusiones y Próximos pasos



CONTEXTO - OBJETIVOS

El objetivo es entrenar un chatbot con RAG sobre los apuntes y guías docentes de la universidad para que los alumnos puedan resolver sus dudas de forma autónoma.



De esta manera, el sistema aprovecha el contenido oficial y contextualizado sin exponer información delicada a APIs externas.



ARQUITECTURA GENERAL

1. Ingesta de datos

- Lectura de PDFs → pdfplumber → limpieza exhaustiva (clean_text)
- Dividir en chunks con RecursiveCharacterTextSplitter de LangChain
- Generar embeddings y poblar ChromaDB

2. Pipeline de entrenamiento

- Script generate_training_data.py:
 - Genera pares (pregunta, resumen) con T5 y BERT2BERT
 - Exporta train_data.json
- Adaptadores LoRA cuantizados (QLoRA)

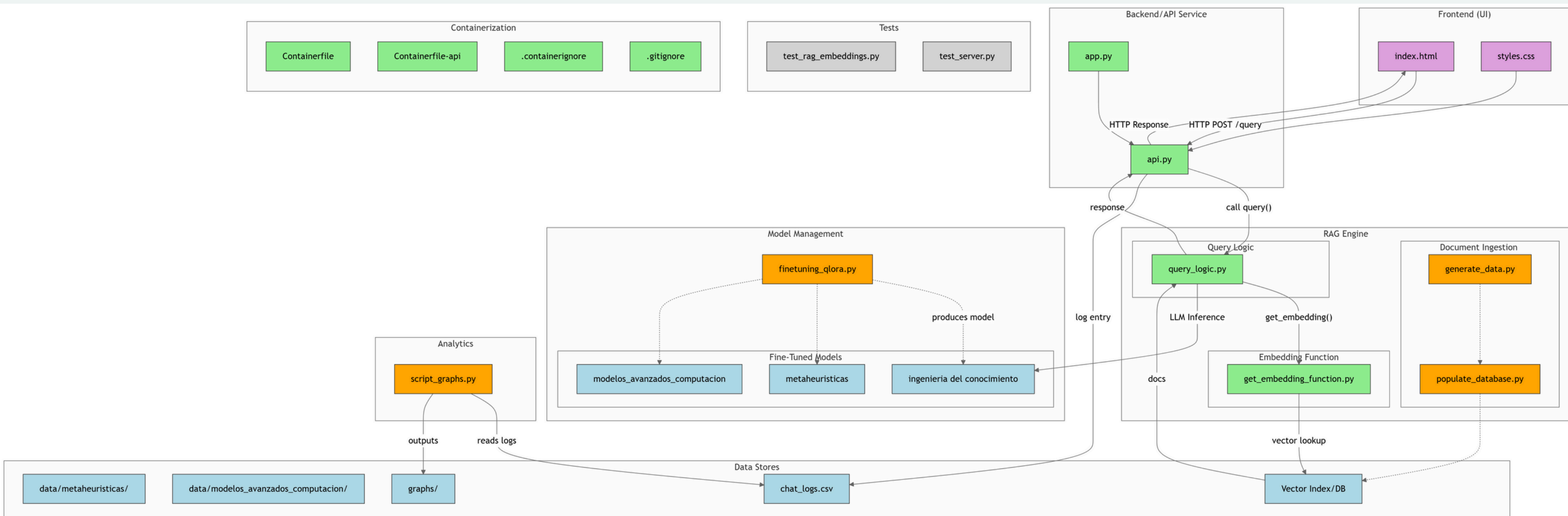
3. Servicio de inferencia

- query_logic.py: carga modelo base o LoRA, consulta RAG, construye prompt con historial

4. API y front

- FastAPI (api.py) con endpoints /chat y /graphs
- Frontend web simple en templates/index.html + static/





FLUJO DE TRABAJO

Usuario envía la pregunta

API recibe la pregunta (message, subject, mode)

query_logic.py genera la respuesta según los parámetros obtenidos

Se responde a la pregunta con fuentes y modelo



Asignaturas

Lista de asignaturas disponibles

- Ingeniería de Servidores
- Cálculo
- Algorítmica
- Sistemas Operativos
- Modelos Avanzados de Computación
- Metaheurísticas
- Ingeniería del Conocimiento

Desplegable para elegir modo de respuesta

Responde con: RAG Base

Ver Estadísticas

Cambiar correo

@correo.ugr.es

Botón modo claro/oscuro

Mensaje del usuario + modo de respuesta

Hola!
Modo: RAG

Hola! ¿En qué puedo ayudarte hoy?
Ver fuentes

Respuesta del bot + botón para consultar fuentes

El correo del profesor de prácticas del subgrupo 2 es < >.

- Ocultar fuentes
- Presentacion.pdf:2:66
 - Presentacion.pdf:2:59
 - Presentacion.pdf:3:67
 - Presentacion.pdf:3:60
 - Presentacion.pdf:1:58

Lista de los 5 fragmentos más relevantes para la respuesta. Y PDFs donde se encuentra la respuesta

cual es el correo del profesor de prácticas del subgrupo 2?
Modo: RAG

que es una maquina de turing no determinista?
Modo: RAG

Una Máquina de Turing No Determinista (MTND) es un modelo matemático de computadora que se basa en la Máquina de Turing, pero con una pequeña modificación. En una Máquina de Turing Determinista (MTD), la configuración de la máquina en un momento dado está determinada únicamente por la configuración actual y el símbolo leído. Por otro lado, en una Máquina de Turing No Determinista (MTND), la configuración de la máquina en un momento dado puede evolucionar con cualquiera de las tripletas (q, b, M), donde q es la configuración actual, b es el símbolo leído y M es el movimiento de la máquina. En otras palabras, en una Máquina de Turing No Determinista, la máquina puede tomar una gran variedad de decisiones en función de la configuración actual, el símbolo leído y el movimiento de la máquina, lo que la convierte en menos predecible que una Máquina de Turing Determinista. Esto también significa que la capacidad de una Máquina de Turing No Determinista puede ser mayor que la de una Máquina de Turing Determinista, ya que puede procesar un mayor número de lenguajes.

Ver fuentes

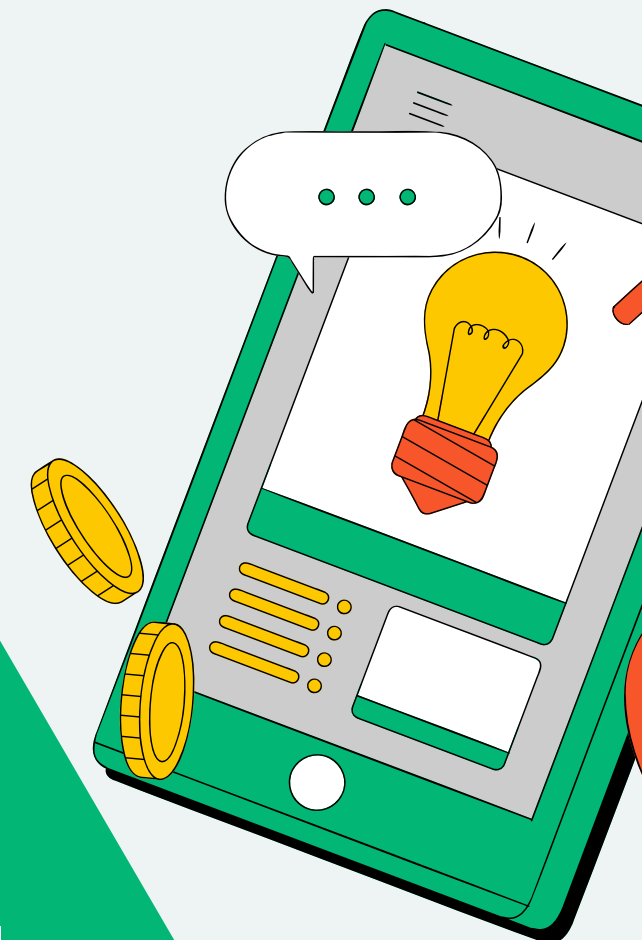
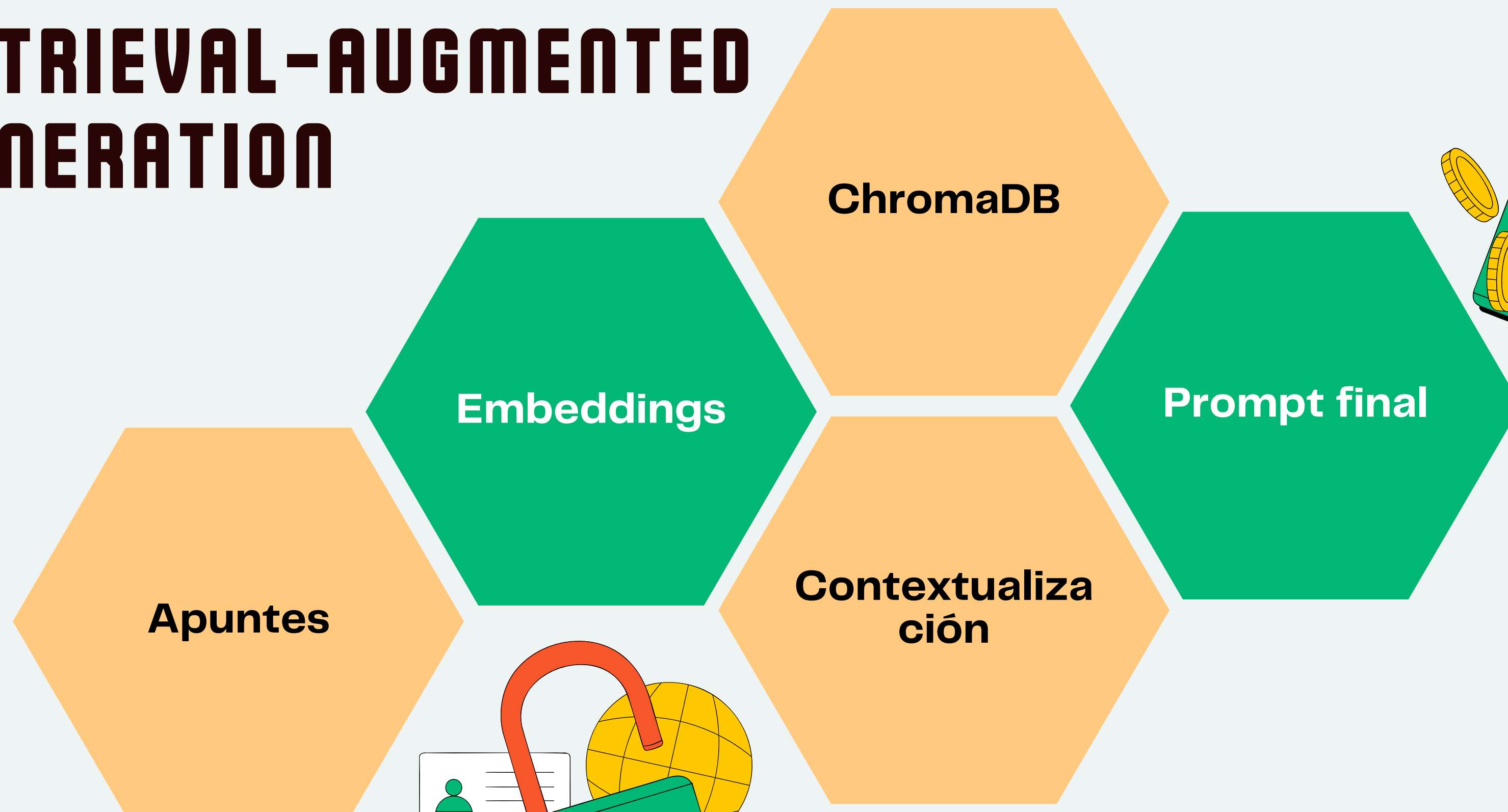
Botón para añadir documentos

Botón de visualización de las estadísticas
Botón configuración del correo + correo asociado

Escribe tu mensaje aquí...

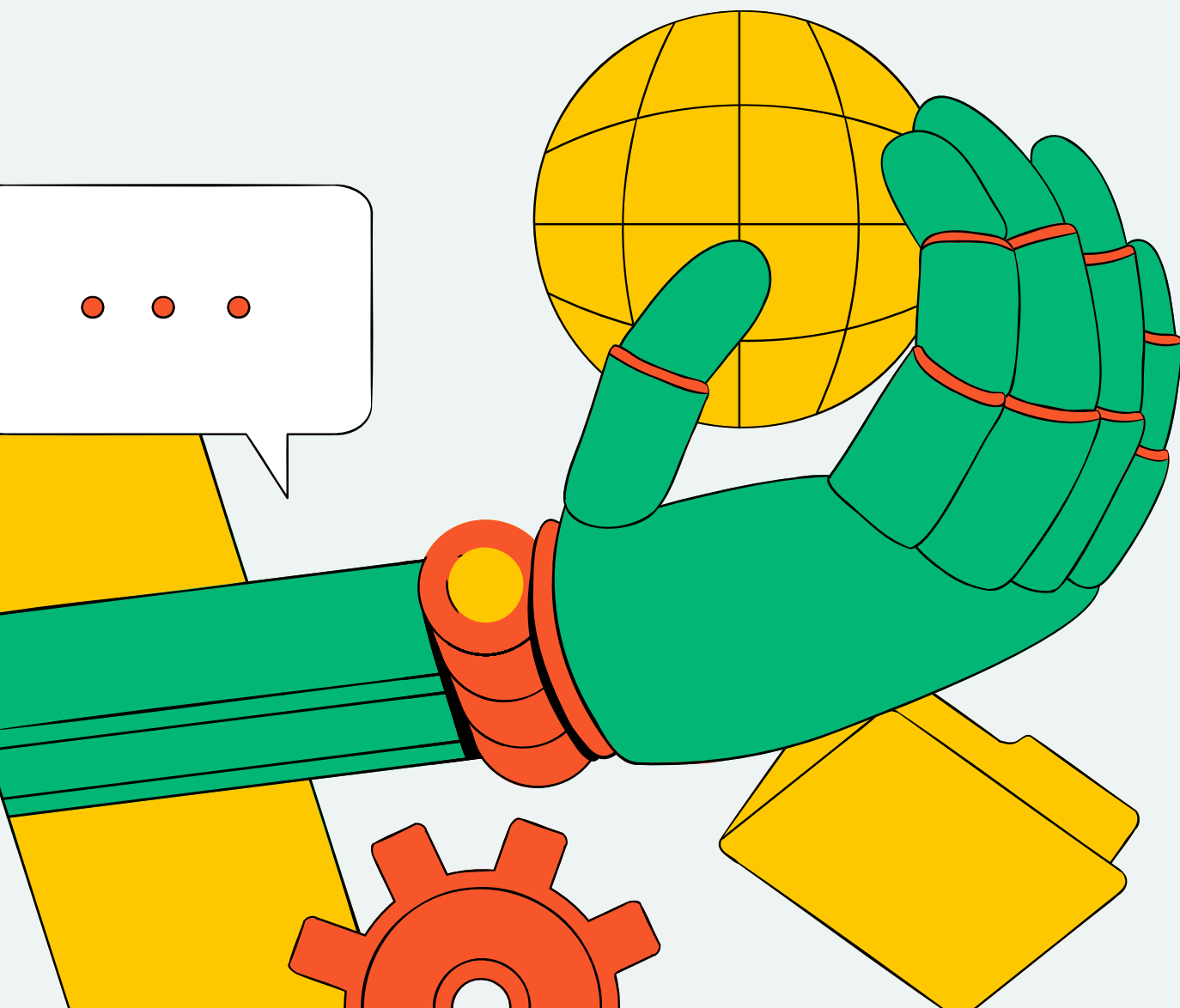
Enviar

RETRIEVAL-AUGMENTED GENERATION



FINE-TUNING CON QLoRA

QLoRA nos permite adaptar modelos de lenguaje de gran tamaño consumiendo muy poca VRAM, ya que emplea adaptadores LoRA cuantizados (4 u 8 bits) que ocupan una fracción del espacio de un modelo completo. Gracias a esta técnica, podemos aprovechar la potencia del LLM sin necesidad de infraestructura hardware excesiva.

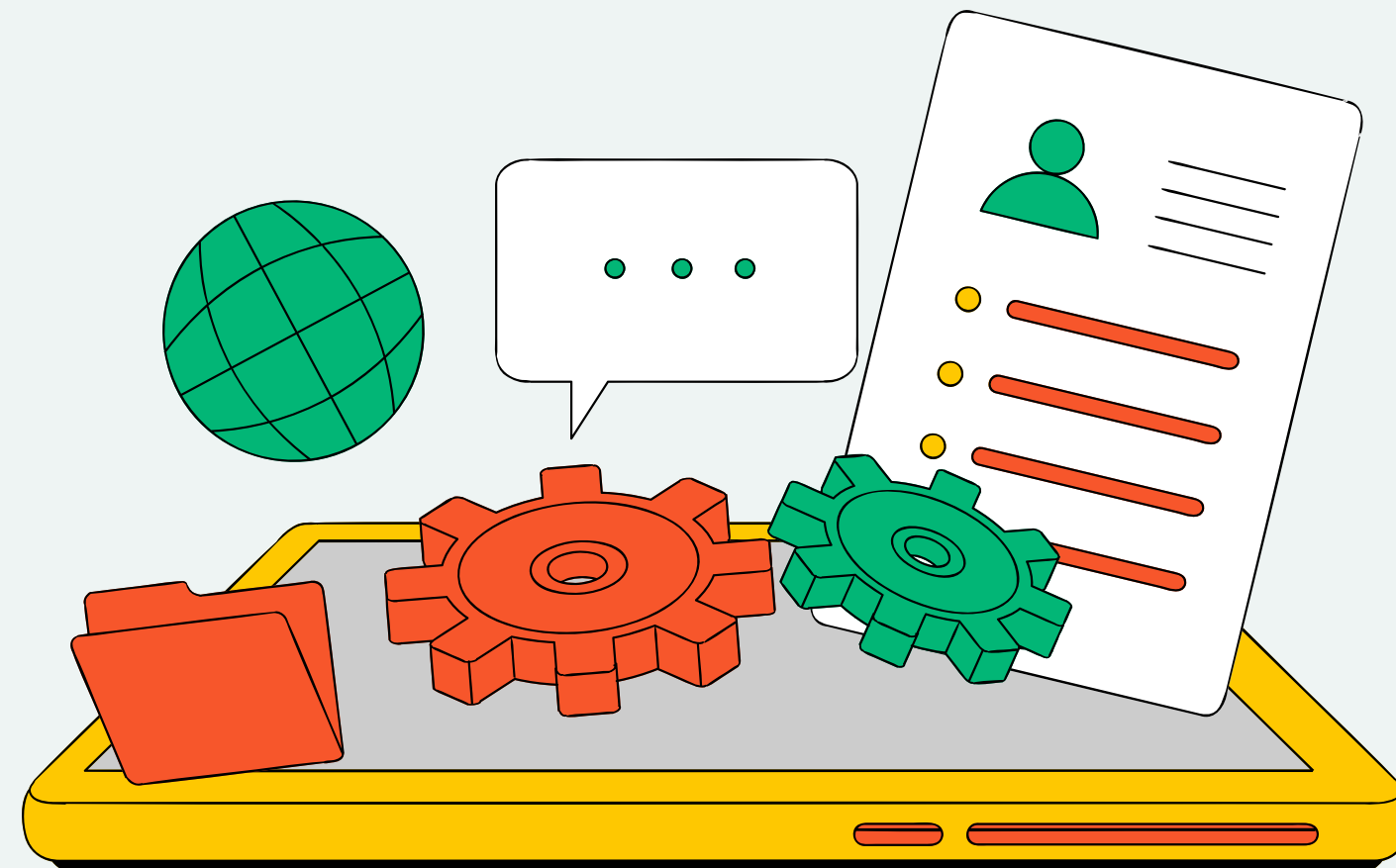


CONTENEDORES CON PODMAN

Seguridad

Compatibilidad

Despliegue ligero



CONCLUSIONES Y PRÓXIMOS PASOS



Chatbot modular

Uso de RAG + QLoRA

Despliegue con Podman y FastAPI

Modelo Multimodal

Inclusión en PRADO



Muchas gracias