

# Introducción a Ollama

David Santiago Fernandez Fandiño  
100911

Universidad ECCI  
*davids.fernandezf@ecci.edu.co*

## 1 Introducción

En los últimos años, los grandes modelos de lenguaje (LLM) han revolucionado el campo de la inteligencia artificial. Uno de estos modelos, ollama, ha demostrado ser una herramienta poderosa para diversas tareas, desde la generación de texto hasta la traducción de idiomas. En este proyecto, exploraremos cómo podemos aprovechar las capacidades de ollama para consumir servicios HTTP de manera eficiente y flexible.

## 2 Objetivos

El presente trabajo tiene como objetivo principal demostrar la viabilidad de utilizar modelos de lenguaje de gran tamaño, como ollama, para interactuar con servicios web a través de protocolos HTTP. Específicamente, se busca:

- Desarrollar una arquitectura que permita a ollama generar solicitudes HTTP de manera dinámica y procesar las respuestas recibidas.
- Evaluar la capacidad de ollama para extraer información relevante de los servicios HTTP y generar respuestas coherentes y útiles.
- Analizar el impacto de diferentes técnicas de prompt engineering en la calidad de las respuestas generadas por ollama.
- Identificar las limitaciones y desafíos asociados con el uso de ollama para este tipo de aplicaciones.

## 3 Conclusiones

En este trabajo hemos explorado la posibilidad de utilizar el modelo de lenguaje ollama para interactuar con servicios web a través de protocolos HTTP. Los resultados obtenidos demuestran que ollama tiene

un gran potencial para automatizar tareas, crear interfaces conversacionales y extraer información de servicios web.

Sin embargo, también hemos identificado algunas limitaciones, como la necesidad de diseñar prompts cuidadosamente y la importancia de contar con datos de entrenamiento de alta calidad.

Como trabajo futuro, se propone investigar las siguientes líneas de investigación:

- Explorar la integración de ollama con otros modelos de IA, como modelos de visión por computadora, para crear aplicaciones más sofisticadas.
- Desarrollar técnicas para mejorar la robustez de ollama ante cambios en las estructuras de los servicios web.
- Analizar el impacto del tamaño del modelo y la arquitectura en el rendimiento y la calidad de las respuestas.

En conclusión, este trabajo representa un primer paso hacia la integración de modelos de lenguaje de gran tamaño con servicios web, abriendo un amplio abanico de posibilidades para el desarrollo de aplicaciones inteligentes y automatizadas.

Repositorio de respaldo .