**Instituto Tecnológico de Costa Rica**

**Escuela de Ingeniería en Computación**

**Proyecto 2: Sistema RAG**

Greivin Mauricio Fernández Brizuela

Damián Obando Cerdas

Sebastián Obando Paniagua

IC-6200 – Inteligencia Artificial

9 de junio de 2024

Contenido

[Introducción 3](#_Toc168612173)

[Objetivos 3](#_Toc168612174)

# Introducción

En la era digital actual, la inteligencia artificial (IA) ha irrumpido en numerosos ámbitos, revolucionando la forma en que interactuamos con la tecnología. Una de las aplicaciones más destacadas de la IA es el desarrollo de chatbots o asistentes virtuales capaces de comprender y responder consultas en lenguaje natural.

El enfoque adoptado se basa en la combinación de modelos generativos preentrenados de vanguardia y la técnica conocida como Generación Aumentada por Recuperación (RAG, por sus siglas en inglés). Los modelos generativos preentrenados, como GPT-3 o LLaMa, han demostrado una capacidad impresionante para generar texto coherente y natural en una amplia gama de tareas. Sin embargo, a menudo carecen de un conocimiento profundo y actualizado sobre temas específicos.

Es aquí donde la técnica RAG entra en juego, permitiendo que el modelo generativo acceda a una base de conocimientos externa durante el proceso de generación de respuestas. Mediante la recuperación y el uso inteligente de información relevante de esta base de datos, el chatbot puede proporcionar respuestas más precisas, detalladas y contextualizadas.

# Objetivos

* Implementar un chatbot usando el modelo preentrenado llama3 con RAG.
* Realizar fine-tuning del modelo con un conjunto de documentos específicos para mejorar la precisión de las respuestas.
* Analizar y comparar el rendimiento del chatbot bajo diferentes configuraciones y prompts.

# Descripción del proyecto

Este proyecto aborda el desarrollo e implementación de un chatbot que combina técnicas avanzadas de procesamiento de lenguaje natural (NLP) con modelos generativos preentrenados y la novedosa técnica de Generación Aumentada por Recuperación (RAG).

En el núcleo del sistema se encuentra el modelo llama3, un potente modelo de lenguaje preentrenado en vastas cantidades de datos de texto. Sin embargo, a pesar de su impresionante capacidad generativa, estos modelos a menudo carecen de conocimientos específicos y actualizados sobre dominios especializados. Es aquí donde entra en juego la técnica RAG.

RAG permite al chatbot no solo generar texto coherente basado en el modelo de lenguaje, sino también recuperar información relevante de una base de datos de documentos cuidadosamente seleccionada y estructurada. Esta base de conocimientos contiene una amplia gama de documentos, como artículos, libros y noticias, que cubren el dominio de interés específico.

El proceso comienza con la selección y preparación de esta base de datos de documentos, asegurando su calidad, diversidad y relevancia. Luego, se configura e integra el modelo llama3 con el sistema RAG, permitiendo una recuperación eficiente de información durante el proceso de generación de respuestas.

Sin embargo, el proyecto no se limita a simplemente combinar estos componentes. Un aspecto clave es el ajuste fino (fine-tuning) del modelo llama3 utilizando un conjunto de datos de entrenamiento específicamente diseñado. Este proceso de entrenamiento supervisado permite adaptar el modelo a las particularidades del dominio y la tarea, mejorando aún más la precisión y contexto de las respuestas generadas.

Además, se llevan a cabo una serie de experimentos meticulosamente diseñados, probando el chatbot con diversos tipos de consultas y situaciones. Estas pruebas permiten evaluar el rendimiento del sistema, identificar áreas de mejora y realizar los ajustes necesarios.

# Metodología