RAG

↗ Parent item | IA

RAG (Retrieval-Augmented Generation) es una técnica que combina la generación de texto con la recuperación de información en tiempo real para mejorar la precisión y relevancia de las respuestas de los modelos de lenguaje natural (LLM, por sus siglas en inglés).

Aquí hay un desglose de cómo funciona:

1. **Retrieval (Recuperación):** En lugar de depender únicamente de la información almacenada en los parámetros del modelo, el sistema recupera datos relevantes de una base de conocimientos externa (por ejemplo, una base de datos, documentos, artículos web, etc.) en respuesta a una consulta específica.
2. **Augmentation (Aumento):** La información recuperada se utiliza para complementar la entrada original del modelo. Esto proporciona contexto adicional y datos actualizados que pueden no estar presentes en el modelo preentrenado.
3. **Generation (Generación):** Con la entrada aumentada, el modelo genera una respuesta más precisa y relevante, utilizando tanto la información recuperada como su conocimiento preexistente.

La principal ventaja de RAG es que permite a los modelos LLM acceder a información actualizada y específica, mejorando así la calidad y precisión de las respuestas en contextos donde la información puede cambiar rápidamente o donde se requiere conocimiento especializado que no está incluido en el entrenamiento inicial del modelo.