Aplicaciones LLM

La técnica RAG

Recuerda: ¿Superar los límites de la context window?

- Entrenando un LLM desde cero con nuestros datos.
  - Enormemente caro. Inviable en la práctica para la mayoría.
- Añadiendo nuestros datos a un LLM ya entrenado (fine-tuning).
  - Muy caro y técnicamente muy complejo. Inviable en la práctica para la mayoría.
- Con la técnica RAG (Retrieval-Augmented Generation) dividimos nuestros datos en pequeños segmentos y así permitimos que el LLM pueda utilizarlos dentro de los límites de su context window.
  - Esta es la técnica utilizada hoy en día por la práctica totalidad de las Aplicaciones LLM.

La técnica RAG

- Preparativos:
  - Divide tus datos en pequeños segmentos.
  - Convierte los pequeños segmentos en números (“embeddings”).
  - Carga los embeddings en una base de datos vectorial.

- Ahora cuando preguntas (“query”) al LLM:
  - El LLM va a la base de datos vectorial y busca (“indexing”) solo (“semantic similarity”) los datos que responden a tu pregunta.

Por lo tanto, al utilizar RAG se dice que:

- La capacidad de hablar (“language generation”) viene del Foundation LLM.
- El conocimiento específico (“knowledge representation”) viene de la base de datos vectorial.
- En otras palabras:
  - El Foundation LLM actúa como una persona que sabe hablar pero no conoce tus datos.
  - La base de datos vectorial actúa como el conocimiento experto que tú añades a ese Foundation LLM para que se comporte como una persona que sabe hablar sobre tus datos.