Aplicaciones LLM

RAG vs. In-Context Learning

Hay cierta confusión entre ambas

- 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.

- En algunos medios se confunde la técnica RAG con la técnica In-Context Learning. A continuación aclararemos la diferencia entre ambas.

- Esta aclaración sólo es relevante si algún estudiante tuviera esta duda. De lo contrario es irrelevante, una mera cuestión teórica.

La técnica RAG

- Objetivo:
  - Combinar la capacidad de recuperación de información con la generación de lenguaje para responder preguntas utilizando información externa.
- Mecanismo:
  - RAG utiliza un sistema de recuperación (retriever) para buscar documentos relevantes o fragmentos de texto en una base de datos (por ejemplo, un corpus de Wikipedia). Luego, utiliza un modelo generador (como BERT o GPT) para formular una respuesta basada en los fragmentos recuperados.
- Ejemplo:
  - Si le preguntas a un modelo RAG sobre un tema específico, primero buscará en su base de datos para encontrar información relevante y luego utilizará esa información para generar una respuesta coherente.

La técnica in-context learning

- Objetivo:
  - Adaptar un modelo pre-entrenado a tareas específicas proporcionando ejemplos en el contexto de entrada.
- Mecanismo:
  - No se re-entrena el modelo. En cambio, se proporciona un contexto que incluye ejemplos de la tarea deseada, y se espera que el modelo generalice a partir de ese contexto para responder adecuadamente.
- Ejemplo:
  - Con modelos como GPT-3 o GPT-4, puedes proporcionar ejemplos de traducciones en el contexto de entrada (por ejemplo, "Inglés: 'Hello' -> Español: 'Hola'") y luego hacer una pregunta de traducción sin proporcionar el ejemplo explícitamente.

En resumen

- "In-context learning" se basa en proporcionar ejemplos en el contexto de entrada para guiar al modelo en la tarea deseada.

- RAG combina la recuperación de información con la generación de lenguaje para responder preguntas utilizando datos externos.

- Ambas técnicas buscan mejorar la capacidad de los modelos de lenguaje para adaptarse a tareas específicas y proporcionar respuestas informadas. Sin embargo, utilizan enfoques y mecanismos diferentes para lograrlo.