LLM Apps

RAG vs. In-Context Learning

There is some confusion between the two

- With the RAG (Retrieval-Augmented Generation) technique, we divide our data into small segments, thus allowing the LLM to use them within the limits of its context window.
  - This is the technique used today by almost all LLM applications.

- In some media, the RAG technique is confused with the In-Context Learning technique. Next, we will clarify the difference between the two.

- This clarification is only relevant if a student had this question. Otherwise, it's irrelevant, a mere theoretical matter.

The RAG technique

- Objective:
  - Combine information retrieval capability with language generation to answer questions using external information.

- Mechanism:
  - RAG uses a retrieval system to search for relevant documents or text snippets in a database (e.g., a Wikipedia corpus). It then uses a generator model (like BERT or GPT) to formulate an answer based on the retrieved snippets.

- Example:
  - If you ask an RAG model about a specific topic, it will first search its database to find relevant information and then use that information to generate a coherent answer.

The In-Context Learning Technique

- Objective:
  - Adapt a pre-trained model to specific tasks by providing examples in the input context.

- Mechanism:
  - The model is not re-trained. Instead, a context is provided that includes examples of the desired task, and the model is expected to generalize from that context to respond appropriately.

- Example:
  - With models like GPT-3 or GPT-4, you can provide translation examples in the input context (e.g., "English: 'Hello' -> Spanish: 'Hola'") and then ask a translation question without providing the example explicitly.

In summary:

- "In-context learning" is based on providing examples in the input context to guide the model in the desired task.
- RAG combines information retrieval with language generation to answer questions using external data.
- Both techniques seek to enhance the ability of language models to adapt to specific tasks and provide informed answers. However, they use different approaches and mechanisms to achieve this.