Desarrollo de Aplicaciones con LLM

Clase 5

Manejo de Memoria en LangChain

- Los modelos de lenguaje tienen una memoria corta.
- Solo pueden recordar la información que se encuentra en su ventana de contexto.
  - GPT-4o: 128k tokens
  - Claude3: 128k tokens
  - Gemini: 1M tokes
- Langchain provee algunas alternativas para mejorar la memoria de los modelos de lenguaje.

LangChain

Manejo de Memoria en LangChain

- Conversation Buffer Memory:
  - Almacena mensajes y los carga en el siguiente mensaje.
- Conversation Buffer Window Memory:
  - Similar al anterior, pero puedes limitar la cantidad de intercambios conversacionales almacenados en la memoria.
  - Por ejemplo, puedes configurarlo para que solo recuerde las últimas 2 preguntas y respuestas de la conversación.
- Conversation Token Buffer Memory:
  - Similar al anterior, pero esta vez puedes limitar la cantidad de tokens almacenados en la memoria.
- Conversation Summary Memory:
  - Almacena un resumen de los intercambios conversacionales anteriores.

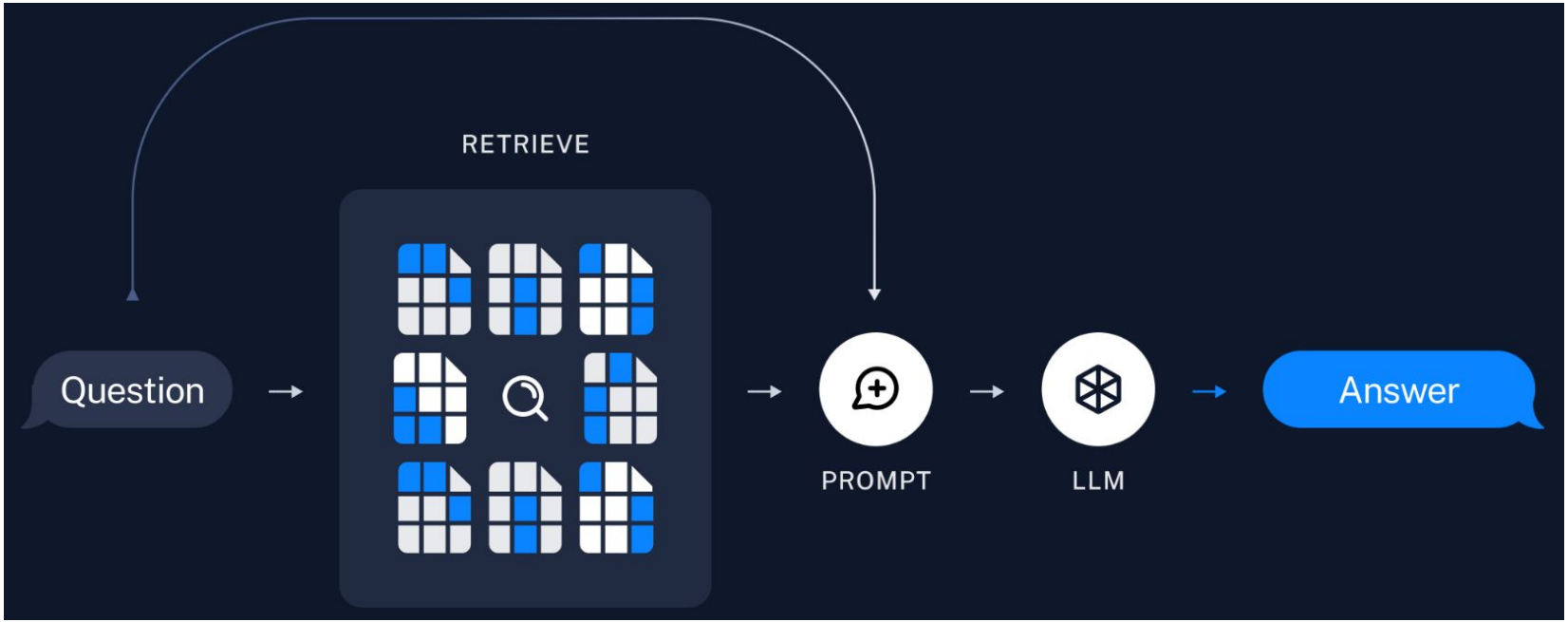
RAG (Retrieval Augmented Generation)

- Generación Aumentada con Recuperación (RAG)
- Componentes de RAG
  - Componente de Recuperación (Recuperador):
    - Este componente busca información relevante en una base de datos o un conjunto de documentos en respuesta a una consulta.
    - Utiliza técnicas como búsqueda por palabras clave, embeddings y modelos de recuperación basados en redes neuronales.
    - El resultado de esta etapa es un conjunto de documentos o fragmentos de texto relevantes.
  - Componente de Generación (Generador):
    - Una vez recuperada la información relevante, el modelo generador (como GPT) toma estos fragmentos y los utiliza para generar una respuesta coherente y contextualmente adecuada.
  - Este modelo puede mejorar la precisión y la especificidad de las respuestas al basarse en la información proporcionada por el componente de recuperación.

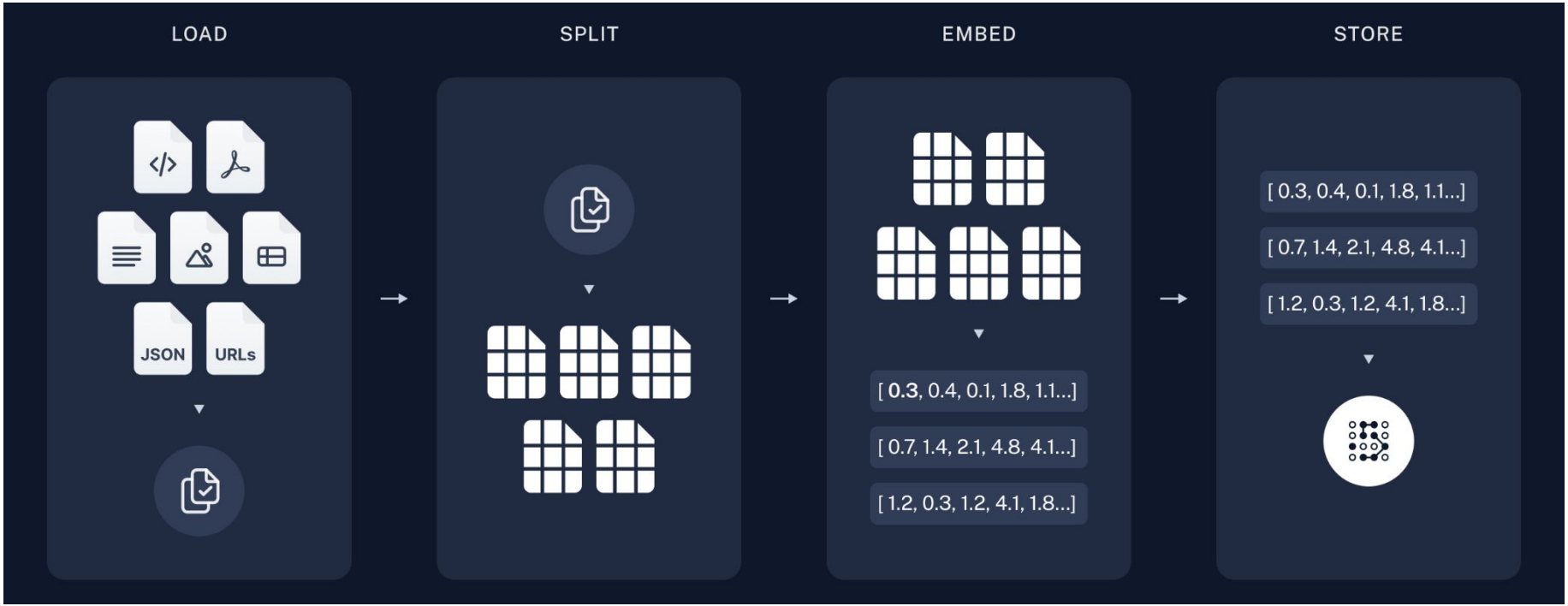
RAG (Retrieval Augmented Generation)

- Generación Aumentada con Recuperación (RAG)
- Componentes de RAG
  - Componente de Recuperación (Recuperador):
    - Este componente busca información relevante en una base de datos o un conjunto de documentos en respuesta a una consulta.
    - Utiliza técnicas como búsqueda por palabras clave, embeddings y modelos de recuperación basados en redes neuronales.
    - El resultado de esta etapa es un conjunto de documentos o fragmentos de texto relevantes.
  - Componente de Generación (Generador):
    - Una vez recuperada la información relevante, el modelo generador (como GPT) toma estos fragmentos y los utiliza para generar una respuesta coherente y contextualmente adecuada.
  - Este modelo puede mejorar la precisión y la especificidad de las respuestas al basarse en la información proporcionada por el componente de recuperación.

RAG (Retrieval Augmented Generation)



RAG (Retrieval Augmented Generation)



Fin.