LLM terms: → not discussed so far

- Inference: refers to when a trained ML model like LLM makes predictions from new unseen data

- Chunking: in RAGs the chunking step breaks down large documents or data sources into smaller, manageable chunks so the retriever can search through large volumes of data while staying in token/input limits. Each chunk usually a paragraph or section is converted to a embedding and stored in vector DB when a search is made only the most relevant chunks are returned

- Retrieval RAG: in Rag retrival is finding relevant info from large dataset to support generation of response, when a query is recived its converted to a vector (embedding) and uses this vector to search the DB of pre indexed embeddings identifing most similar or relevent data points. Techiques like ANN (approximate nearest neighbour are used.

- Streamed vs unstreamed responses: streamed responses mean the LLM/Agent more specifically like ChatGPT starts sending the words as soon as they are generated (user sees text grow in real time) unstreamed responses wait until the whole answer is generated and then sends it all at once but user waits longer especially if long answer.