LLM terms

- Memory (STM vs LTM): Done by LSTM Networks a type of RNN

  - Short term memory (STM) is the facts passed in as part of the prompt, like users name age etc

  - Long term memory (LTM) in A LLM (more specifically Agents) store important information for the future it saves facts like past events, preferences and learned skills for better long term performance, unlik STM it lives beyond session

- Other Architecture patters: we have seen RAG, ReAct, COT but other Agent patterns are: Planner Executor (splits into planner and exutor to divide work, DAG Agents or Direct asyclic graph where each node is a agent dose a task and passes it to the next agent with no cycles (one way), TOT (tree of thought) is a way to organize agents reasoning the root is the main problem and it branches out creating ideas/steps to lead to a sol.

- Vector Database: when implementing something like a RAG or Agents Vector DB stores and Retrives embeddings efficiently, these embeddings are vector representation of data like documents, text, images etc during query the system converts it to a embedding and Searches the vector DB for the most similar/relevant embedding (Ex, related documents/snippits) it dose this using methods like ANN to find similar data points in high dim space. this retrived info from DB is then fed to something like LLM.