

Summary on Day 25:

By 11a.m I have started to working on my project. Day before yesterday I tried by implementing node.js but it doesn't work.so today I have installed Gemini API packages by using the command

Npm install @ google/genai in my project bot

Afterthat I have created frontend page using html and css.

By 12 p.m we went to attend the presentation session.

There sriram take the session on AI like how to use AI and hw prompting is important to get the desired output and what are all the prompting techniques are there and how it is applied and all.

When we gave the prompt in the AI tool sometimes it does not give the exact output that what we needed,so we need prompt engineering.

Tokens: when the client gives the input as a request to AI the inputs are taken as token like each word is considered as a token.

Context limit: The inputs we given are have some limits. When the limit is exceeds then we cannot able to prompt.

Modules: different types of modules are chatgpt,gpt 3,claude,Gemini etc.

Embeddings: Embeddings are numerical representation of concept.

Prompt: prompts are the input we given to Ai like what we needed is given as a input to AI as a request.

Work of LLM:

It is a prediction engine. If we gave the text as input then it predicts what the following token should be, based on the data it was trained.

Before doing prompt engineering, we need to **choose the module** like chatgpt,Gemini,claude etc.

Each module has different token count like for a novel in Gemini it has about 70000 to 100000 token count. Compared to Gemini chatgpt has small number of tokens.

Every prompt that retrieves message from AI by using

- **Temperature**

- **Top k**
- **Top p**

Temperature: if we gave a single word in LLM, it predicts the probability of what the next token. So **lower** the temperature in prompts that expects more deterministic response. **higher** the temperature leads to unexpected results.

Top k and top p: it **restricts** the predicted probability of next token. Similar to temperature it also have some controls.

Setting values to temperature, top k, top p:

- If temperature is 0, top k and top p becomes irrelevant.
- If temperature is extremely high, then temperature becomes irrelevant.
- If top-k is 1(higher), then temperature and top p becomes irrelevant.
- If top-p is 0, then temperature and top k becomes irrelevant.

Prompting techniques:

Zero shot: it is the simplest prompt. it is like general prompting.

One shot: in one shot the prompt which provides a single example which helps the models to understand what you are asking for.

Few shot: it is similar to one shot with many examples.

Step-back prompt: is a technique for improving the performance by prompting. In general we first give the prompt and after it displays the answer then we again give prompt relevant to the specific task.

Chain of thought (cot): is a technique for improving reasoning. In general we give prompt and to improve reasoning we add the prompts as step-by-step.

Tree of thought: it is similar to chain of thought but with multiple reasoning paths simultaneously. It is used in complex problems.

Memory banks: what are all the inputs given by the user in first stored in the memory banks after that only it response to the user.

He explained all these by using real time examples. This session makes me to understand how to use AI. By following these principles how we can retrieve the exact thing of what we needed. Using these techniques we can handle AI in simpler and in an most efficient way.

This session takes two hours of class. After the lunch again I start working on my project.