Prompt Engineering: -

Base Models - Predict the next word given a set of text.
Chat Models - Ideal for powering chatbots.
Specialized Models - Specialized for specific tasks like coding

Few Shots learning - We give the model a few examples and thus teach it how to perform.

Overview of Llama Models:-

Llama2 - 7B, 13B and 70B parameter models

Larger the model - more capacity to learn from the training data, but also more computationally expensive.
Upon instruction tuning we get Llama chat models
Instruct models - Exhibit more human like behaviour

Code Llama - Llama2 model for coding tasks (3 sizes similar to Llama 2)

Formatting prompts in Llama2:-

We surround the prompt using Instruction tags.
Put verbose = True to display the prompt.

“[INST] Prompt [/INST]”

The base / foundation models don’t understand Instruction tags, so put add_inst = False.

Temperature - Amount of creativity in the response.
Consistent responses - Temp = 0
Creative - Temp = 1
MAX_TOKENS - Max No. Of tokens in response

Input tokens + Max tokens has a limit, so if we give a really long input, it may cross that limit with the output tokens.

MULTI TURN CHAT PROMPTS:-

To give context to the LLM, we pass the previous prompt and model response in the new prompt as well.
<s> - Start tag, </s> - End Tag, to enclose a prompt and response pair.

<s> [INST] Prompt 1 [/INST]
Response 1 </s>
<s> [INST] Prompt 2[/INST]
Response 2 </s>
<s> [INST] Prompt 3 [/INST]

Note no end tag at the end.

PROMPT ENGINEERING TECHNIQUES:-

We can guide the model to improve its response by providing it with a specific set of instructions.
1. One way is to provide examples of the task to be carried out.
2. Tell the model to assume a persona / role
3, Specifying how to format responses
4. Including additional info / data for the model to use in its response

Giving examples of the task - In Context Learning
Can be used for zero / few shot prompting.
Role - Make the LLM behave as someone to generate better promoting

Zero-shot Prompting

You are prompting the model to see if it can infer the task from the structure of your prompt. In zero-shot prompting, you only provide the structure to the model, but without any examples of the completed task.

Few-shot Prompting

In few-shot prompting, you not only provide the structure to the model, but also two or more examples.
You are prompting the model to see if it can infer the task from the structure, as well as the examples in your prompt.

Role Prompting

Roles give context to LLMs what type of answers are desired.
Llama 2 often gives more consistent responses when provided with a role.

SUMMARISING GIVEN A CONTEXT:-

```
context = """
<paste context in here>
"""
query = "<your query here>"

prompt = f"""
Given the following context,
{query}

context: {context}
"""
response = llama(prompt,
                 verbose=True)
print(response)
```

REASONING BASED PROMPTING:-

Chain of thought prompting - ’Think step by step’ or ‘Explain your reasoning’.

```
prompt = """
15 of us want to go to a restaurant.
Two of them have cars
Each car can seat 5 people.
Two of us have motorcycles.
Each motorcycle can fit 2 people.

Can we all get to the restaurant by car or motorcycle?
```

```
Think step by step.
Explain each intermediate step.
Only when you are done with all your steps,
provide the answer based on your intermediate steps.
"""
response = llama(prompt)
print(response)
```

Since LLMs predict their answer one token at a time, the best practice is to ask them to think step by step, and then only provide the answer after they have explained their reasoning.