Parameters

- Temperature
- Sampling
  - Top p
  - Top k
- Repitition Penalty
- Max tokens.

Advanced Prompting techniques

- Zero Shot Prompting
- One Shot Prompting
- Few Shot Prompting
- Role Based Prompting
- Chain-Of-Thought
- Self Consistency.

Zero Shot Prompting

↳ Only provide the task to LLM, without any example.

Ex: Translate this sentence to Hindi

→ Quick responses

→ Instructions are clear

One Shot Prompting

↳ Provide the task to LLM with exactly one example.

⇒ Convert the sentence in a polite form.

Ex: Sentence: Open the door.
Polite: Please, Open the door

Now, Sentence: Call me later.
↳ Please, Call me later.

When we need to give minimal guidance to our model.

→ Slightly ambiguous
→ format / tone / -

Few - Shot Prompting

↳ Task (Instruction) + Few examples

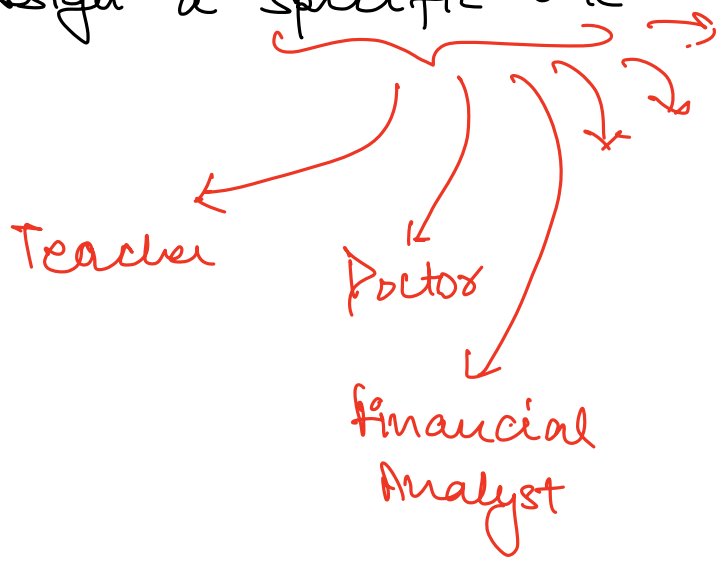
(3-5)

→ Classification Problems.
→ Complex / difficult to understand tasks.

Role Prompting

↳ Assign a role / persona to model to control tone, depth & perspective of the response.

→ Inside the prompt, assign a specific role to the model.



Zero-Shot ⇒ 0 examples

One-Shot ⇒ 1 example

few-Shots ⇒ 3-5 examples

Role Prompting ⇒ Assign a role to the model.

Chain-of-Thought (CoT)

↳ Prompt to ask the model to show the step-by-step reasoning before giving the final answer.

Instead of

Give me the answer.

With CoT:

Think step by step & then answer

⇒ In CoT, we are asking model to break down a problem into intermediate steps to reach to the final response.

→ Solving Maths | DSA | Puzzle problems.

→ Multi-step reasoning process.

Self-Consistency Prompting

Instead of trusting one answer, self-consistency asks the model:

1) Generate multiple answers.
2) Compare the answers.
3) Select the most frequent | consistent result.

⇒ Self-Consistent is a technique where multiple outputs are generated for the same instruction & one of these multiple answers, we select the most frequent one.

→ Expensive < Cost
Latency

Temperature → 0.0 - 2.0

↓

Controls Creativity of the model.

less Temp. ⇒ More predictable | consistent

High Temp ⇒ More creative | random.

Sampling

- Top-K
- Top P

Apple is going to launch new

0.4	0.3	0.15	0.1	0.05
Iphone	Macbook	Ipad	Airpods	Imac

Top-K ⇒ K = 3.

Top-(P) ⇒

↳ 0.8.

0.4	0.3	0.15	0.1	0.05
Iphone	Macbook	Ipad	Airpods	Imac
0.4	0.7	0.85	0.95	1