Agenda.

→ Prompt Engineering.

→ Element of a prompt

→ Parameters of a prompt

Prompt Engineering is art of writing a prompt, choosing the right words to guide the LLM's performance & adjusting the output till we get the desired output.

→ Instructions we give to LLM to harness it's full potential & generate the output.

Elements of a Prompt

- → Instruction
- → Context
- → Constraints
- → Format
- → Variables.

Instruction

↳ What to do in a very basic & straight forward way.

Prompt: Generate a packing list for a trip.

Context.

↳ Some additional | background information about the instruction provided.

Prompt: Generate a packing list for a 5 days business trip to USA

Constraint

↳ Limitations | Restrictions we need to put on the model to generate the response.

→ What to include vs What not to include

→ Tone

→ Audience

→ Length - - - - - - - -

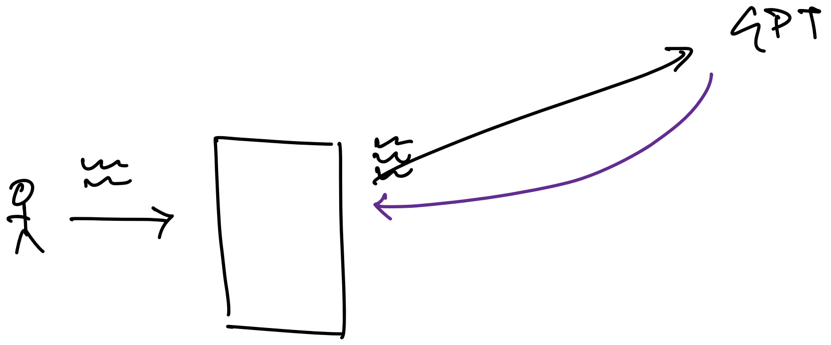
Prompt : Generate a packing list for a 5 days business trip to USA for 1 person who is 30 year old & baggage limit is 10KG, . . .

Format.

→ Structure or layout of the response generated by LLM.

→ How to organise & style the response.

Prompt : Generate a packing list for a 5 days business trip to USA for 1 person who is 30 year old & baggage limit is 10KG, give the response in the email format.



JSON

```
{
        "id"    :  "1234" ,
        "name"  :  "iphone 17 pro"
        "brand" :  "Apple"
        "price" :   ____
        "__"    :   ____

}
```

Variables.

→ {user_name} → Curly Brackets.

Prompt :

{sender} = Deepak

{receiver} = Dinesh

{duration} = 2

{Destination} = USA

{baggage_limit} = ____

Write an email to {receiver} from {sender} with a packing list for {duration} days trip to {Destination} - - - - - - - .

Give the response in email format.

Parameters of Prompt.

↳ These params can be used to control | manage model's output.

1) Temperature
2) Sampling
3) Repitition penalty
4) Max Tokens.

Temperature

↳ Helps us to control the predictability or randomness of the response.

→ Range : 0-2

low temperature ⇒ More deterministic

High temperature ⇒ More creative | diverse.

Temperature		Example Use-Case
0 – 0.3	⇒	Code | Algo | Maths | - - - -
0.3 – 0.6	⇒	Explanations.
0.6 – 1	⇒	Posts | Blogs | - - -
1 – 1.4	⇒	Stories | fictions | - - -
> 1.5	⇒	Rarely Used.

```
Client = OpenAI()

response = client.response.create(
                 "model" : "gpt-4",
                 "prompt" : "__________"
                 "temperature" : "⊗"
           );
```

Sampling

↳ Controls how the model picks up the next Tokens.

Top -K Sampling

↳ Select top K tokens based on the probability.

API

REST → 0.85

Http → 0.81

fastAPI → 0.72

Graph API → 0.59

Pharma API → 0.41

Top- K

K=1 ⇒ Only REST

K=2 ⇒ REST & HTTP

K=3 ⇒ _ _ _

Top -P.

↳ Chooses tokens with probability > P.

Repitition Penalty

↳ Control the repitition of words or phrases the model has already generated.

→ It reduces the redundancy.

Repitition Penalty

1	→	No penalty
1.1-1.4	→	mild repition Control
1.4-1.7	→	strong repition Control
> 2	→	

Max Tokens.

↳ Maximum no. of tokens our model can generate as a response.