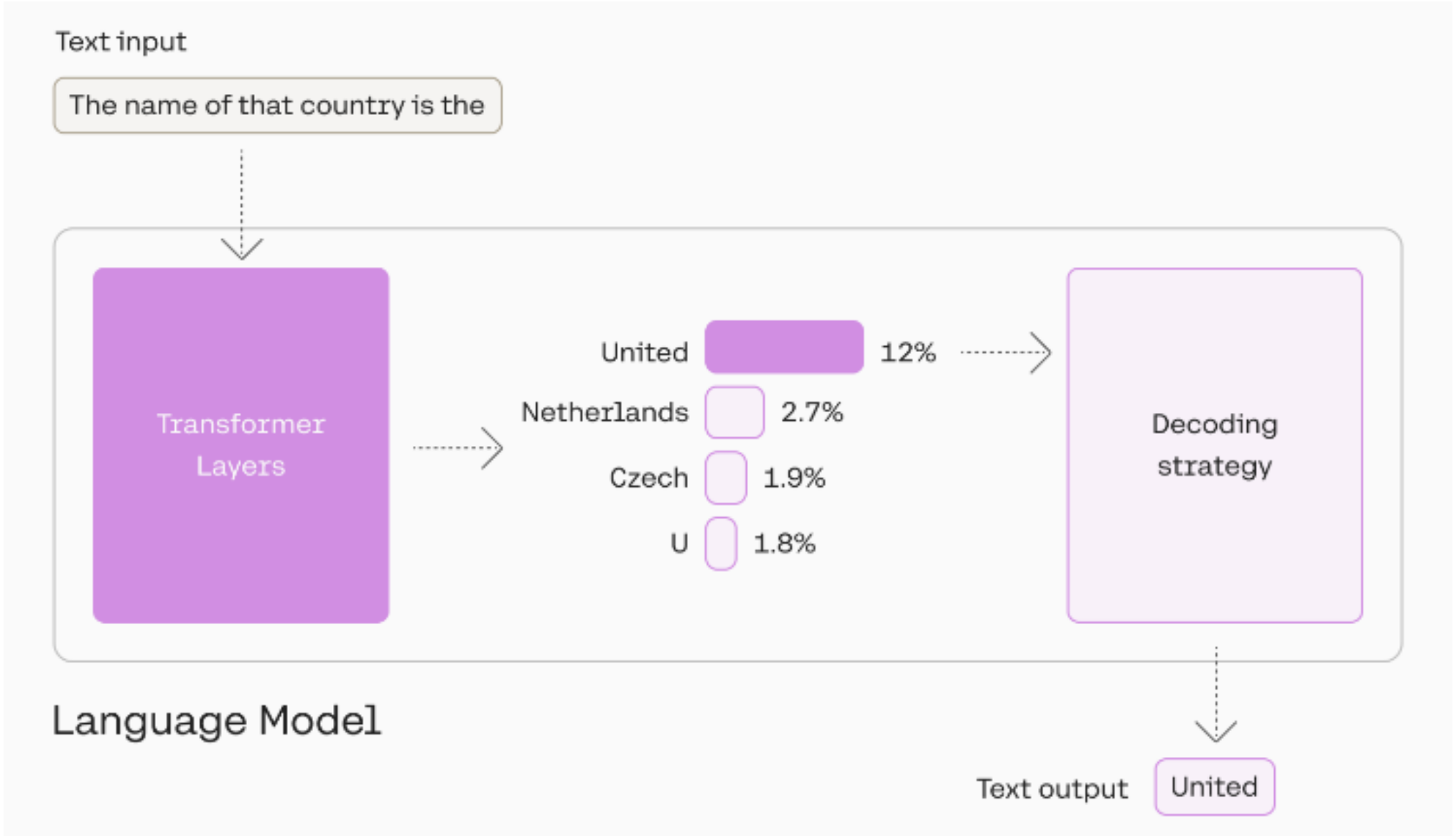
OpenAI 參數說明

Temperature

- 決定模型輸出的多樣性
- 數值介於 0-2
- 通常與 top_p 參數合併使用
- 官方建議 temperature & top_p 一次只更動一個
- Temperature 越高,模型輸出的創意越高,但出錯率也越高
- 出錯是指模型自創答案的可能性就會越高
- Temperature越小(例如小於0.2)模型輸出確定性結果的機率越高
- Temperature越小模型越容易輸出重複的結果

Top_p

- 和 temperature 一樣,決定模型輸出的多樣性
- 預設值為1
- Top_p 指定為 0.1, 代表只有累積機率為10%的字會被模型納入考慮
- Top_p越小,模型的確定性就越高
- Top_p越大,模型的多樣性就越高
- 官方建議 Top_p 與 temperature 選一個來調整即可



https://docs.cohere.com/docs/temperature

MaX Tokens

代表模型做多生成多少個 token
要注意 max tokens 的設置,因為這個與 $$有關
GPT 4 有 8K 與 32K 兩種模型(後者較貴)
8K指的是輸入以及輸出總共有8000個 tokens

	輸入	輸出
Model	**Prompt**	**Completion**
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

n

- 每次要生成的結果個數
- 預設值為1
- 如果想要一次請GPT生成多個可能的結果,可以將預設值設定超過1的數字
- 這個值越大,會快速消耗可能的 token, $$也會消失得更快

Stop

- 阻斷串列
- 可以用來阻止 GPT 再次生成結果
- 例如 top =[‘3’]則代表GPT如果發現下一個要生成的字是 3 時,就會停止生成
- STOP的好處是可以減少 token的消耗

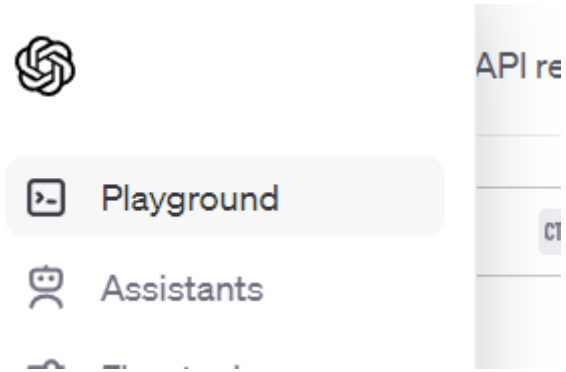
Frequency_penalty

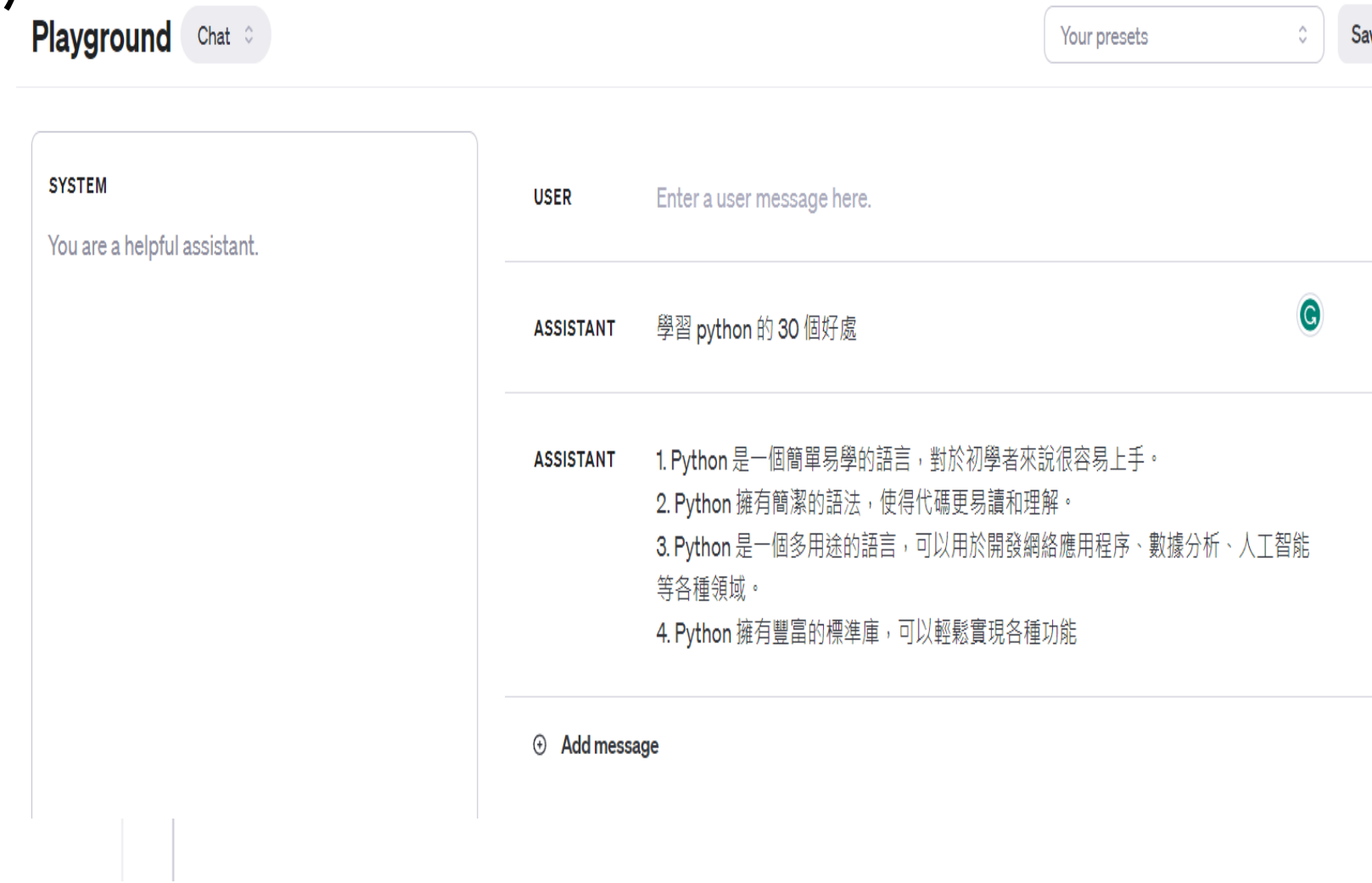
- 預設值為0
- 可能值介於 [-2, 2] 之間
- 數值越高,模型傾向於輸出變化性越高的結果
- 模型會懲罰出現頻率較高的字,所以模型所挑的字會越有變化

Presence_penalty

- 預設值為0
- 可能值介於 [-2, 2] 之間
- 和 frequency penalty 一樣,都是用來控制模型的多樣性
- 和 frequency penalty不同的地方在於 presence penalty 有考慮到 prompt 的影響. 模型會儘量挑出 prompt 中沒有出現的字
- Presence penalty 是以字有沒有出現來決定要步要給予懲罰. Frequency penalty 則是用頻率來決定

參數的使用說明


API re
Playground
Assistants


Playground
Chat
Your presets
SYSTEM
You are a helpful assistant.
USER
Enter a user message here.
ASSISTANT
學習 python 的 30 個好處
ASSISTANT
1. Python 是一個簡單易學的語言，對於初學者來說很容易上手。
2. Python 擁有簡潔的語法，使得代碼更易讀和理解。
3. Python 是一個多用途的語言，可以用於開發網絡應用程序、數據分析、人工智能等各種領域。
4. Python 擁有豐富的標準庫，可以輕鬆實現各種功能
Add message

Best of

- 預設值為1
- 很耗$$
- 一般常和 n 參數一起使用
- 例子:
  - Best of =10, n=1
  - 模型會在 server 端產生 10 個結果
  - 挑選一個最好的回傳給你
- 使用這個參數的好處是不需要重複送 prompt 給server 取回結果並挑選出最好的
- 只要送一次 prompt 給 server 讓 server 幫你挑最好的

Inject start (輸出格式化)

USER	Chinese:你好嗎? English: How are you? Chinese: 這個蘋果好吃嗎? English:
ASSISTANT	Is this apple tasty?