

No of components \Rightarrow Level 3 \rightarrow

Level 1 \rightarrow Level 2 \rightarrow Level 3

I II III

<u>Framework</u>	<u>Lm</u>	<u>Vector Database</u>
\Rightarrow 1) openai	openai	1) pinecone
2) Langchain	Cloud	2) Weaviate
3) Llama index	Gemini	3) Chroma db
		4) FAISS



→ Factors : 1) Requirements - Tex, Image, Multimodal

2) Architecture & Model Size

3) Speed & Latency →

4) Hardware & Resource

5) Cost & Licensing

(Gemini) 6) Responsible Model → Ethical

⊗ Accuracy

⊗ Context Window

⊗ Multimodality support

⊗ Model Size → Accuracy

⊗ Small is Size

* GPT4

* Mistral (Anthropic)

⊗ LLAMA

* Bloom

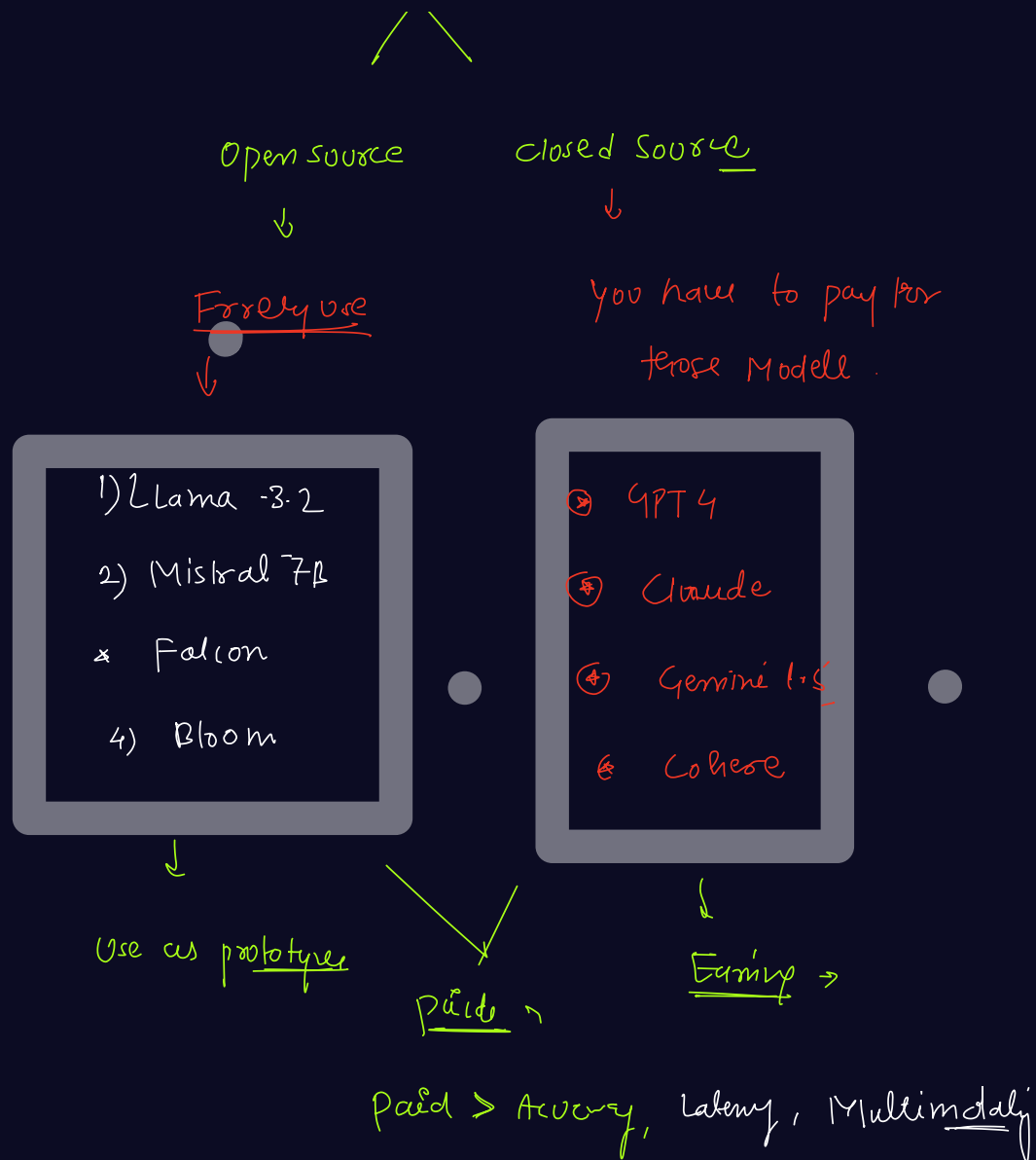
→ Falcon

+ ⊗ BERT

* Gemini

* Claude

* Cohere



④ How to Use Those Models :-

- 1) API Access Free -> gpt, playground
- 2) Download locally. → Ollama, LM studio
- 3) Cloud → free → API. (Aos, gcp)

4) playground

1) Ans Bed Rock

2)

④ → Website, GitHub, Hugging Face