Generative AI with Google cloud

7 Days → 1. Generative AI with Google

2. Vertex AI → Generative AI

3. Build transformative LLM powered application on google cloud.

4. Gemini on vertex AI

5. How to Implement Application with Gemini

6. RAG Implementation with Gemini pro, vector search

7. Fine Tuning Foundation model on GCP

Prerequisite:

→ python

→ Basics of Generative AI

- → LLM
- → Langchain
- → vector DB
- → GCP Account — 300$ ←

→ Langchain

→ llamaIndex

→ open source LLm → GPU based machine

4 GB

RTX→3050 ti

16 GB

LLama2 → 7b → 13b 70b

out of memory

13 min

meta AI

LLmops → Deploy open LLm

AWS → sagemaker

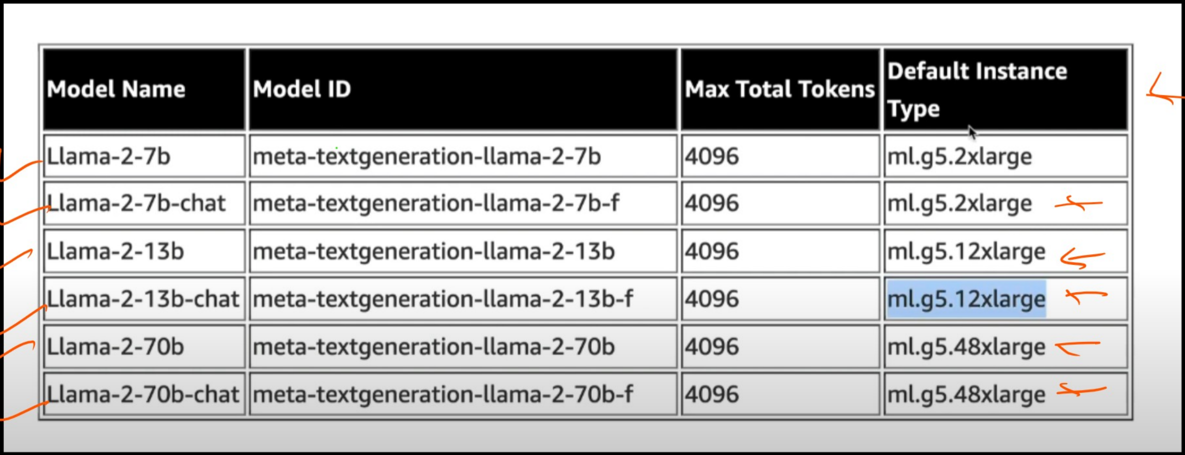
DLC

API Gateway

Lamda

API

openAI → GPT-3,4 → API key Req → ← Res m.



Model Name	Model ID	Max Total Tokens	Default Instance Type
Llama-2-7b	meta-textgeneration-llama-2-7b	4096	ml.g5.2xlarge
Llama-2-7b-chat	meta-textgeneration-llama-2-7b-f	4096	ml.g5.2xlarge
Llama-2-13b	meta-textgeneration-llama-2-13b	4096	ml.g5.12xlarge
Llama-2-13b-chat	meta-textgeneration-llama-2-13b-f	4096	ml.g5.12xlarge
Llama-2-70b	meta-textgeneration-llama-2-70b	4096	ml.g5.48xlarge
Llama-2-70b-chat	meta-textgeneration-llama-2-70b-f	4096	ml.g5.48xlarge

Google

2017 → Transformers

2018 → BERT

2019 → T5 → Text to Text Transfer Transformer
LLM 10B P model

2020 → LaMDA → LLM

2021 → AlphaFold

20222 → Palm —

2023 → Bard —

2024 →

Generative AI

↓

Gcp

↓

Vertex AI