fundamental

1. Roadmap
2. intro
3. Google Gemini, hugging face, openAI API

- Google → Google AI studio
- hugging face → token
- openAI API → { GPT 3.5, GPT 4 }

NLP Related foundational

1. text cleaning
2. encoding & embedding
3. transformer
4. BERT | GPT

Prompt → LLM → output

{ gpt, gemini, mistral, llama. }

Prompt: set of instruction

LLM ↑ trained (huge)

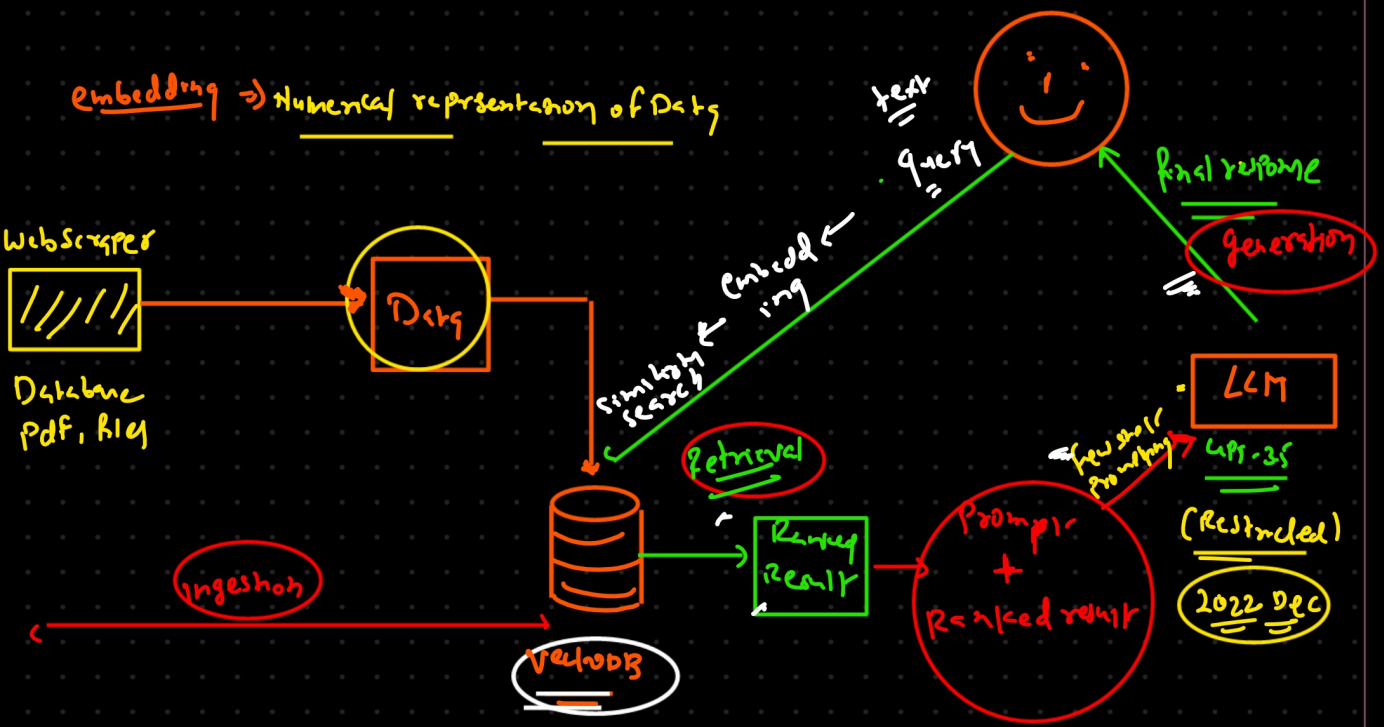
text cleaning

1. RAG (external resource)
2. finetuning ⇒ (task specific)
   - finetuning → (Parameter)
   - full finetuning
     - { instruction fine tuning }
     - supervised fine tuning
   - PEFT
     - LoRA
     - QLoRA
3. RLHF
4. quantization

Cosine Similarity   Dot Product   Jaccard Similarity

Embedding ⇒ ① [ ] ② [ ]

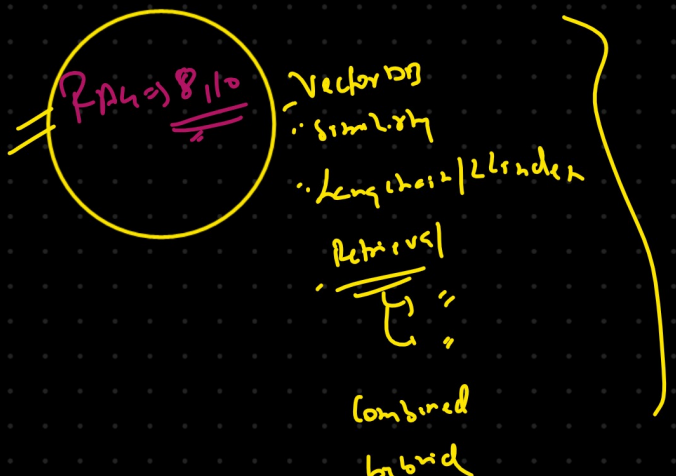
Proof



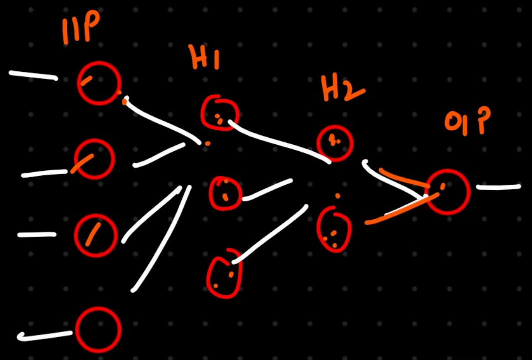
RAG

1. Ingestion
2. retrieval
3. generation

RAG ⇒ Retrieval Augmented Generation



Finetuning ⇒ technically



4×3=12 3×2 2×1

12 + 3 6 + 2 2 + 1

15 + 8 + 3 ⇒ 26

Parameter ⇒ (weight + bias)

↑ trainable

Params = 26

↑ trainable params

ANN or NN

↓

Basic unit

↳ CNN

↳ RNN

↳ transformer

↳ GAN

↳ RL

① FP (Dot Prod + Act + Loss)

② BP (optimizer)

↳ GD, Adam, RMS Prop, Adagrad

minimize Loss ↓↓↓

1000 billion → 1.76 Trillion (size of model + Data)

Model	Provider	Open-Source	Speed	Quality	Params	Fine-Tuneability
GPT-4	Open AI	☒	★☆☆	★★★★	1.76T	☒
GPT-3.5 Turbo	Open AI	☒	★★☆	★★★☆	175B	✓
GPT-3	Open AI	☒	★☆☆	★★★☆	175B	☒
Claude 3 Opus	Anthropic	☒	★★☆	★★★★	Not specified	☒
Claude 3 Sonnet	Anthropic	☒	★★☆	★★★☆	Not specified	☒
Claude 3 Haiku	Anthropic	☒	★★★	★★☆☆	Not specified	☒
Command Nightly	Cohere	☒	★★☆	★★★☆	52B	✓
BERT	Google	✓	★★★	★☆☆☆	345M	✓
T5	Google	✓	★★☆	★☆☆☆	11B	✓
PaLM	Google	☒	★☆☆	★★☆☆	540B	✓
Gemini Nano	Google	☒	★★☆	★★☆☆	1.76T	Coming soon
Gemini Pro	Google	☒	★★☆	★★☆☆	175B	Coming soon
Gemini Ultra	Google	☒	★☆☆	★★★★	175B	Coming soon
LLaMA	Meta AI	✓	★★☆	★★☆☆	65B	✓
Llama 2 7B	Meta AI	✓	★★☆	★★☆☆	7B	✓
Llama 2 13B	Meta AI	✓	★★☆	★★☆☆	13B	✓
Llama 2 70B	Meta AI	✓	★☆☆	★★★☆	70B	✓
Mistral 7B	Mistral AI	✓	★★☆	★★☆☆	7.3B	✓
Mixtral 8×7B	Mistral AI	✓	★★☆	★★★☆	46.7B	✓
Orca	Microsoft	✓	★★☆	★★★☆	13B	?
Falcon 40B	Falcon LLM	✓	★★☆	★★☆☆	40B	✓
Falcon 180B	Falcon LLM	✓	★★☆	★★★☆	180B	✓

NN → weight + Bias ⇒ Parameter = Data

↑ (BP)

trainable Parameter

LLM

{ GPT ⇒ 175B
LLama ⇒ 70B
mistral ⇒ 45B
⋮ }

Finetuning ⇒ Retrain the Parameter of Network (model)

Full finetuning

Set of Parameter (PEFT)

Eg

Biggerusecase

Domain

= Finetuning

= RAG

External Data source ∃ eng.

Weight trial (Resource)

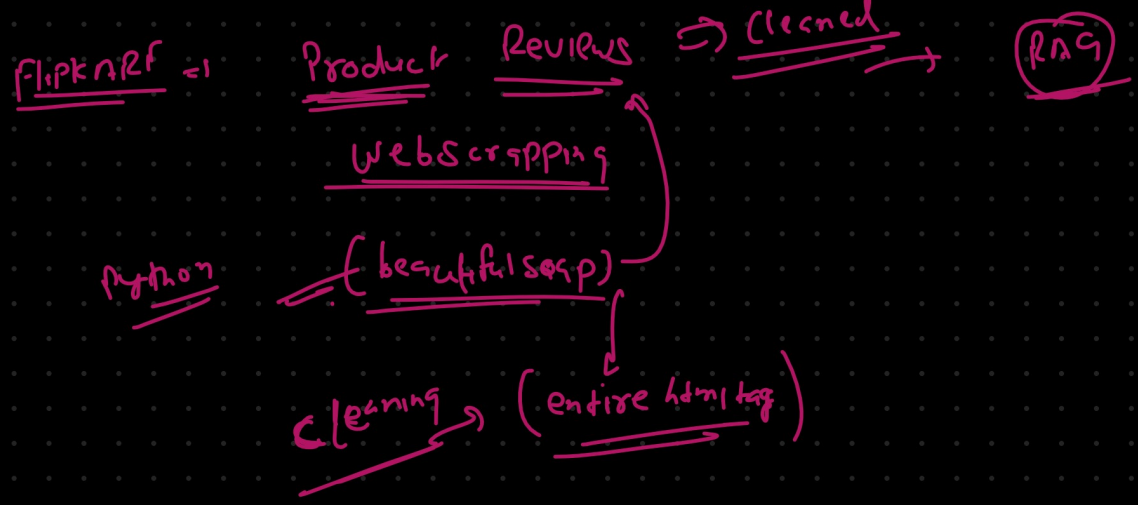
95/100

90/100

Combined approach

AI

Finetuning ⇒ RAG App


Flipkart
Product
Reviews
Cleaned
RAG
Webscrapping
python
beautifulsoup
cleaning
entire html tag