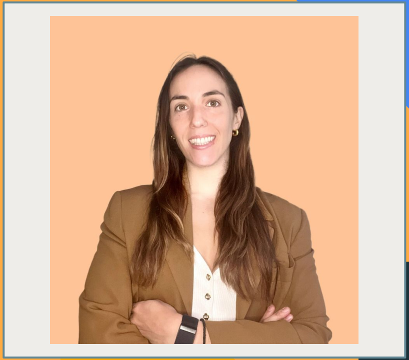
Life Beyond ChatGPT

From RAG to fine tuned models with enterprise data

Presenter



Maria Zervou

Sr. Specialist Solutions Architect

Agenda

- New Wave of Deep Learning
- Customisation Phases of GenAI
- Build a RAG application
- Live demo
- Fine Tuning Concepts
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary – Call to Action

Look out for....

Look out for....

Expert

Newbie

Agenda

- **New Wave of Deep Learning**
- Customisation Phases of GenAI
- Build a RAG application
- Live demo
- Fine Tuning Concepts
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary - Call to Action

What are GenAi Models

Large Language Models

Diffusion Models

What are GenAi Models

LARGE AI MODELS

To Train Generative AI models we need…..

GPUs

How many do you think we need?

To Train Generative AI models we need…..

GPUs

We need a lot of GPUS to train <u>your own</u> Generative AI models

Transformer Models

Fueling the next wave of Deep Learning

How Gen Ai will disrupt you?

How it will disrupt you?

Data + GenAi = Huge business value

Create Conversational Interfaces for everything

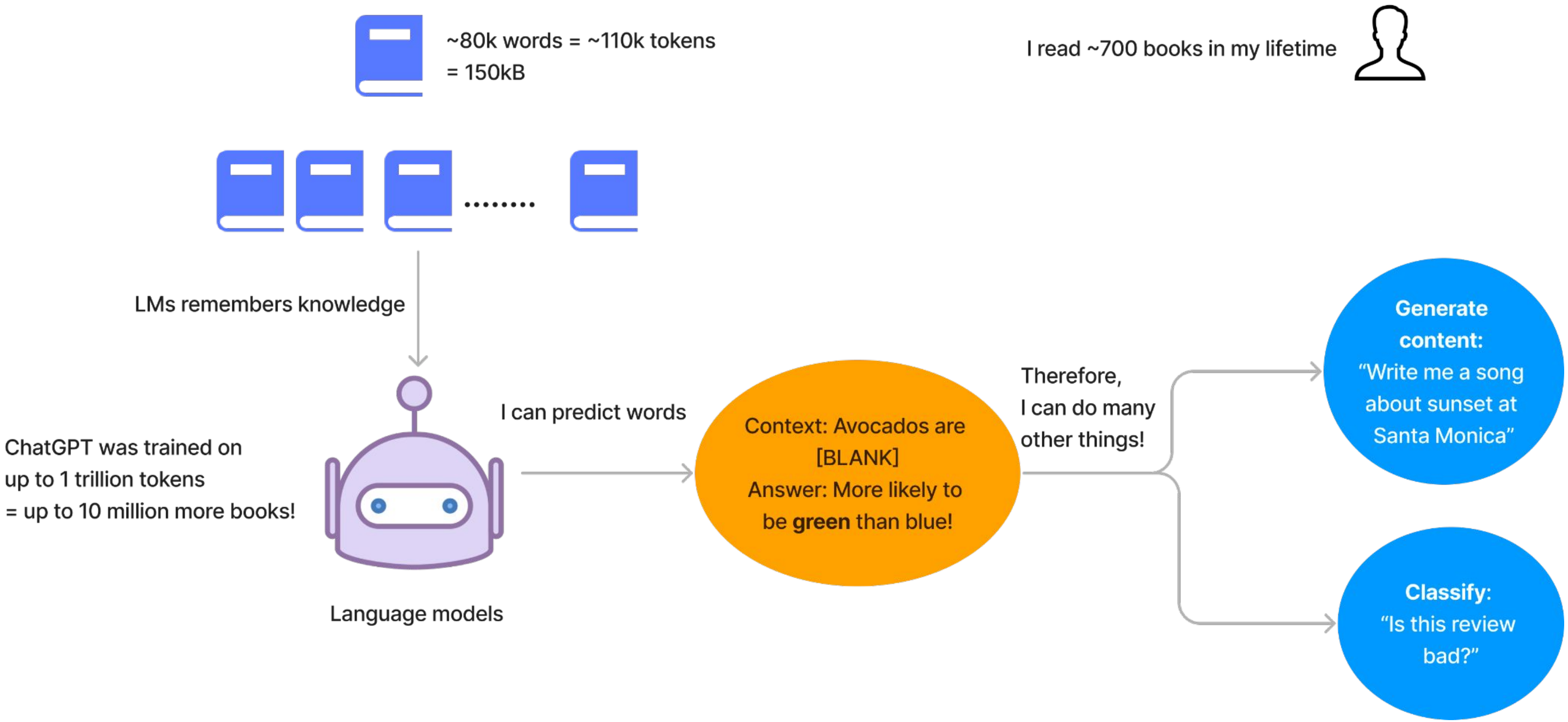
How it will disrupt you?

Data + GenAi = Huge business value

Create Conversational Interfaces for everything

Human Level Comprehension but faster #Notjustachatbot

Human Level Comprehension



How it will disrupt you?

Data + GenAi = Huge business value

Create Conversational Interfaces for everything

Human Level Comprehension but faster #Notjustachatbot

Generate Human Quality Text and Images

So where do we start?

Agenda

- New Wave of Deep Learning
- **Customisation Phases of GenAI**
- Build a quick RAG application with your own data
- Live demo
- Fine Tuning Concepts
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary - Call to Action

Customisation Phases of GenAi

Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)

Customisation Phases of GenAi

Mild

Extra Hot

Foundational model with Prompts
Retrieval Augmented knowledge (RAG)
Fine-tune foundational model on your data
Fully retrain foundational models (your own “GPT”)

Kpis Time

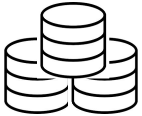
Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Training Cost



Data Size

Know-how

Customisation

Kpis Time

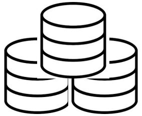
Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)

Training Cost



Data Size

Extra Hot

Know-how

High

Customisation

Proprietary LLMs

Proprietary LLMs

PaLM 2 · cohere · ChatGPT · OpenAI · ANTHROPIC

Pros:

- Access to state-of-the-art models (e.g. GPT4).
- Easy to use: No need for advanced knowledge on LLM.

Proprietary LLMs

Proprietary LLMs

ChatGPT PaLM 2 cohere OpenAI ANTHROPIC

Pros:

- Access to state-of-the-art models (e.g. GPT4)
- Easy to use: No need for advanced knowledge on LLM.

Cons:

- Model updates and service SLA depend fully on the providers.
- Quality of inference can vary significantly over time [ref].
- You **may** need to send your data to the providers (but not always).

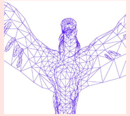
Open Source LLMs

stability.ai
Stable Diffusion

Open Source LLMs

mosaicML
MPT

Hugging Face



databricks
Dolly

Open source models customers can get from an open repository like Hugging Face.

Pros:

- Access to many general & specialized models.
- Data and models stay within your environment.
- Flexibility in tuning latency and throughput by changing the inference cluster configuration
- Transparency in the source code, model weights, and training dataset
- Fine-tuning is possible

Open Source LLMs

Open source models customers can get from an open repository like Hugging Face.

Pros:

- Access to many general & specialized models.
- Data and models stay within the customers' environment.
- Flexibility in tuning latency and throughput by changing the inference cluster configuration
- Transparency in the source code, model weights, and training dataset
- Fine-tuning is possible

Cons:

- Requires expertise for tuning models and selecting the right infrastructure.

Foundational Models

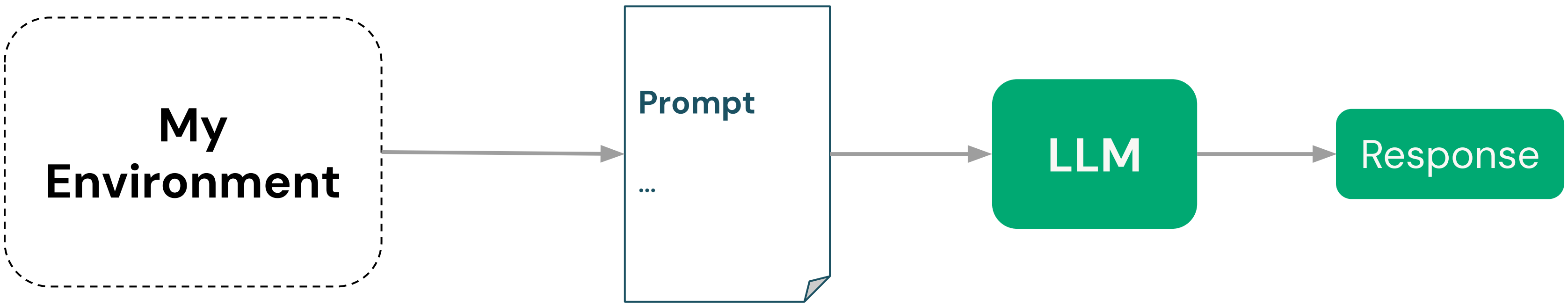
Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Prompt Engineering



Kpis Time – Foundational Models

Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Training Cost

Data Size

Extra Hot

Know-how

High

Customisation

Kpis Time – Foundational Models

Foundational model with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Mild

Low

Extra Hot

High

Training Cost

Data Size

Know-how

Customisation

Customisation Phases of GenAi

Saas with Prompts

Foundational model with Prompts	- **Ready to be plugged** to your applications - Can modify your output with your own **Prompts** - Can **generate "any answer"** - You **don't control data inside** the model knowledge base - You **cannot control the model and version** - You **cannot control ownership**

Customisation Phases of GenAi – RAG

Foundational model as SaaS with Prompts

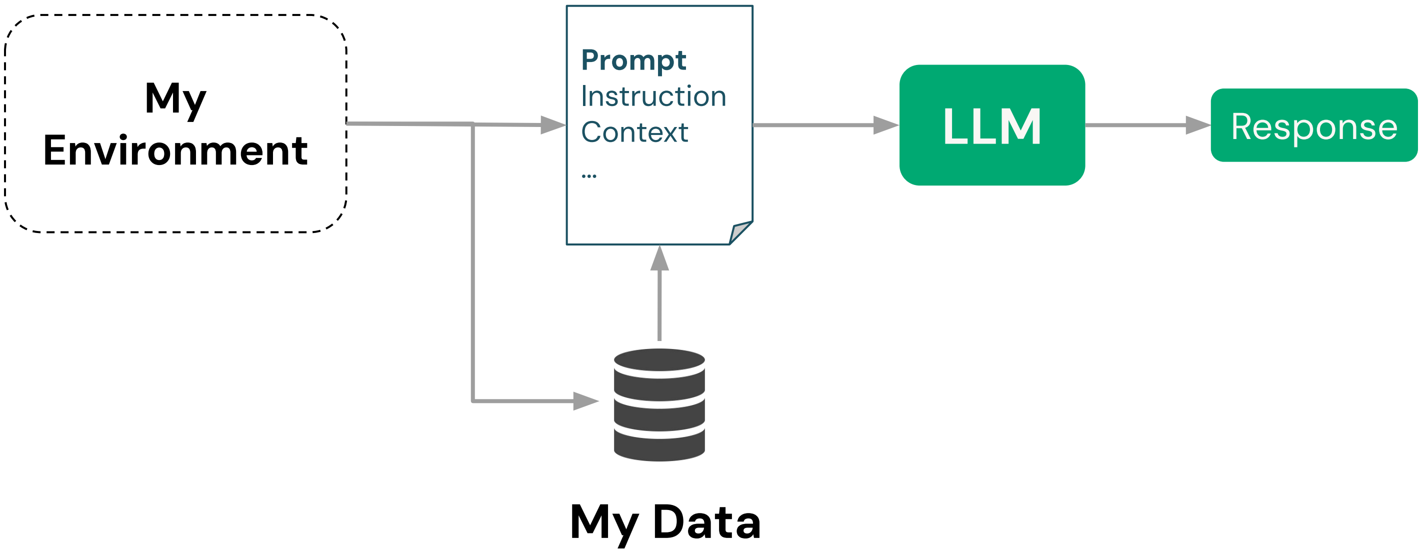
Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Provide your data as you are calling the model

RAG Architecture

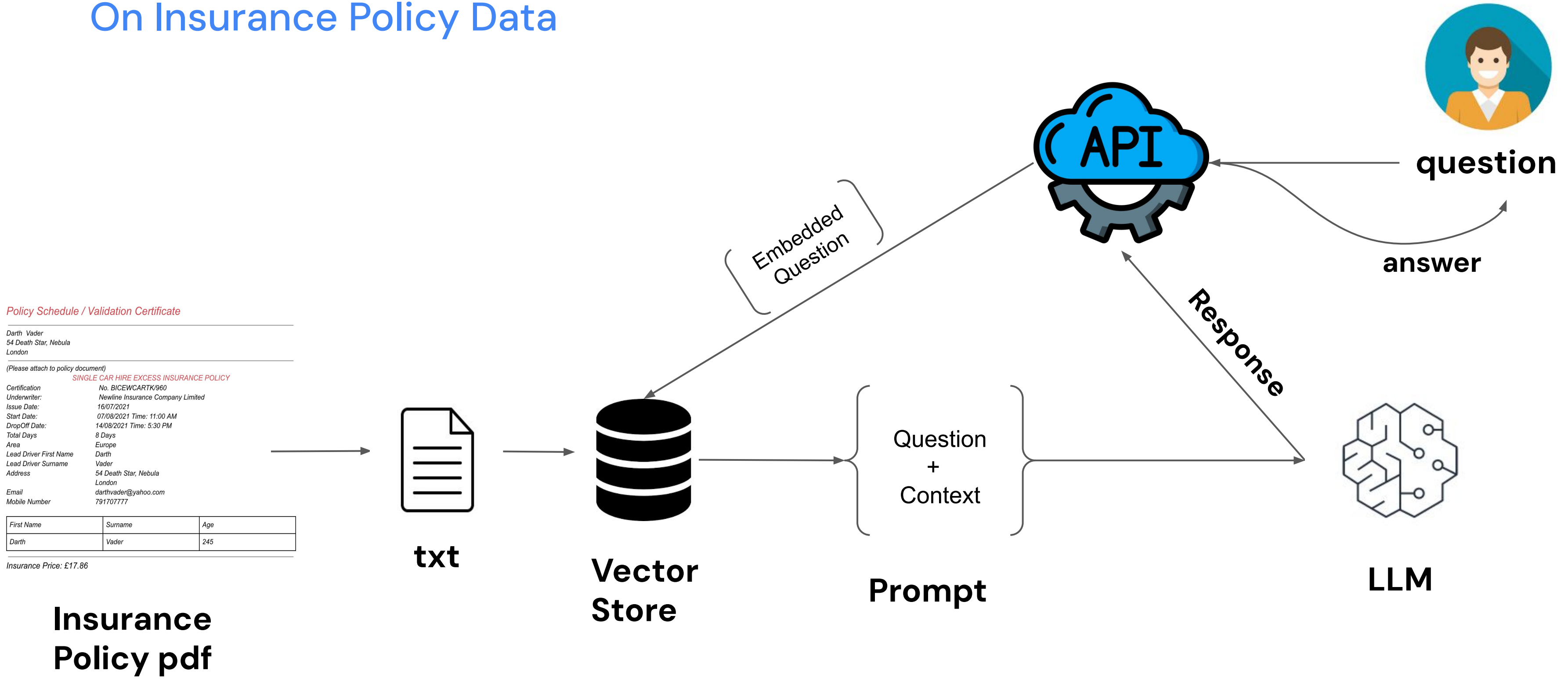


Agenda

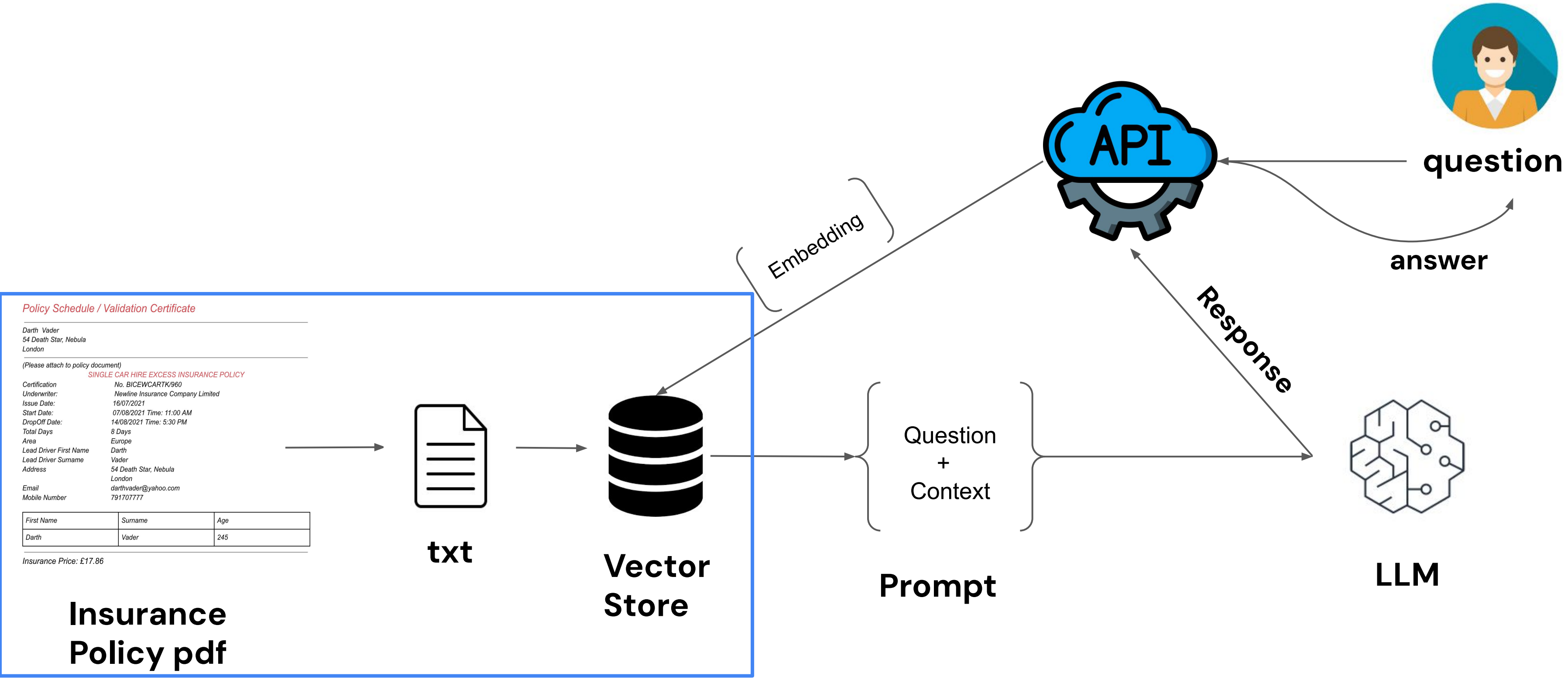
- New Wave of Deep Learning
- Customisation Phases of GenAI
- **Build a quick RAG application with your own data**
- Fine Tuning Concepts
- Live demo
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary – Call to Action

RAG Architecture for Q/A Bot

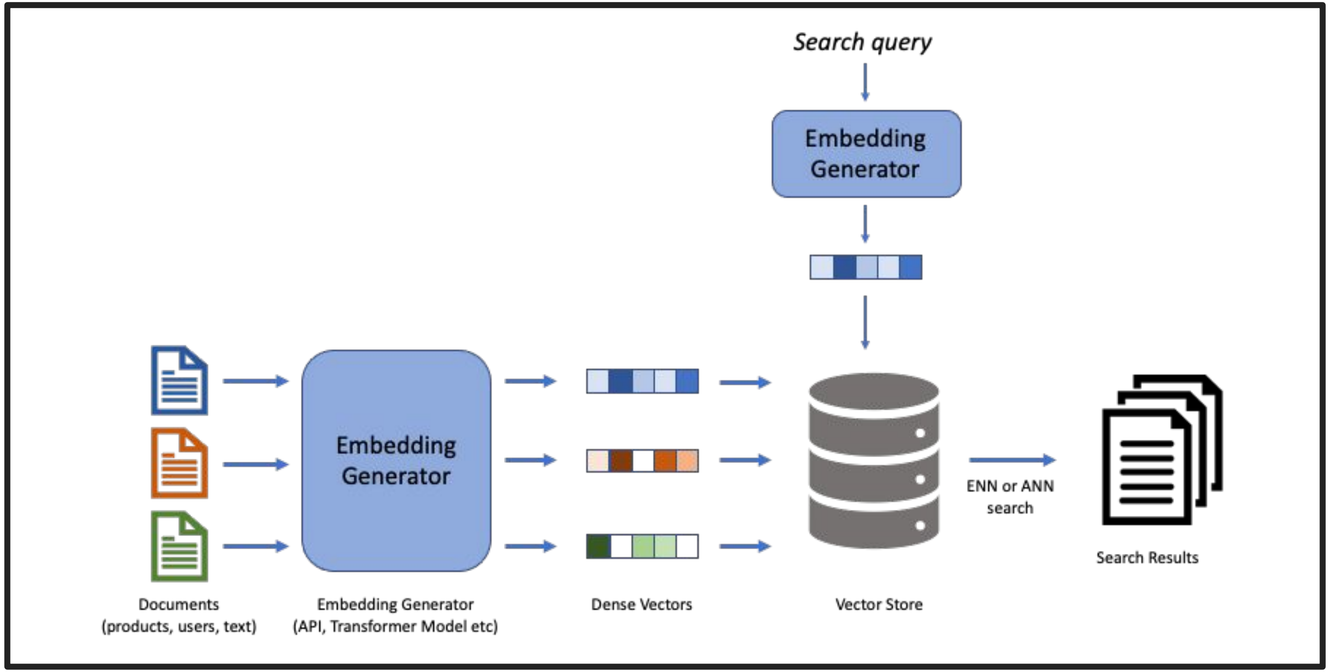
On Insurance Policy Data



Document Indexing

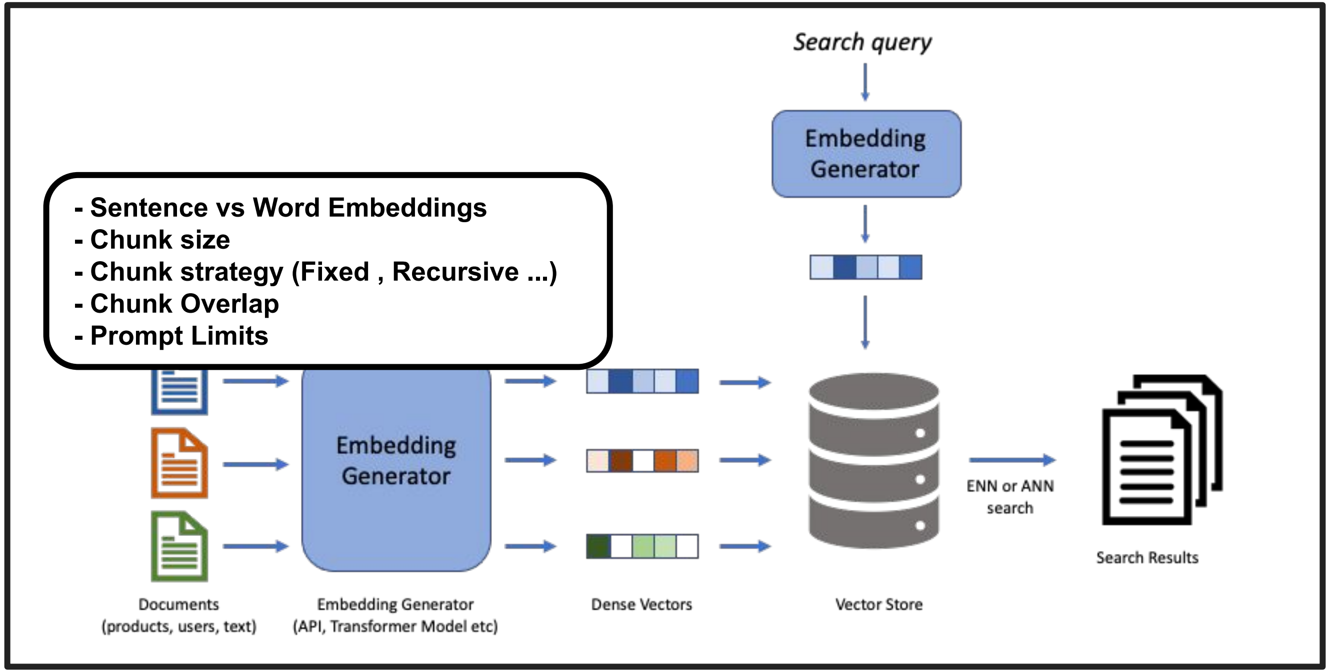


Document Indexing

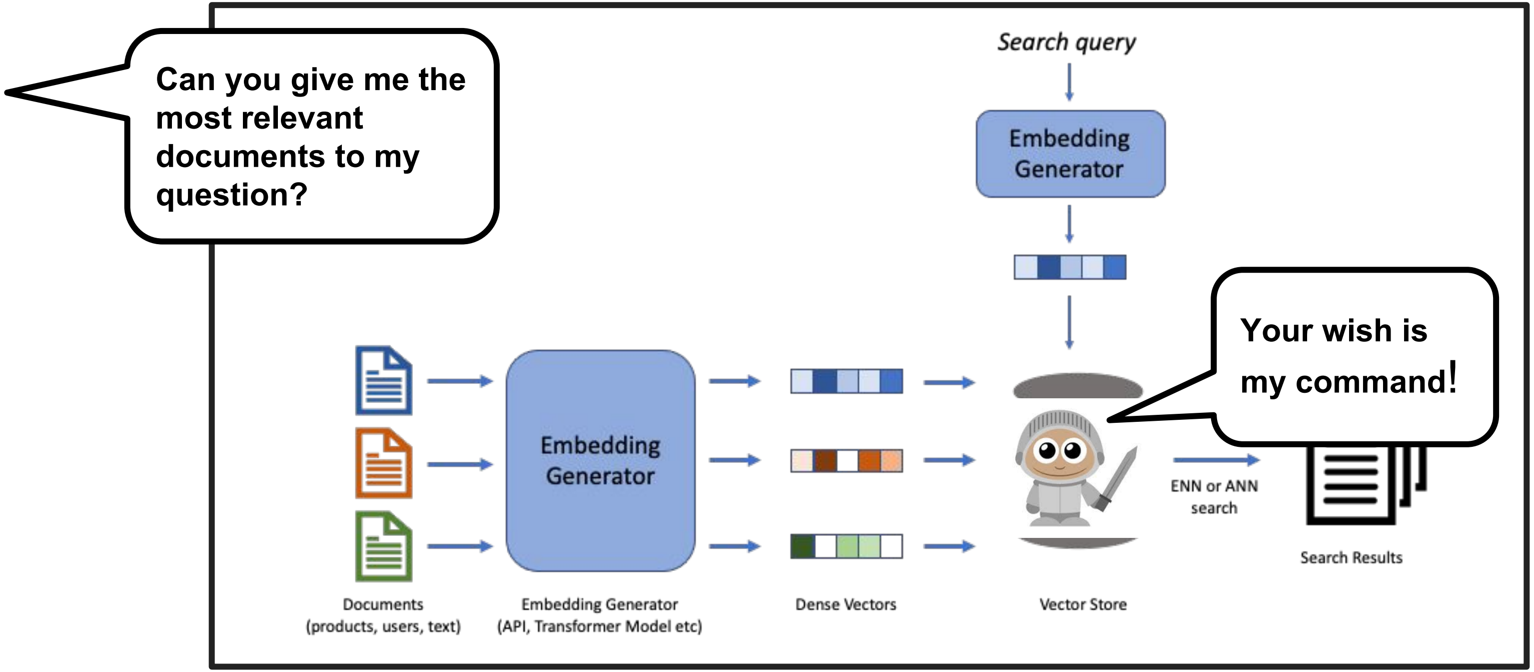


Embeddings

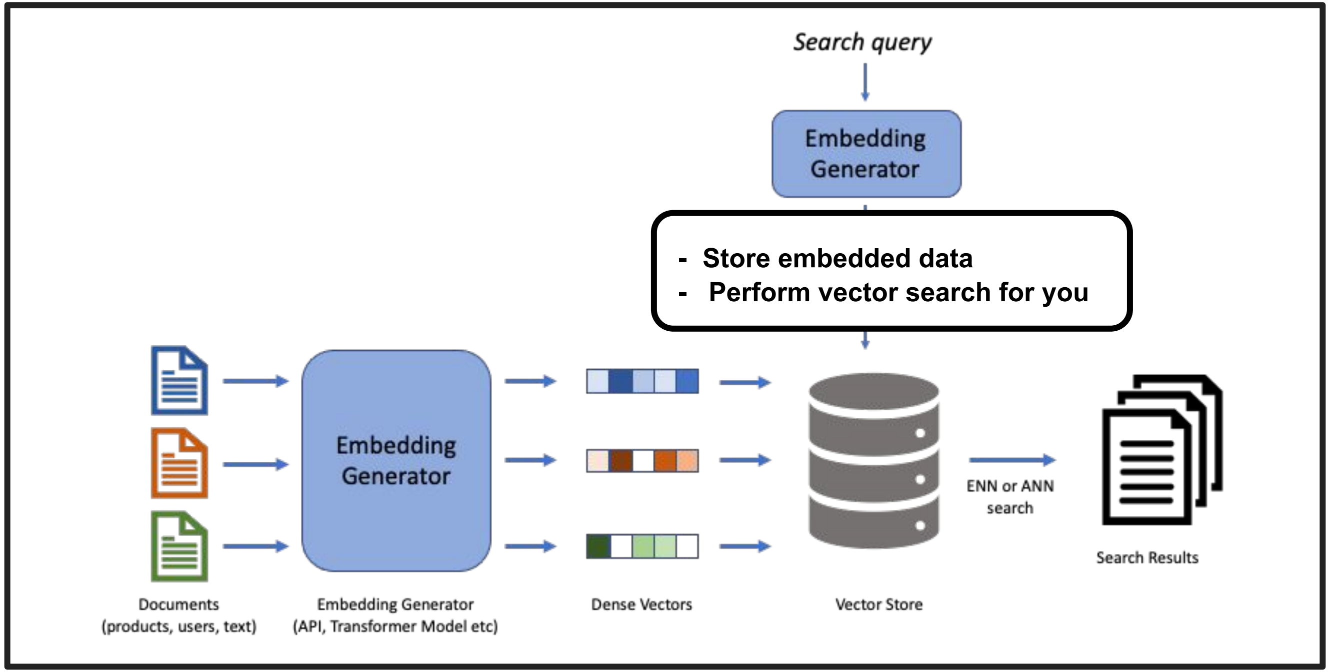
Translate text to fixed size vectors



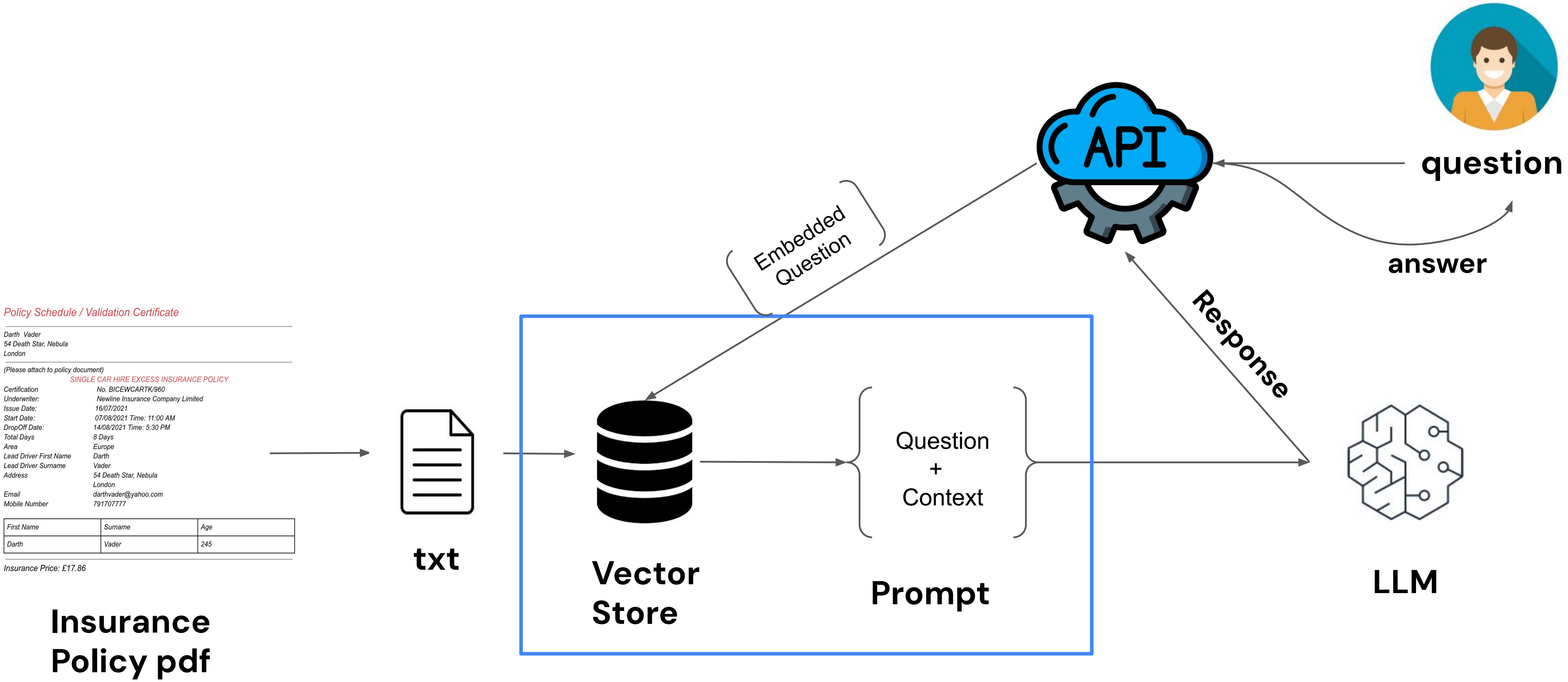
Vector Index



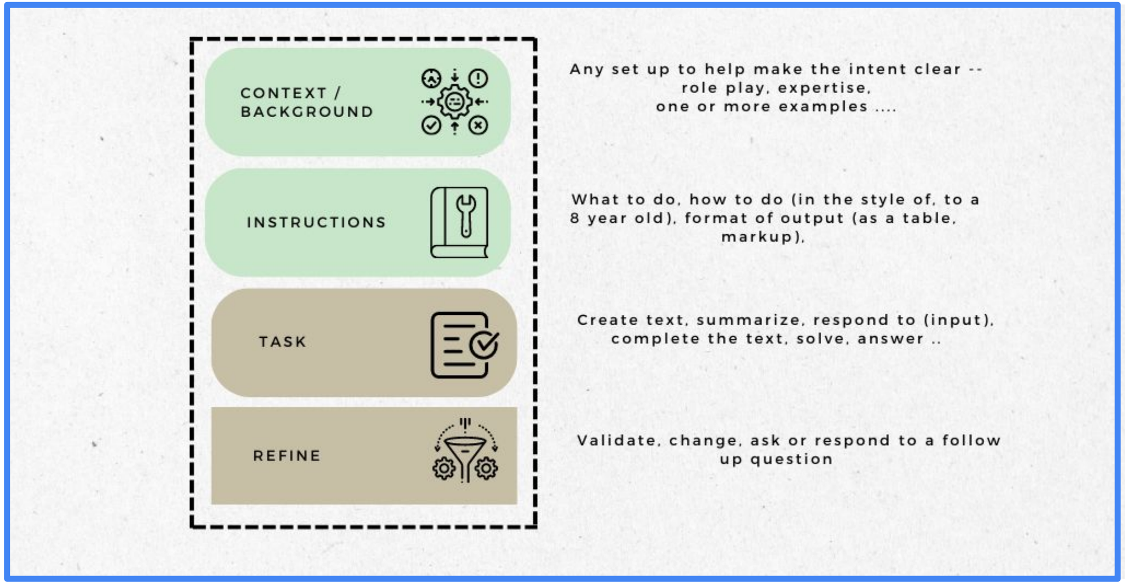
MVP Steps – Vector Store



Prompt Engineering



Prompt Engineering – Anatomy of a Prompt



Prompt Engineering- Anatomy of a Prompt

```
You are a helpful assistant and you are good at helping to answer a
question based on the context provided, the context is a document.

If the context does not provide enough relevant information to
determine the answer, just say I don't know. If the context is
irrelevant to the question, just say I don't know. If you did not
find a good answer from the context, just say I don't know. If the
query doesn't form a complete question, just say I don't know.

If there is a good answer from the context, try to summarize the
context to answer the question.
```

The **"act as"** hack

Prompt Engineering – Anatomy of a Prompt

```
You are a helpful assistant and you are good at helping to answer a
question based on the context provided, the context is a document.

If the context does not provide enough relevant information to
determine the answer, just say I don't know. If the context is
irrelevant to the question, just say I don't know. If you did not
find a good answer from the context, just say I don't know. If the
query doesn't form a complete question, just say I don't know.

If there is a good answer from the context, try to summarize the
context to answer the question.
```

Instructions

Prompt Engineering – Anatomy of a Prompt

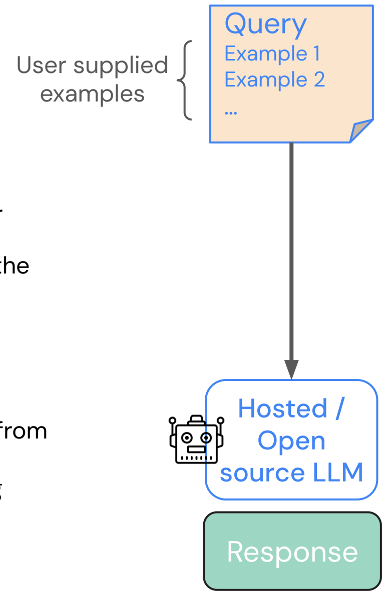
```
You are a helpful assistant built by Databricks, you are good at
helping to answer a question based on the context provided, the
context is a document.

If the context does not provide enough relevant information to
determine the answer, just say I don't know. If the context is
irrelevant to the question, just say I don't know. If you did not
find a good answer from the context, just say I don't know. If the
query doesn't form a complete question, just say I don't know.

If there is a good answer from the context, try to summarize the
context to answer the question.
```

Task:
Summarisation

Prompt engineering patterns

“Add Context to the Query”

Few shot prompting



Prompts can be:

Natural language sentences or questions. Code snippets or commands. Combinations of the above.

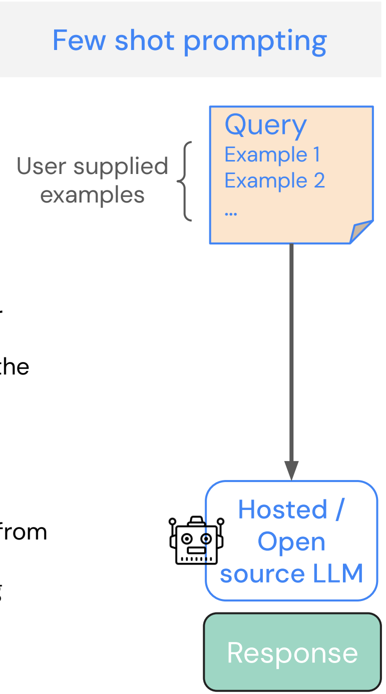
Emojis....basically any text!

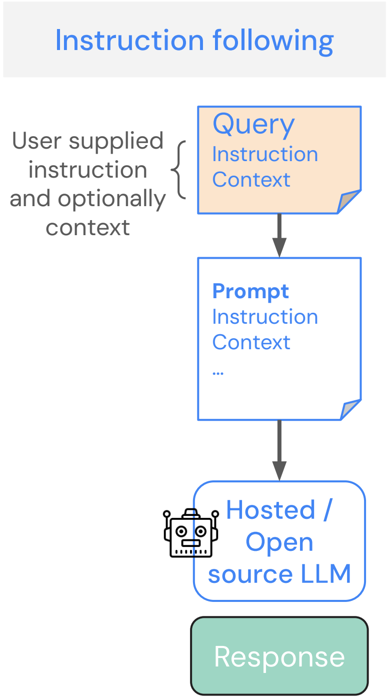
Prompts can include outputs from other LLM queries.
This allows nesting or chaining LLMs, creating complex and dynamic interactions.

Note: The patterns can be combined

Prompt engineering patterns

“Add Context to the Query”





Prompts can be:

Natural language sentences or questions. Code snippets or commands. Combinations of the above.

Emojis....basically any text!

Prompts can include outputs from other LLM queries.
This allows nesting or chaining LLMs, creating complex and dynamic interactions.

Note: The patterns can be combined

Prompt engineering patterns

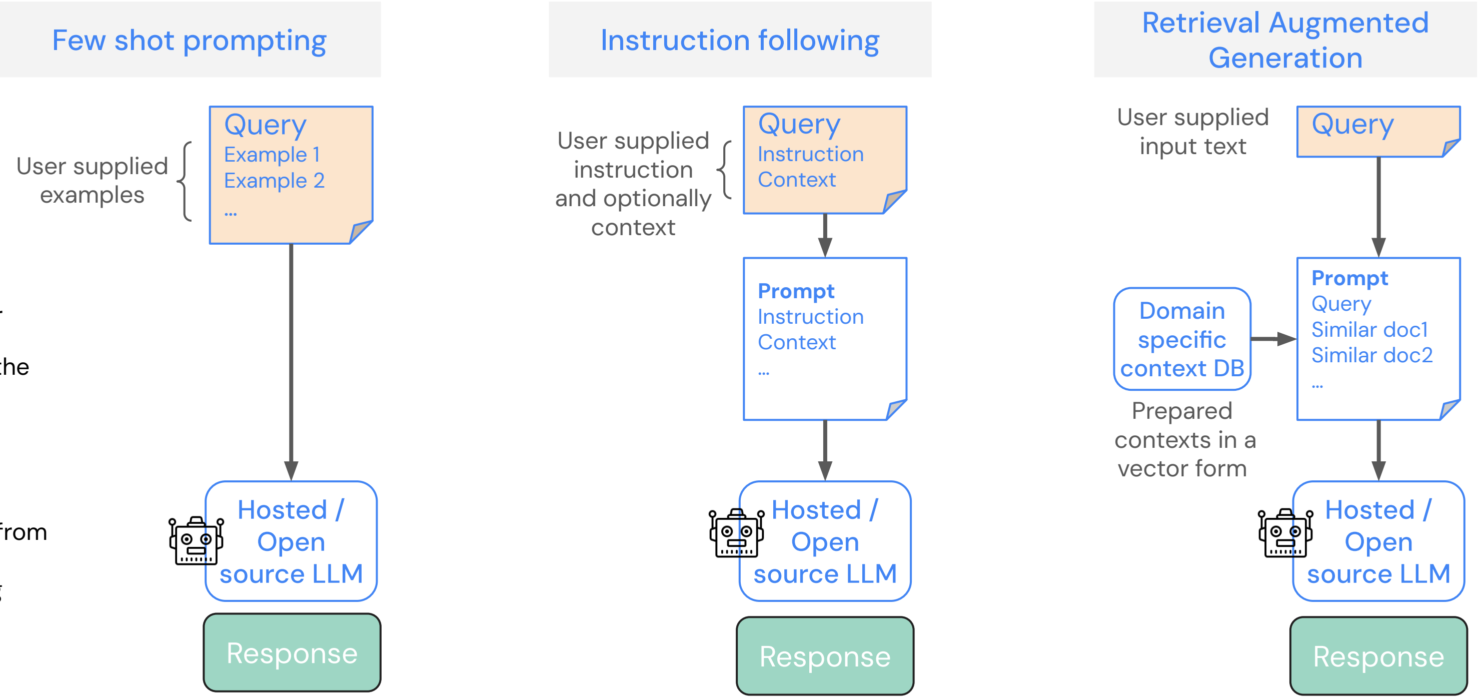
"Add Context to the Query"

Prompts can be:

Natural language sentences or questions. Code snippets or commands. Combinations of the above.

Emojis....basically any text!

Prompts can include outputs from other LLM queries.
This allows nesting or chaining LLMs, creating complex and dynamic interactions.



Note: The patterns can be combined

Prompt engineering patterns

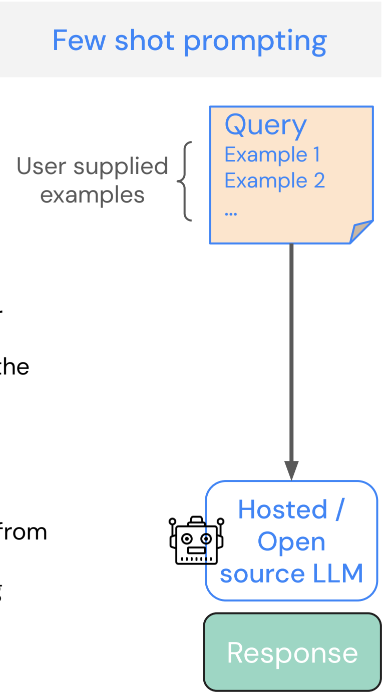
“Add Context to the Query”

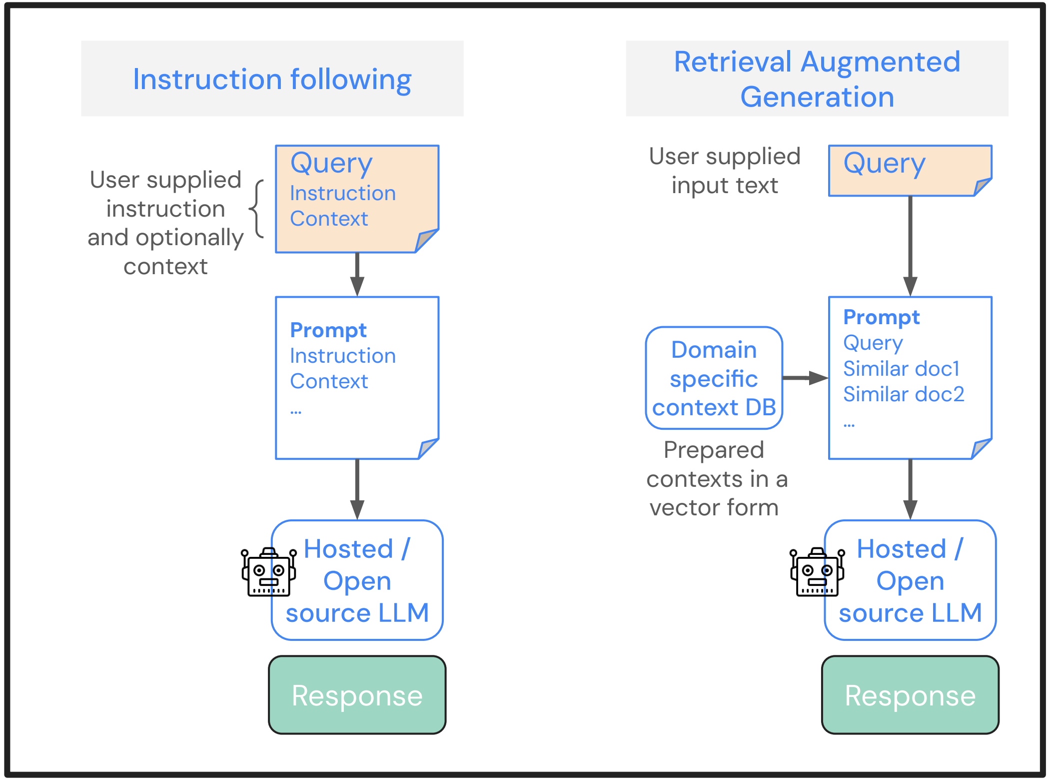
Prompts can be:

Natural language sentences or questions. Code snippets or commands. Combinations of the above.

Emojis....basically any text!

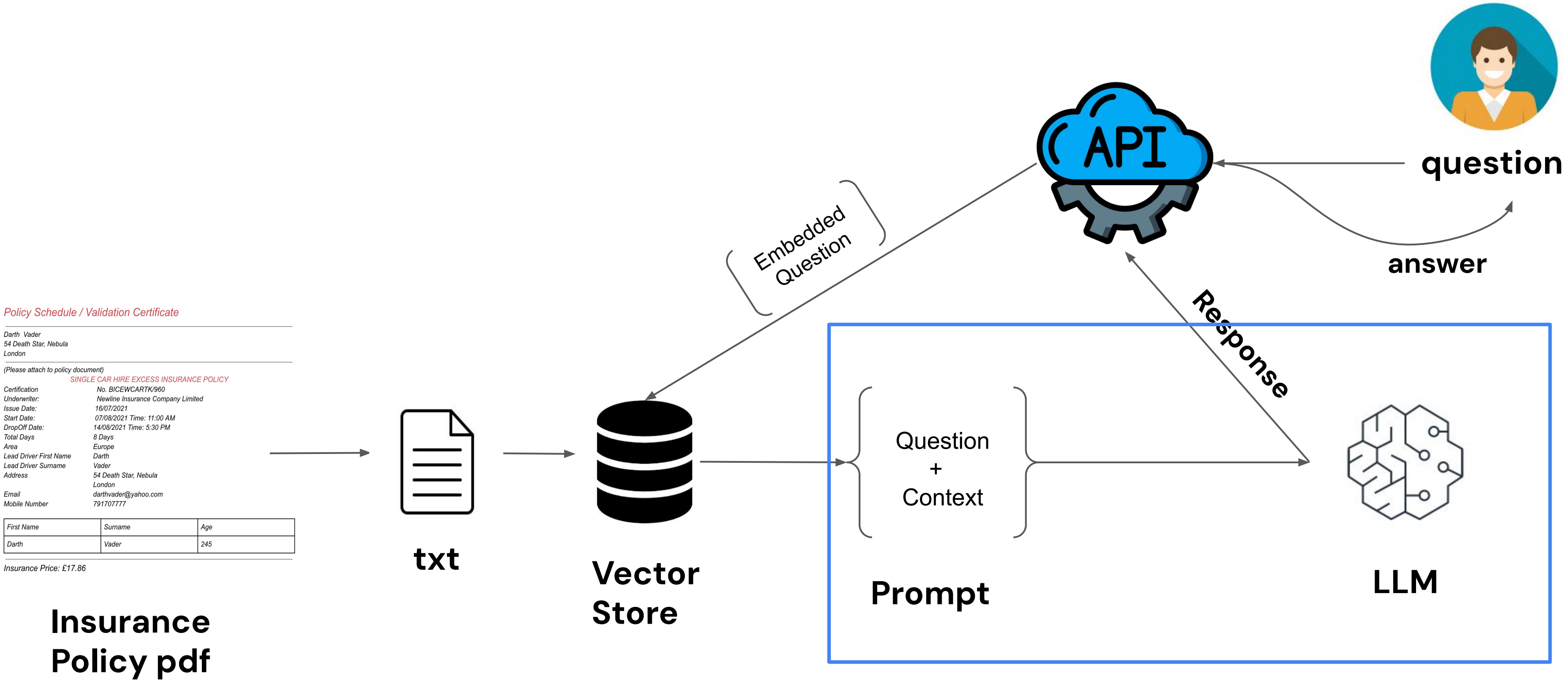
Prompts can include outputs from other LLM queries.
This allows nesting or chaining LLMs, creating complex and dynamic interactions.





Note: The patterns can be combined

LLM Time


Policy Schedule / Validation Certificate
Darth Vader
54 Death Star, Nebula
London
(Please attach to policy document)
SINGLE CAR HIRE EXCESS INSURANCE POLICY
Certification No. BICEWCARTK/960
Underwriter: Newline Insurance Company Limited
Issue Date: 16/07/2021
Start Date: 07/08/2021 Time: 11:00 AM
DropOff Date: 14/08/2021 Time: 5:30 PM
Total Days 8 Days
Area Europe
Lead Driver First Name Darth
Lead Driver Surname Vader
Address 54 Death Star, Nebula
London
Email darthvader@yahoo.com
Mobile Number 791707777
First Name Surname Age
Darth Vader 245
Insurance Price: £17.86
Insurance Policy pdf
txt
Vector Store
Embedded Question
API
question
answer
Response
Question + Context
Prompt
LLM

A Typical LLM Release

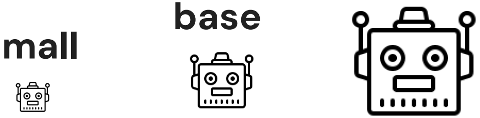
Multiple **sizes** (foundation/base model):

Size means memory required to load / train the model

A Typical LLM Release

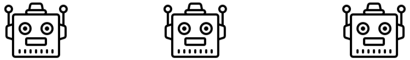
Multiple **sizes** (foundation/base model):

small **base** **large**



Size means memory required to load / train the model

Multiple **sequence lengths**:



512 **4096** **62000**

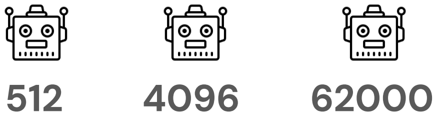
Length you can learn from / use to generate text.

A Typical LLM Release

Multiple **sizes** (foundation/base model):

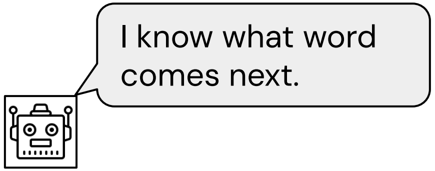
Size means memory required to load / train the model

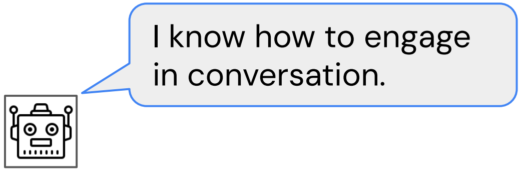
Multiple **sequence lengths:**

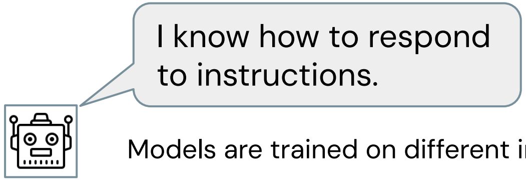


Length you can learn from / use to generate text.

Flavors/fine-tuned versions (**base**, **chat**, **instruct**):



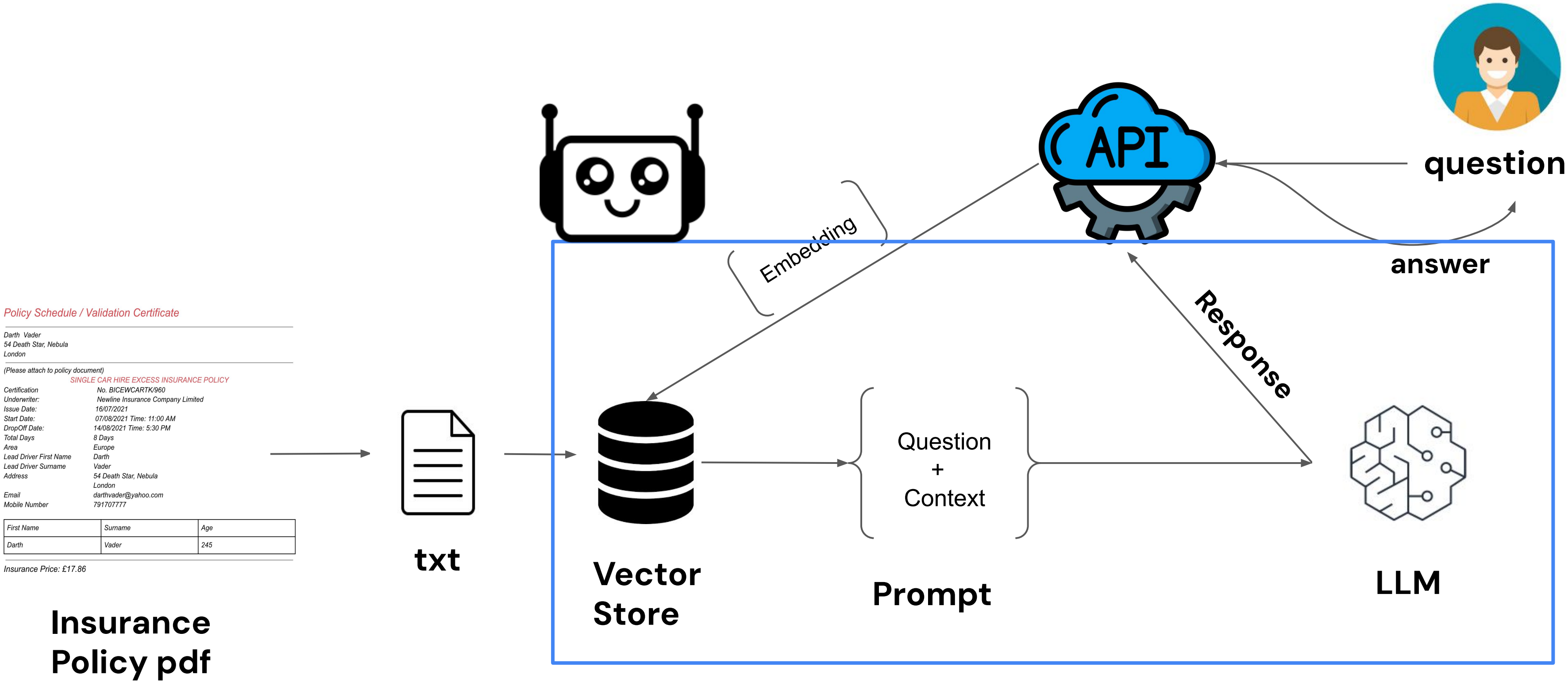




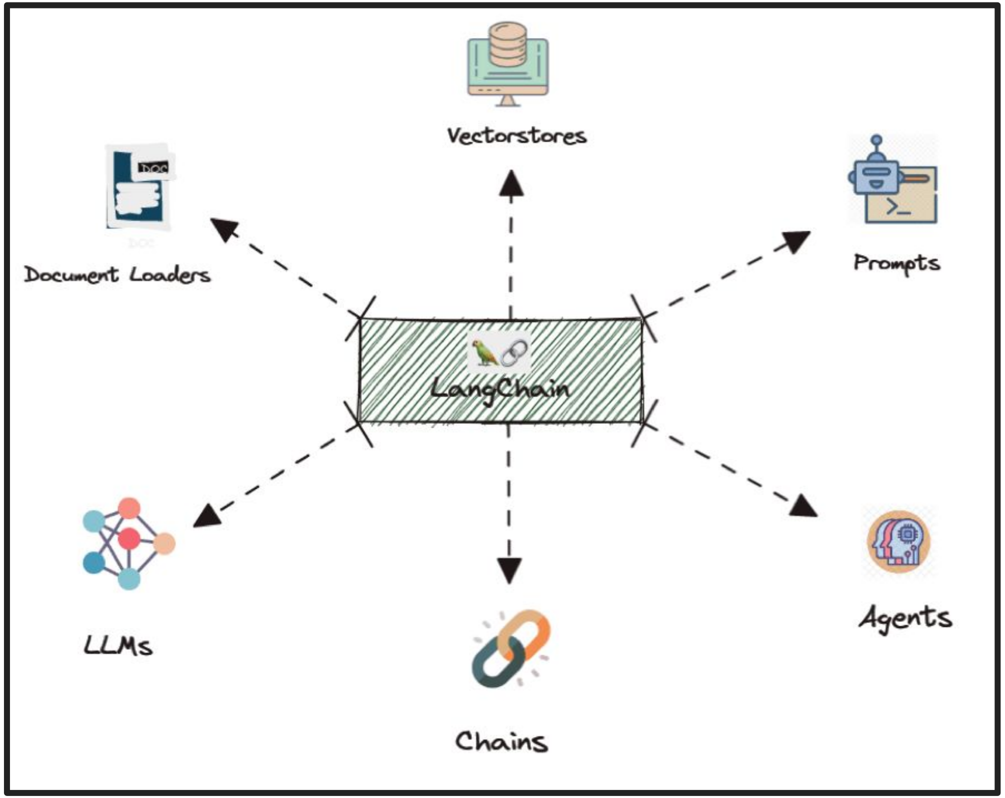
Models are trained on different instructions.

What model are we gonna use?

Application Time



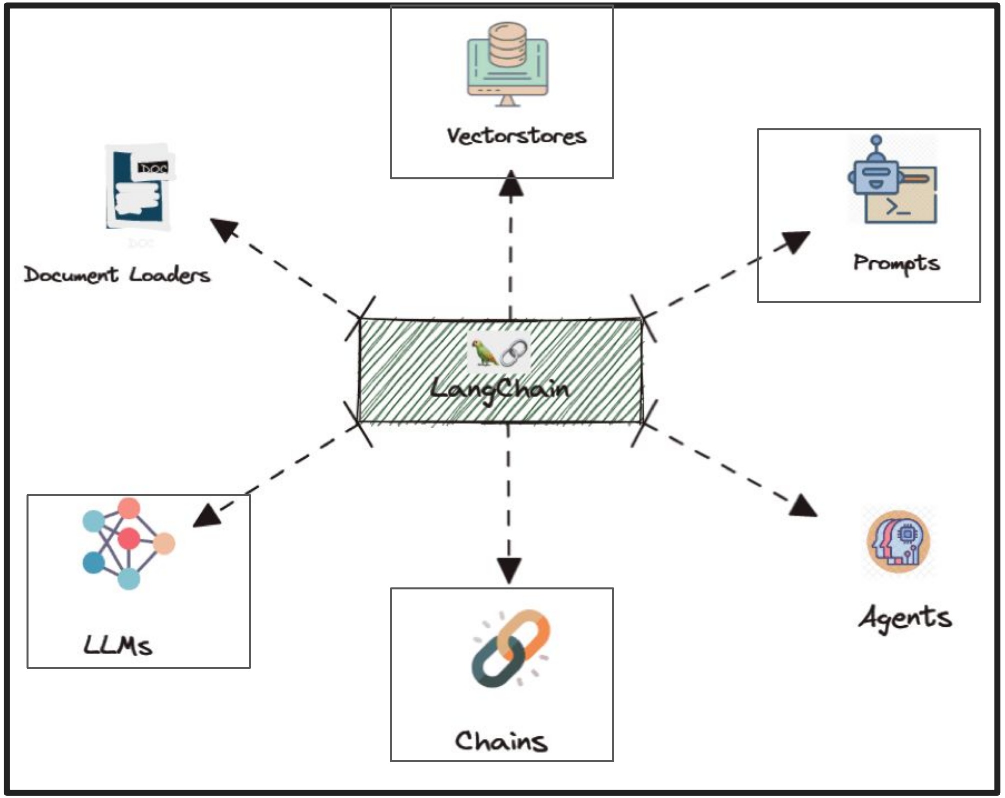
Chaining



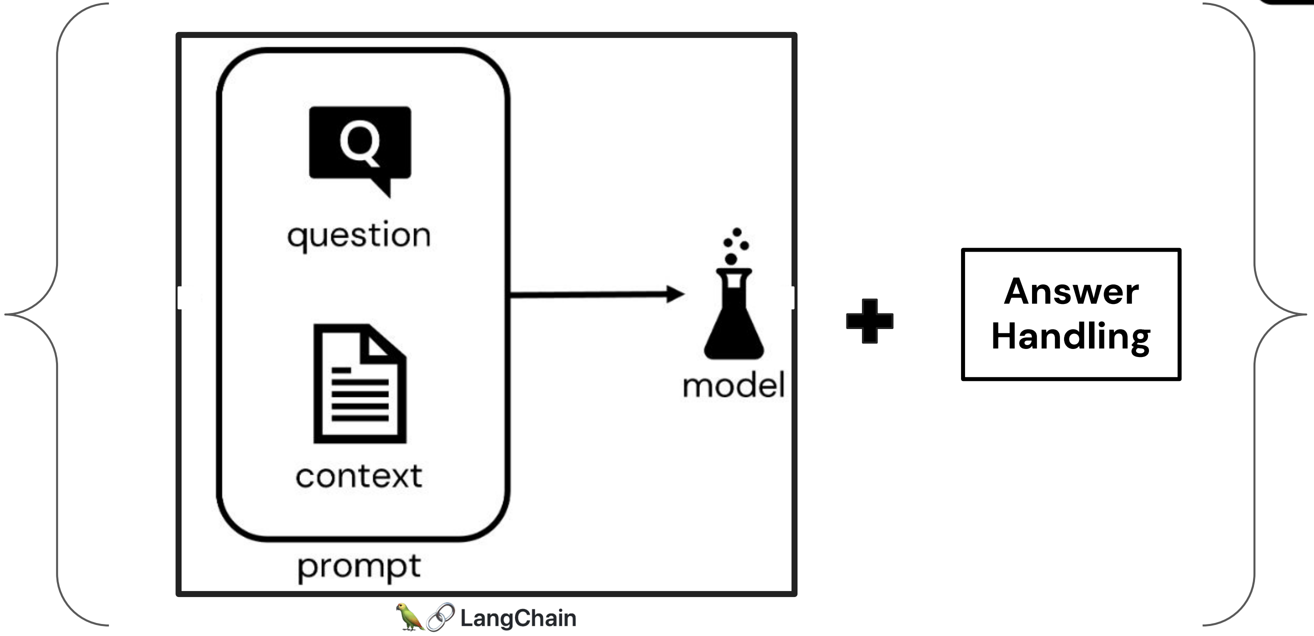
Vectorstores
Document Loaders
Prompts
LangChain
LLMs
Chains
Agents

Chaining

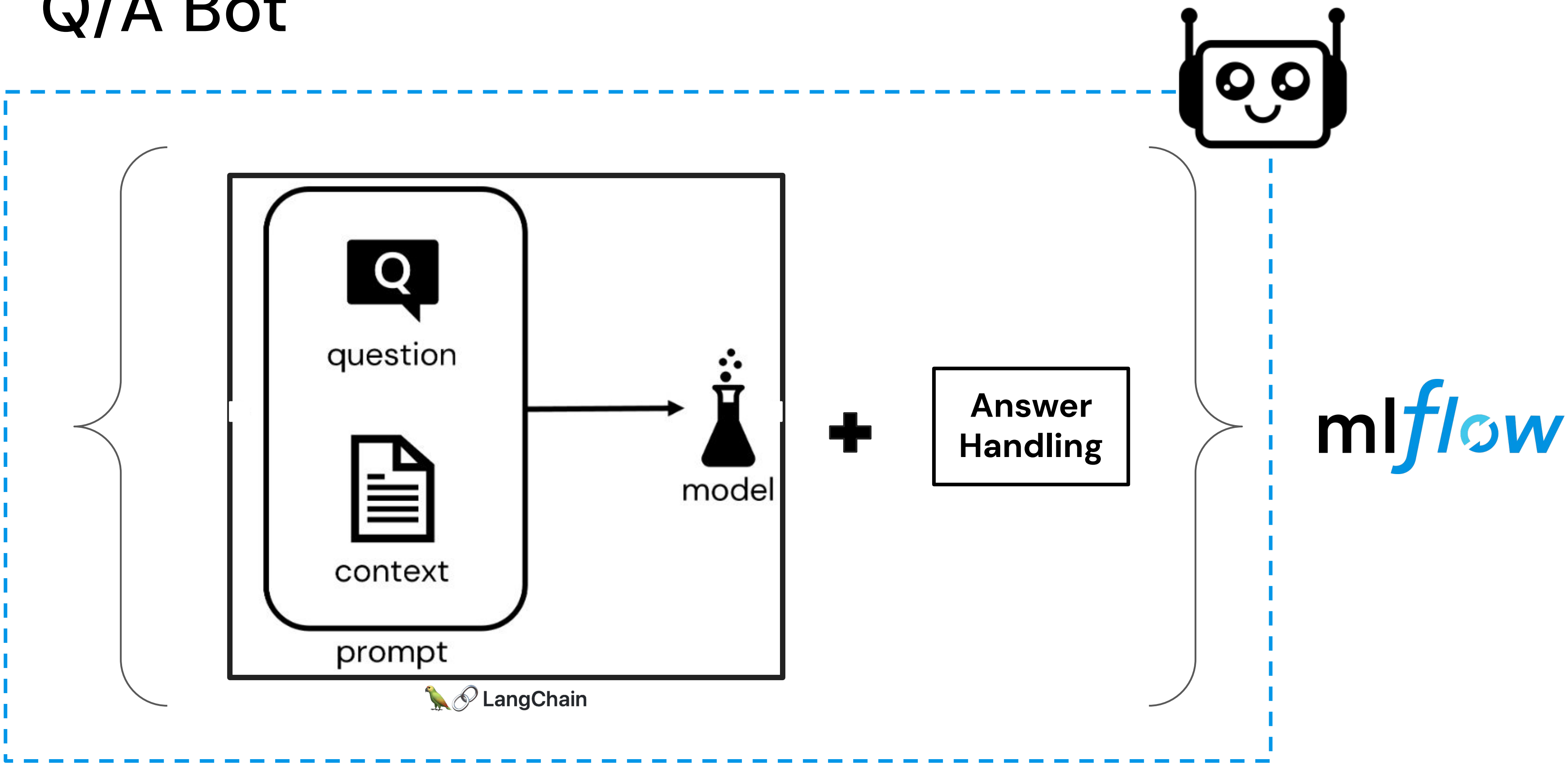
LangChain



Q/A Bot



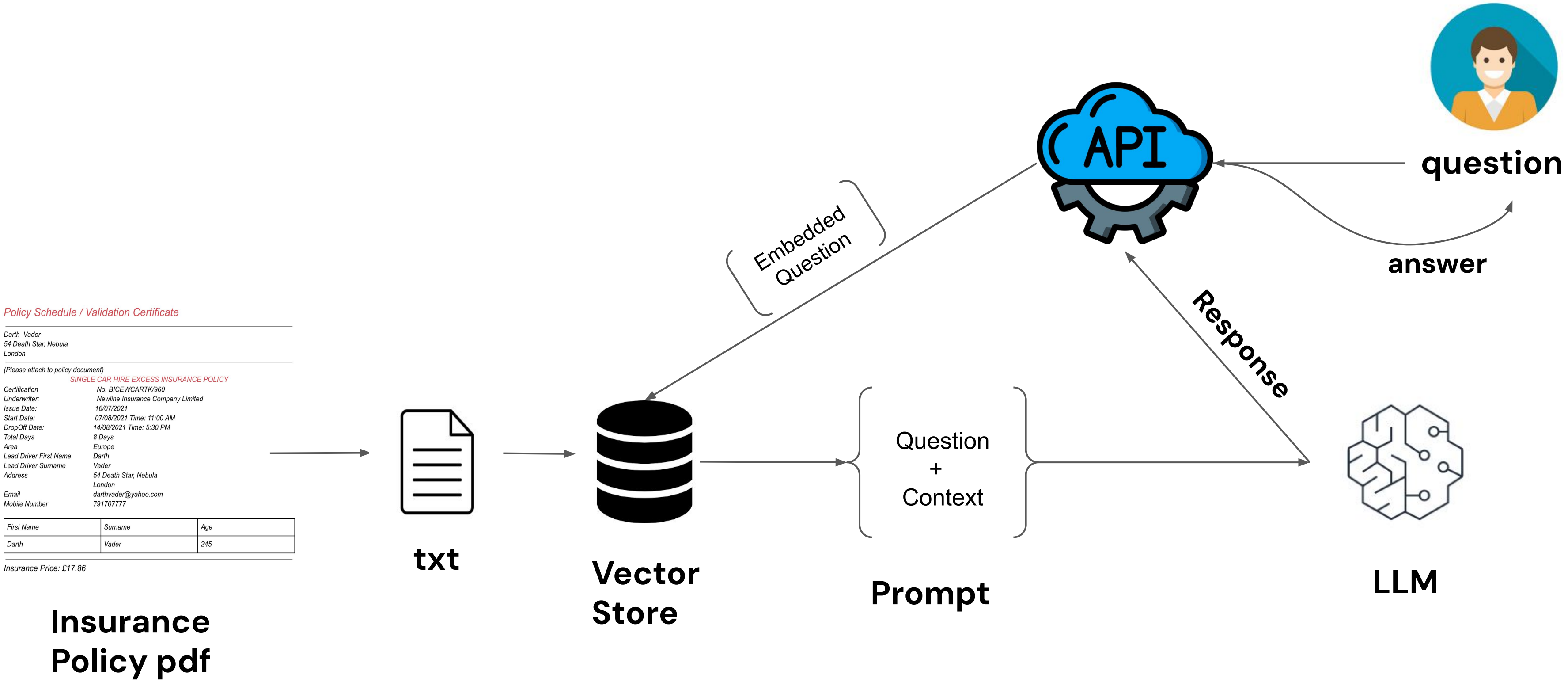
Q/A Bot


question
context
prompt
model
LangChain
Answer Handling
mlflow

Agenda

- New Wave of Deep Learning
- Customisation Phases of GenAI
- Build a RAG application
- **Live demo**
- Fine Tuning Concepts
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary - Call to Action

RAG Architecture



KPIs Time

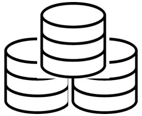
Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")

Training Cost



Data Size

Extra Hot

Know-how

High

Customisation

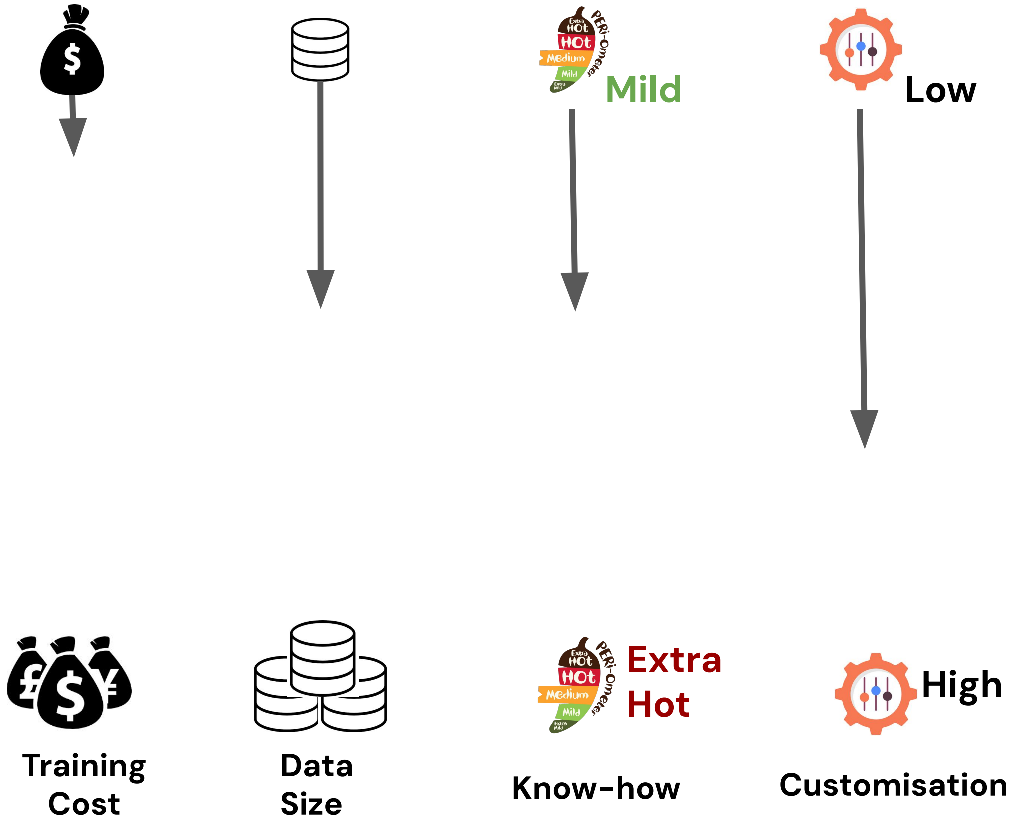
Customisation Phases of GenAi

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)



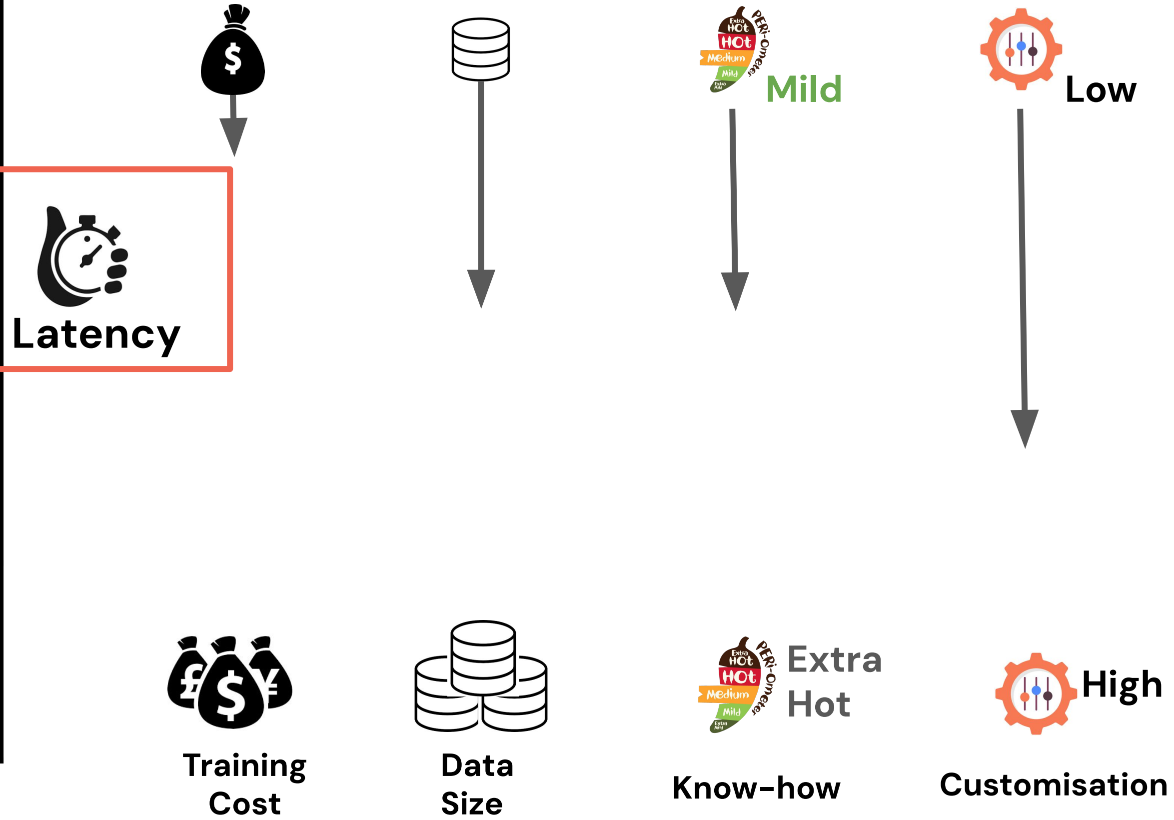
Customisation Phases of GenAi

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)



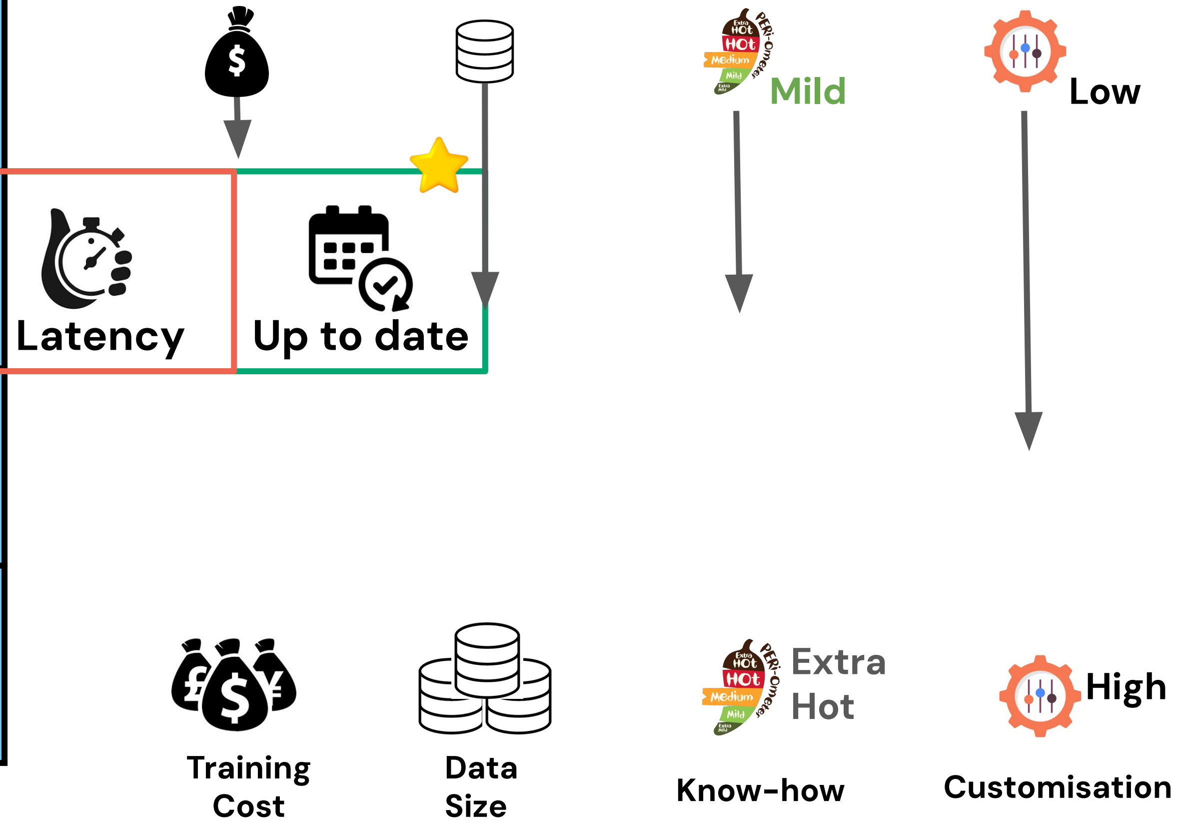
Customisation Phases of GenAi

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)



Phases of GenAi– RAG

Retrieval Augmented knowledge (RAG)	- **Augment knowledge** of a GanAI model with **your own data** - You can add filters to prompts **(avoid jailbreaking and hallucinations)** - **Can control the model and version** - Can **control ownership** - Still requires some **prompt engineering** - You **don't control data inside** the model knowledge base - It can **add latency** to your app

Final thoughts on RAG – Pros

- **Augment knowledge** of a GenAI model with your own data
- You can **add filters to prompts** (avoid jailbreaking and hallucinations)
- Can **control the model and version**
- Can **control ownership**

Final thoughts on RAG – Cons

- It can get **expensive**
- You **don't control data inside** the model knowledge base
- It is not 100% clear **how the prompt affects the answer**
- **Domain specific Q/A** may **not work** well with RAG

Agenda

- New Wave of Deep Learning
- Customisation Phases of GenAI
- Build a RAG application
- Live demo
- **Fine Tuning Concepts**
- 5 mins on Pretraining
- Do you really want to discuss cost?
- Summary - Call to Action

Customisation Phases of GenAi – Fine Tuning

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

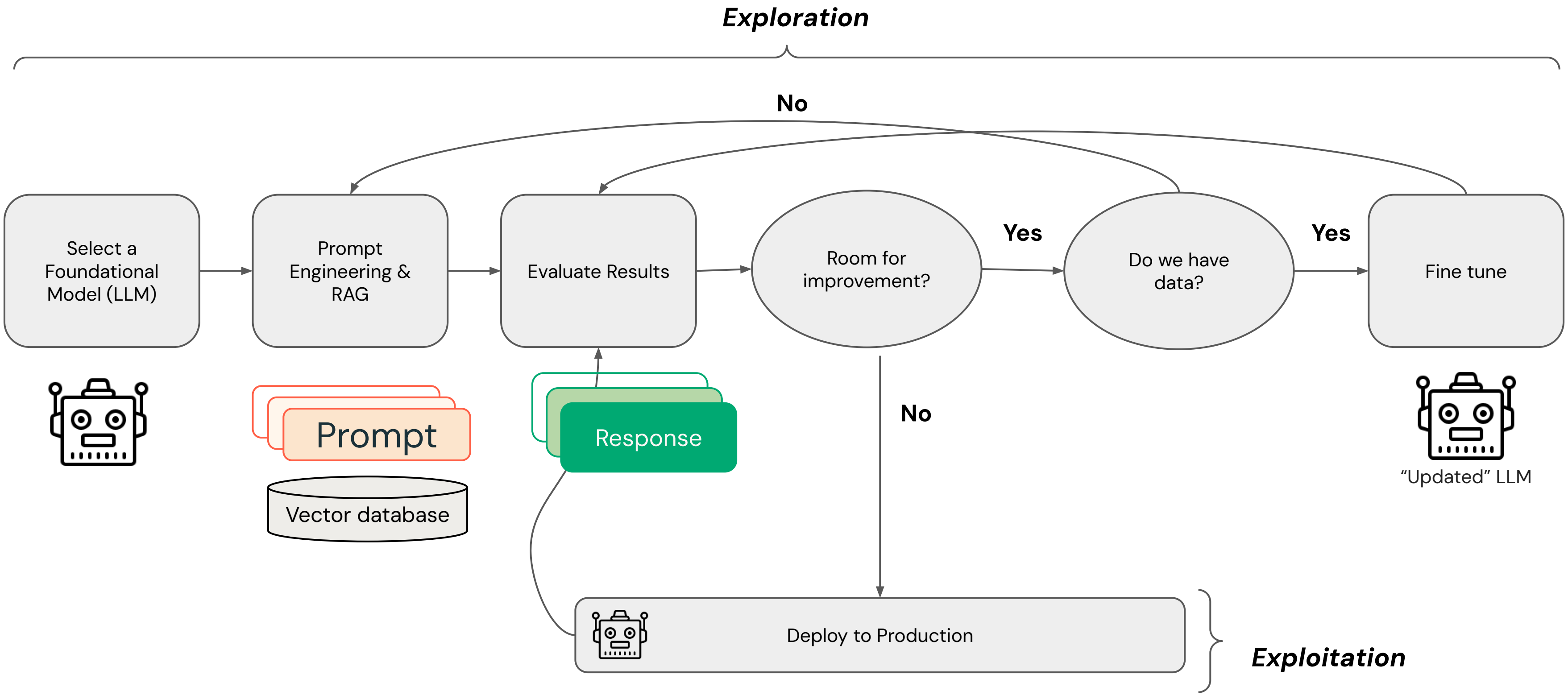
Fully retrain foundational models (your own “GPT”)

Tune a model on your data

When should I fine tune models?

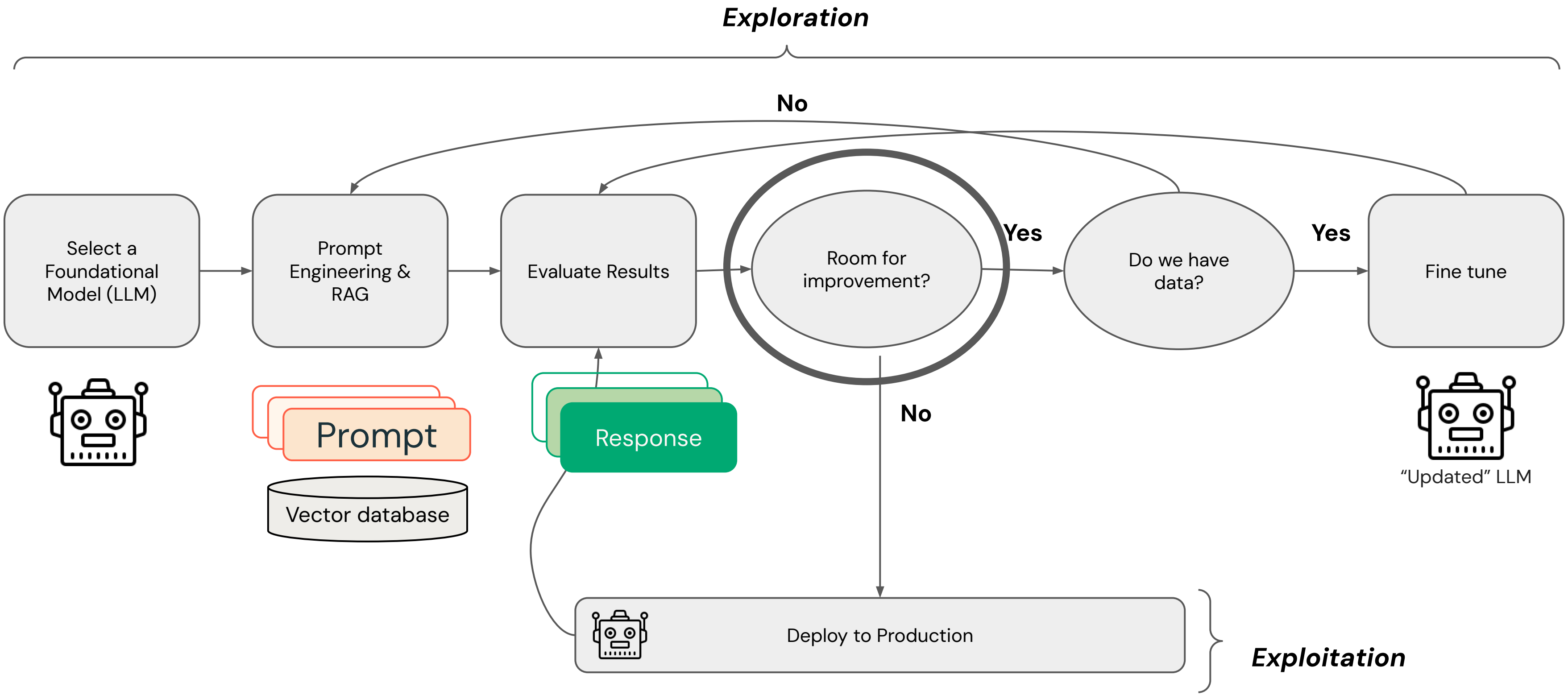
Initial Process

Experimentation & Exploitation Strategy

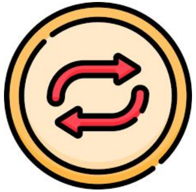


Initial Process

Experimentation & Exploitation Strategy

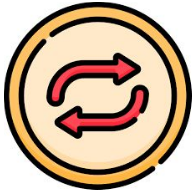


When should I fine tune?

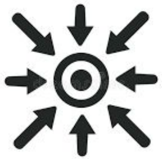


Repetition in the prompt – Token budget

When should I fine tune?

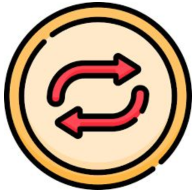


Repetition in the prompt – Token budget

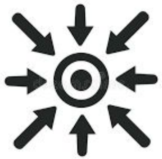


Promising few-shot

When should I fine tune?



Repetition in the prompt - Token budget



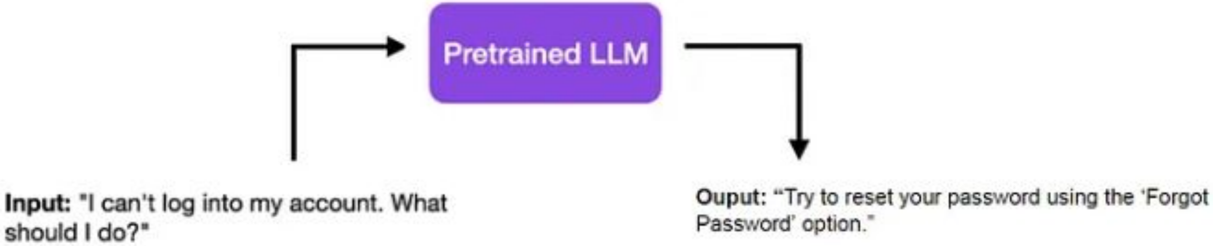
Promising few-shot

Change the Behaviour

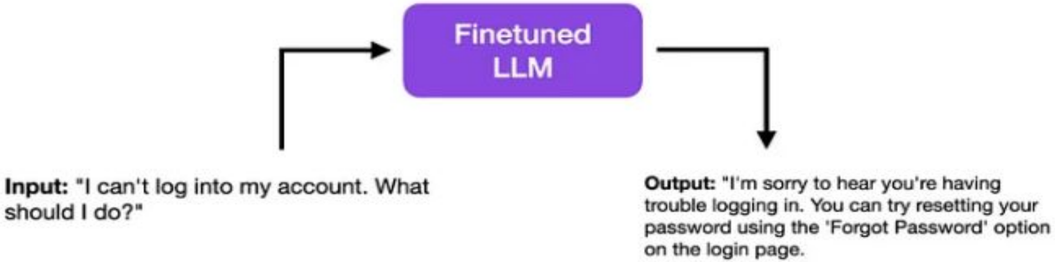
When should I fine tune?

It's NOT for new concepts

Fine Tuning - with an example



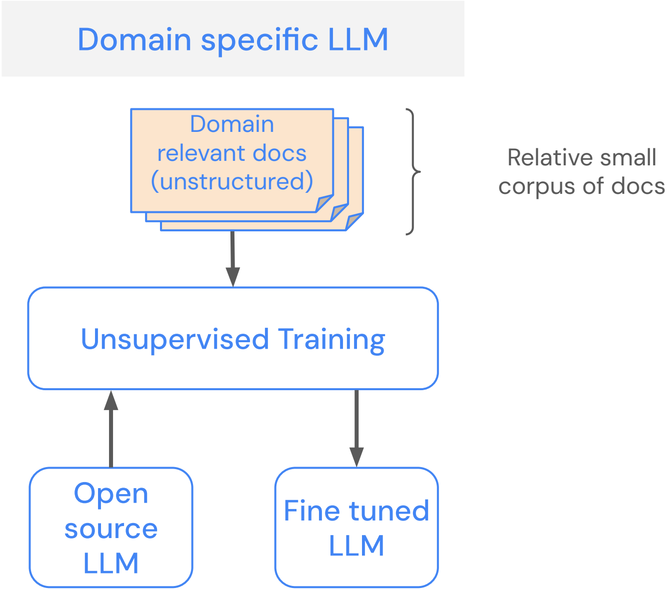
Fine Tuning - with an example



How are we fine tuning

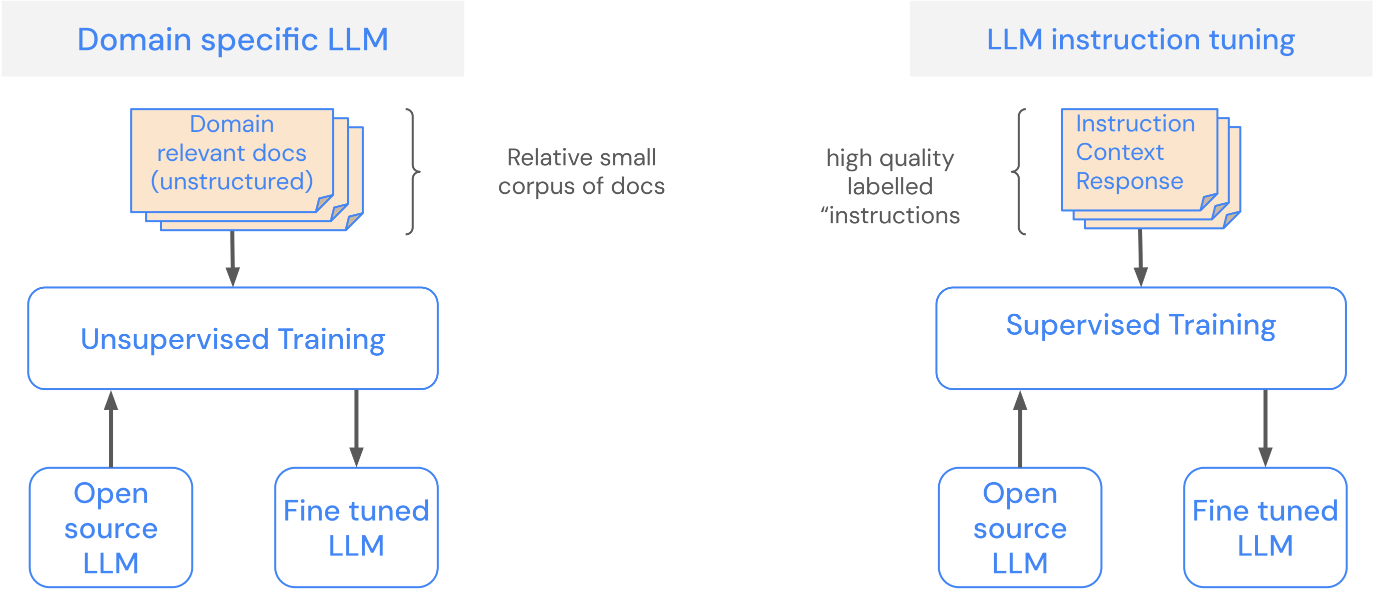
Fine-tuning Types - Domain Specific Tuning

"Adjust the model behavior"



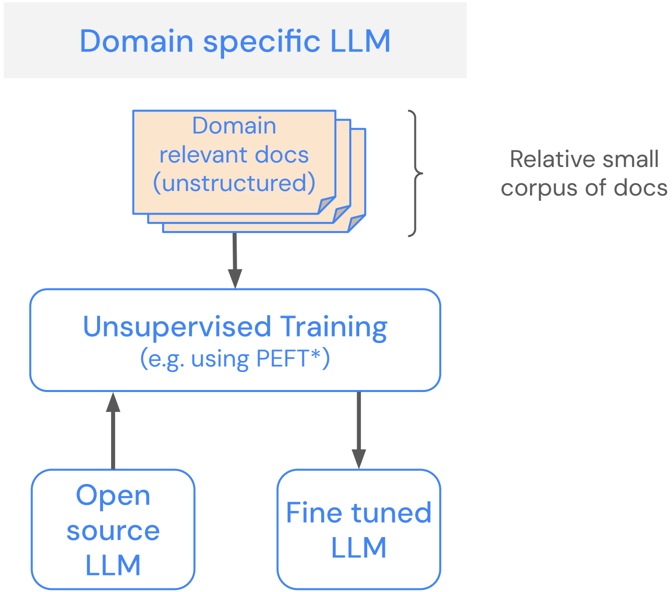
Fine-tuning Types - Instruction Tuning

“Adjust the model behavior”



Fine-tuning Types - Domain Specific Tuning

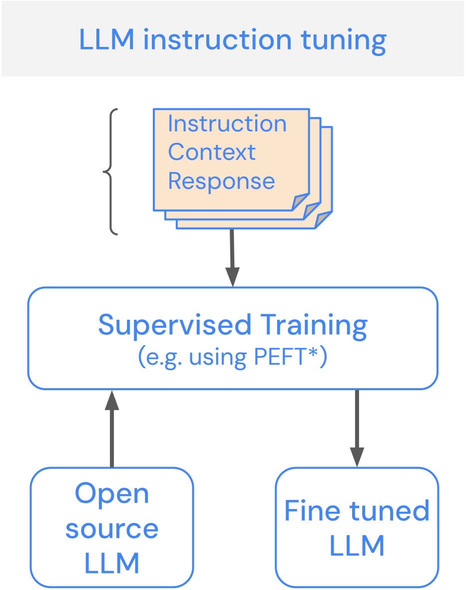
"Change the model behavior"



- Fine tune on small corpus

Fine-tuning Types - Instruction Tuning

“Adjust the model behavior”



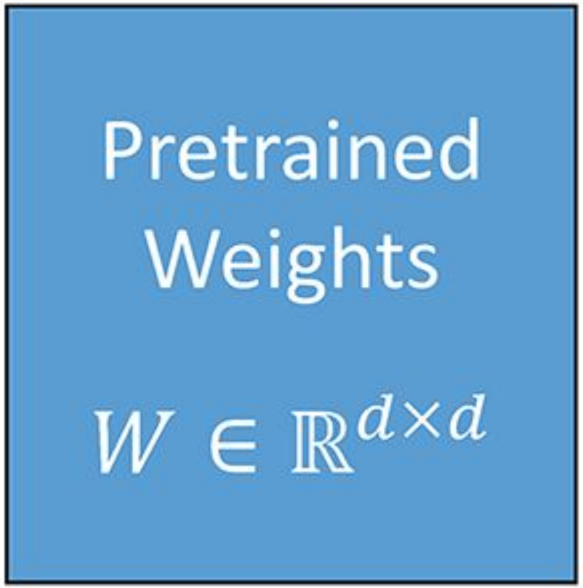
* Parameter-Efficient Fine-Tuning

Notes:

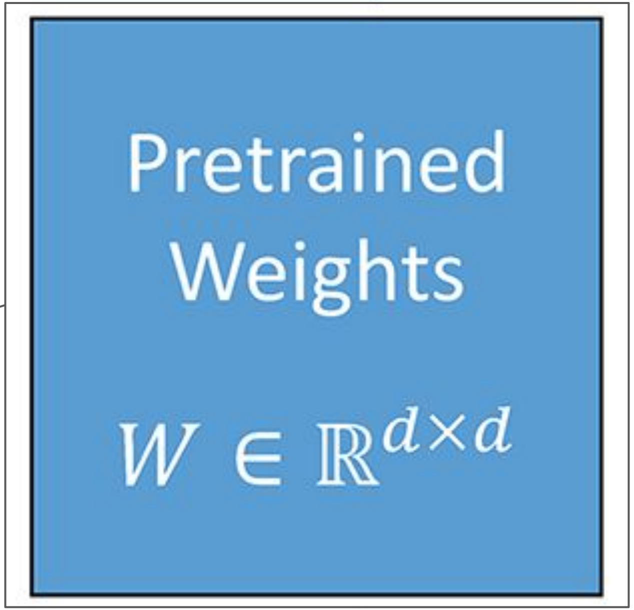
- **LLM instruction tuning** requires high quality labelled “instruction → response” data sets (increases effort & costs)
- Best results can be expected when combining both into two subsequent stages:

What am I fine tuning?

What am I fine tuning?

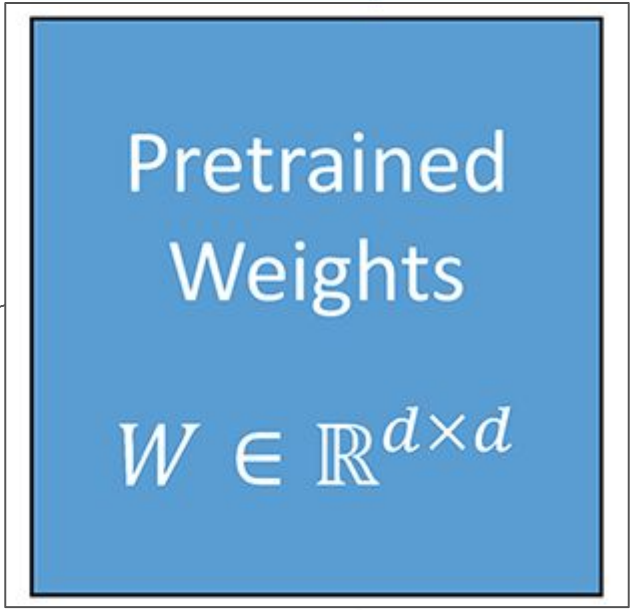


What am I fine tuning?



All of it?

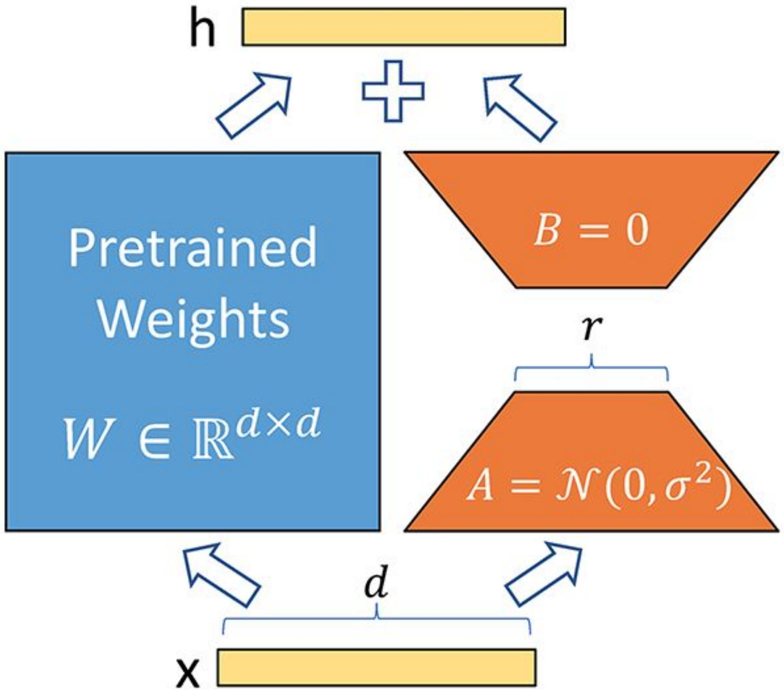
What am I fine tuning?



All of it?

1. Accelerate
2. Deepspeed

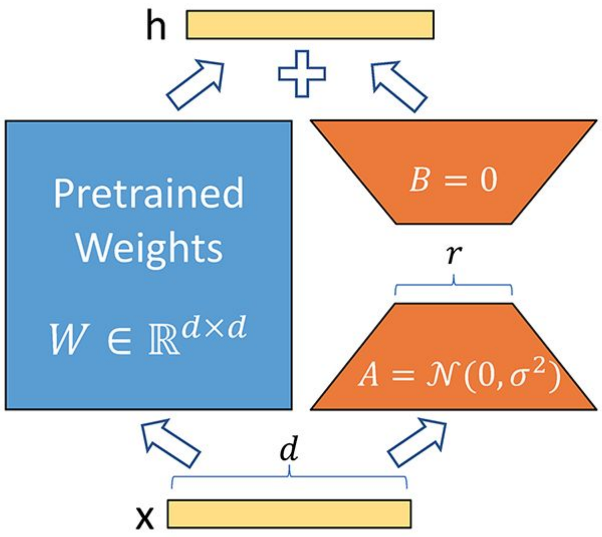
What am I fine tuning?



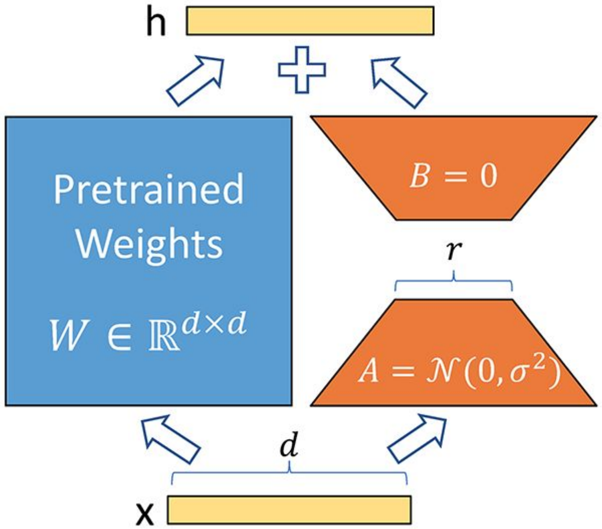
What am I fine tuning?

PEFT library → Parameter efficient Fine Tuning:

- Lora → Add Adapters with weights, which are the only parameters being fine-tuned, while freezing the rest
- Qlora → As above but quantized version

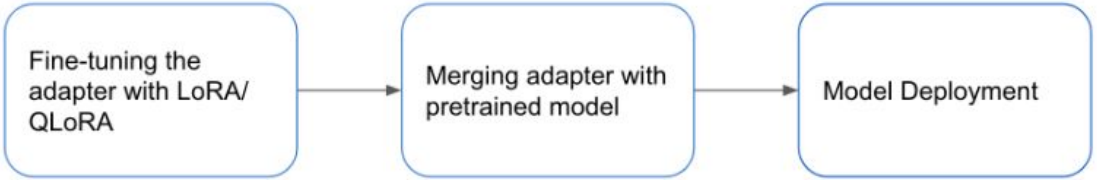


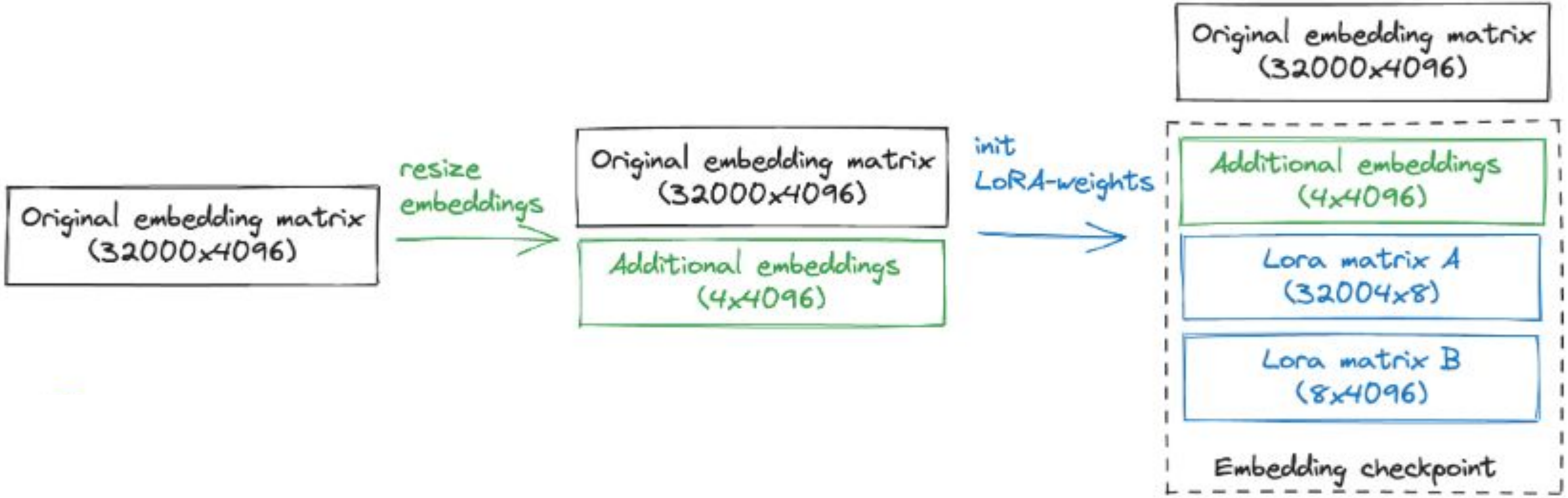
What am I fine tuning?



PEFT library → Parameter efficient Fine Tuning:

- Lora → Add Adapters with weights, which are the only parameters being fine-tuned, while freezing the rest
- Qlora → As above but quantized version




Original embedding matrix
(32000x4096)
resize
embeddings
Original embedding matrix
(32000x4096)
Additional embeddings
(4x4096)
init
LoRA-weights
Original embedding matrix
(32000x4096)
Additional embeddings
(4x4096)
Lora matrix A
(32004x8)
Lora matrix B
(8x4096)
Embedding checkpoint

Checkpoint Sizes

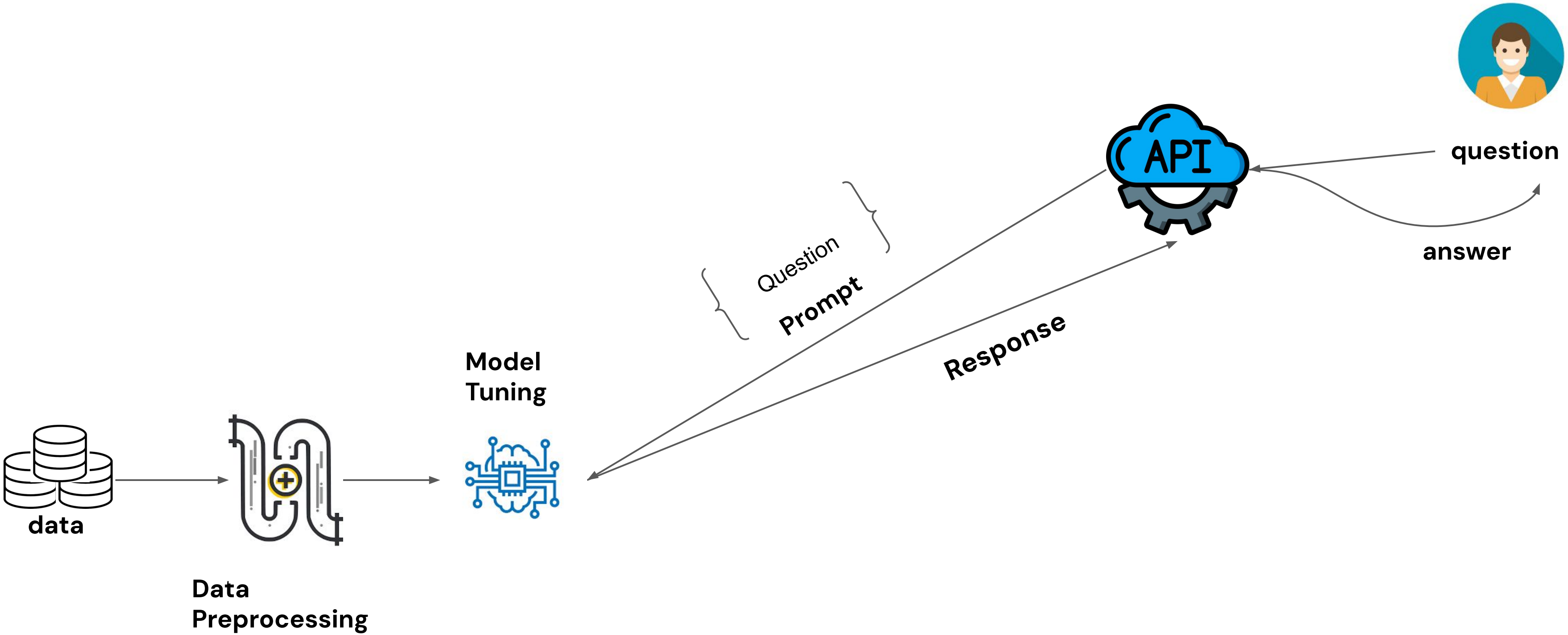
Ref

Number of trainable parameter / Checkpoint size	LoRA: q_proj and v_proj	LoRA: all layers	Full-parameter
7B	4194304 / 8MB	20566080 / 41MB	7B / 14GB
13B	6553600 / 13MB	31887424 / 64MB	13B / 26GB
70B	16384000 / 33MB	104190016 / 201MB	70B / 140GB

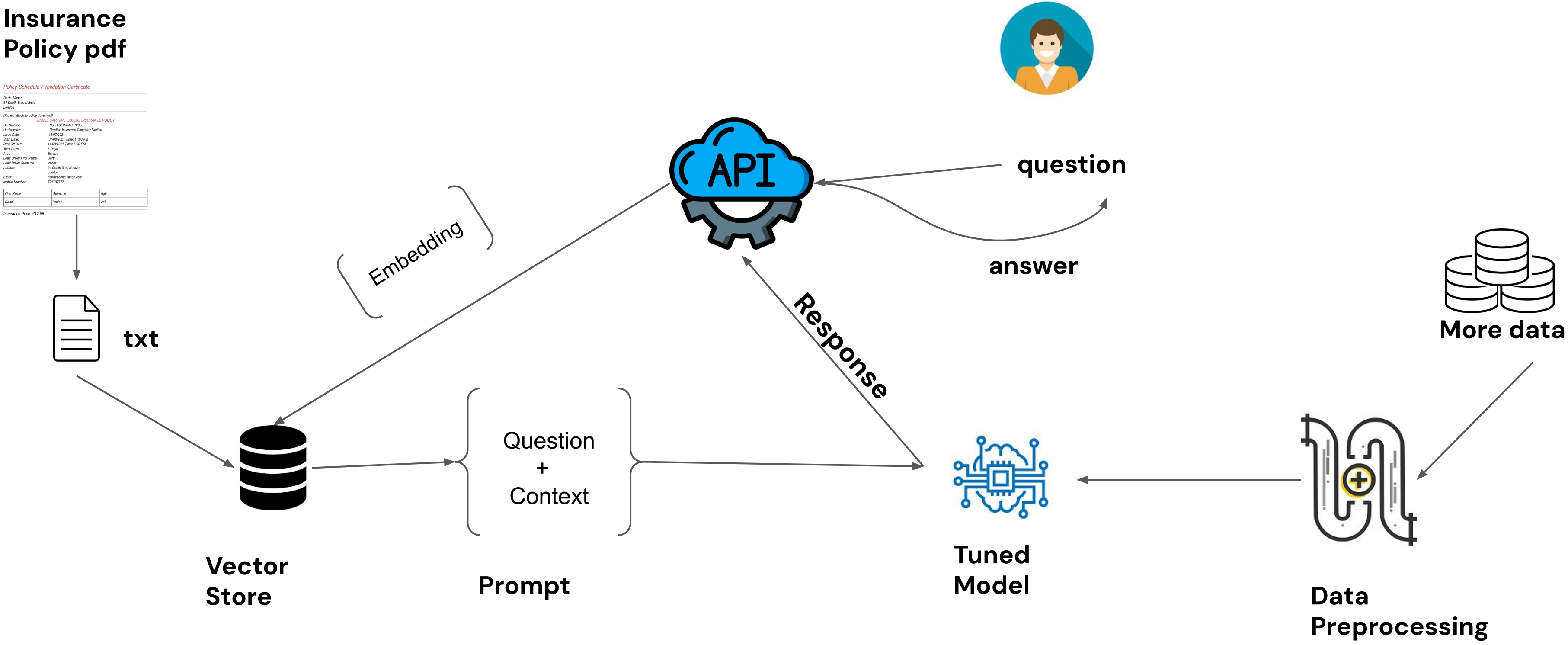
Final thoughts on Lora/Qlora

- The principal trade-off with LoRA is straightforward: you may give up some model quality, but you gain the ability to **serve many models more efficiently.**
- Cannot secure A100s? With LoRA you can still fine-tune models on **smaller GPUs** (reduced memory usage while training).
- Compared to regular checkpoints, LoRA checkpoints are significantly **smaller**, facilitating more **scalable serving, especially when managing multiple fine-tuned models.**

Fine Tuning Architecture



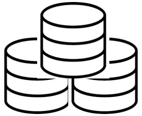
Fine Tuning Architecture With RAG



Kpis Time

Foundational model as SaaS with Prompts
Retrieval Augmented knowledge (RAG)
Fine-tune foundational model on your data
Fully retrain foundational models (your own "GPT")

Training Cost



Data Size

Extra Hot

Know-how

High

Customisation

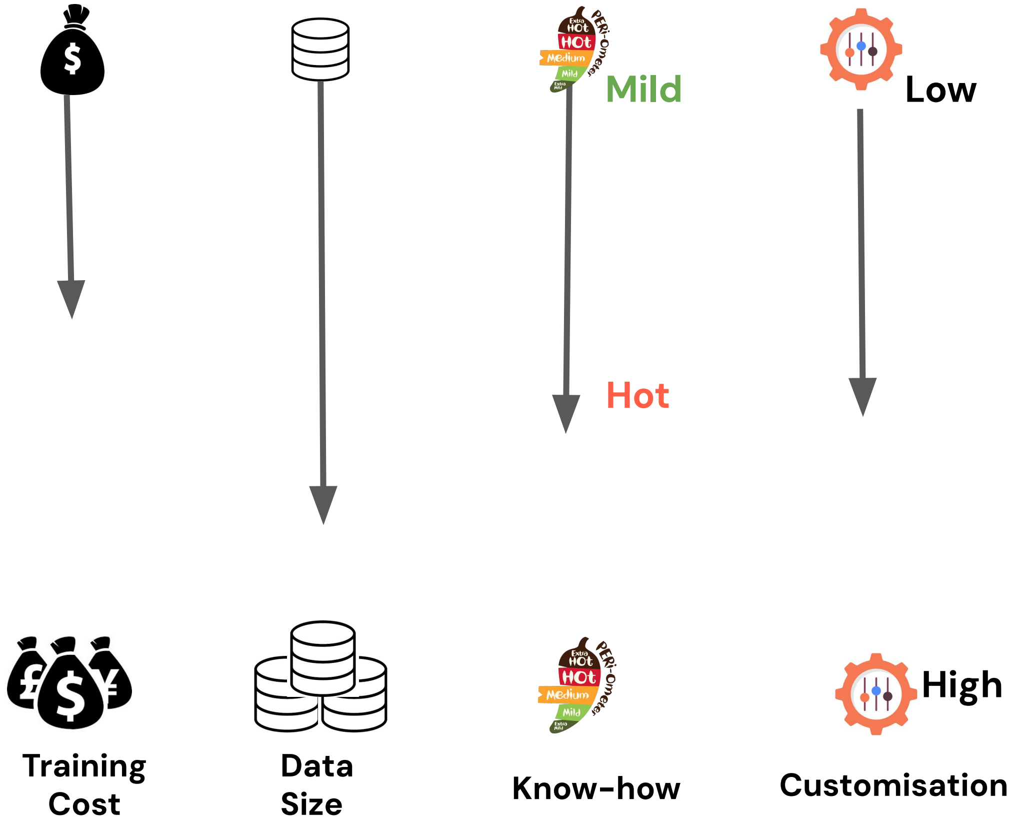
Phases of GenAi

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own "GPT")



Phases of GenAi- Fine Tuning

Fine-tune foundational model on your data	- Can **update certain "parts"** of the model - Can **win with a smaller model** - Can still **add RAG and** add filters to prompts **(avoid jailbreaking and hallucinations)** - **Can control the model and version** - Can **control ownership** - Still requires some **prompt engineering** - You **don't control data inside** the model knowledge base - **No guarantee this can improve quality** - Requires **computational resources and technical skills**

Final thoughts on Fine Tuning – Prons

- Trained on domain specific knowledge so more accurate responses (may, may not)
- You can **lock down the version of the model and IP**
- Can **still add RAG** and add filters to prompts (avoid jailbreaking and hallucinations)

Final thoughts on Fine Tuning – Cons

- You need to **gather the data** and make sure **they are of good quality**
- Fine-tuning typically results in creating a **niche model for a niche use-case**
- Model management and infrastructure, and serving
- Original pretraining data **may dominate**
- Can't create new capability, just bring domain specific knowledge to the model
- No **guarantee** this can improve quality

Agenda

- New Wave of Deep Learning
- Phases of GenAI
- Build a quick RAG application with your own data
- Live demo
- Fine Tuning Concepts
- Live demo
- **5 mins on Pretraining**
- Do you really want to discuss cost?
- Summary – Call to Action

Customisation Phases of GenAi – Pretraining

Foundational model as SaaS with Prompts

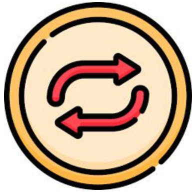
Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

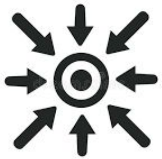
Fully retrain foundational models (your own “GPT”)

Create a model on your data

Why Pretraining:



Full customisation



Full IP

Competitive Advantage

Why pretrain models?

Control

To Train Generative AI models we need…..

GPUs

We need a lot of GPUS to train <u>your own</u> Generative AI models

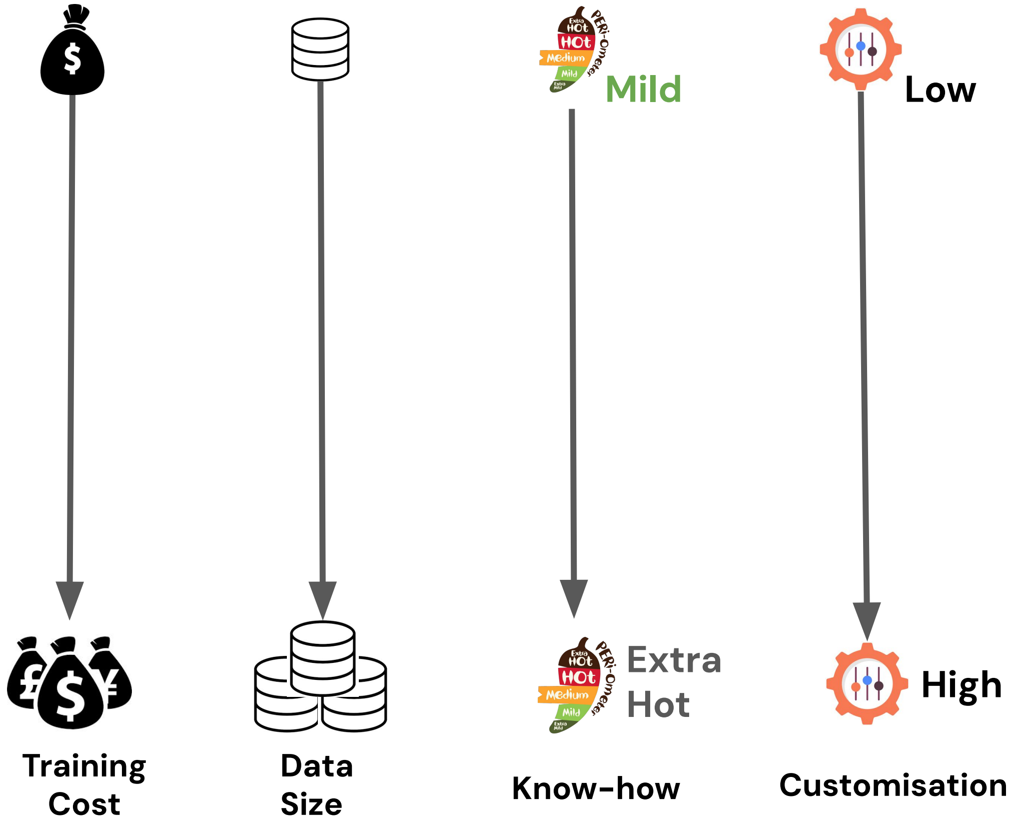
Kpis

Foundational model as SaaS with Prompts

Retrieval Augmented knowledge (RAG)

Fine-tune foundational model on your data

Fully retrain foundational models (your own “GPT”)



Phases of GenAi– RAG

Retrieval Augmented knowledge (RAG)	- **Owner** of your mini GPT style - The way to go for a very **particular use cases or small models** - Requires **resources** both technical and computational - Requires a **lot of data** or labels (100K+) - If you don't know what you are doing, it will **never converge**

Final thoughts on Fine Tuning – Prons

- Owner of your **mini GPT style**
- You **control data** inside the model knowledge base

Final thoughts on Fine Tuning – Cons

- Requires **resources both technical and computational**
- Requires **a lot of data**
- If you don't know what you are doing, **it will never converge**

Agenda

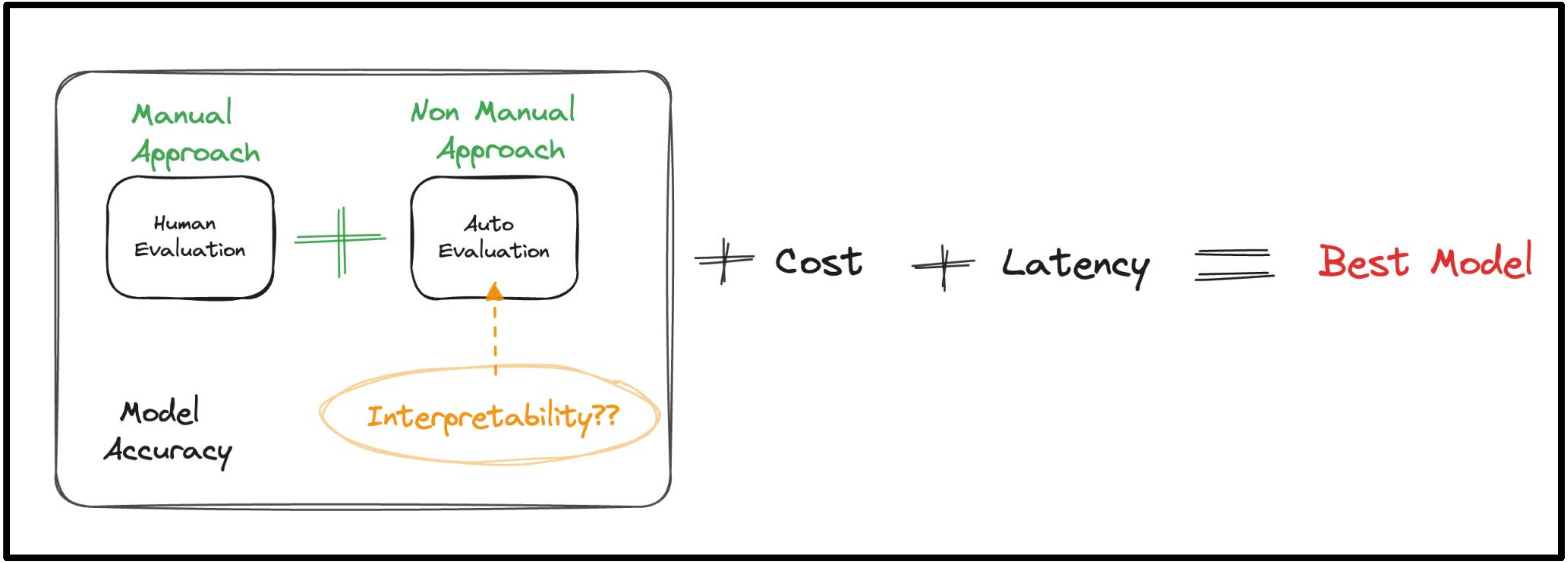
- New Wave of Deep Learning
- Phases of GenAI
- Build a quick RAG application with your own data
- Live demo
- Fine Tuning Concepts
- Live demo
- 5 mins on Pretraining
- **Do you really want to discuss cost?**
- Summary – Call to Action

LLM Projects: Cost

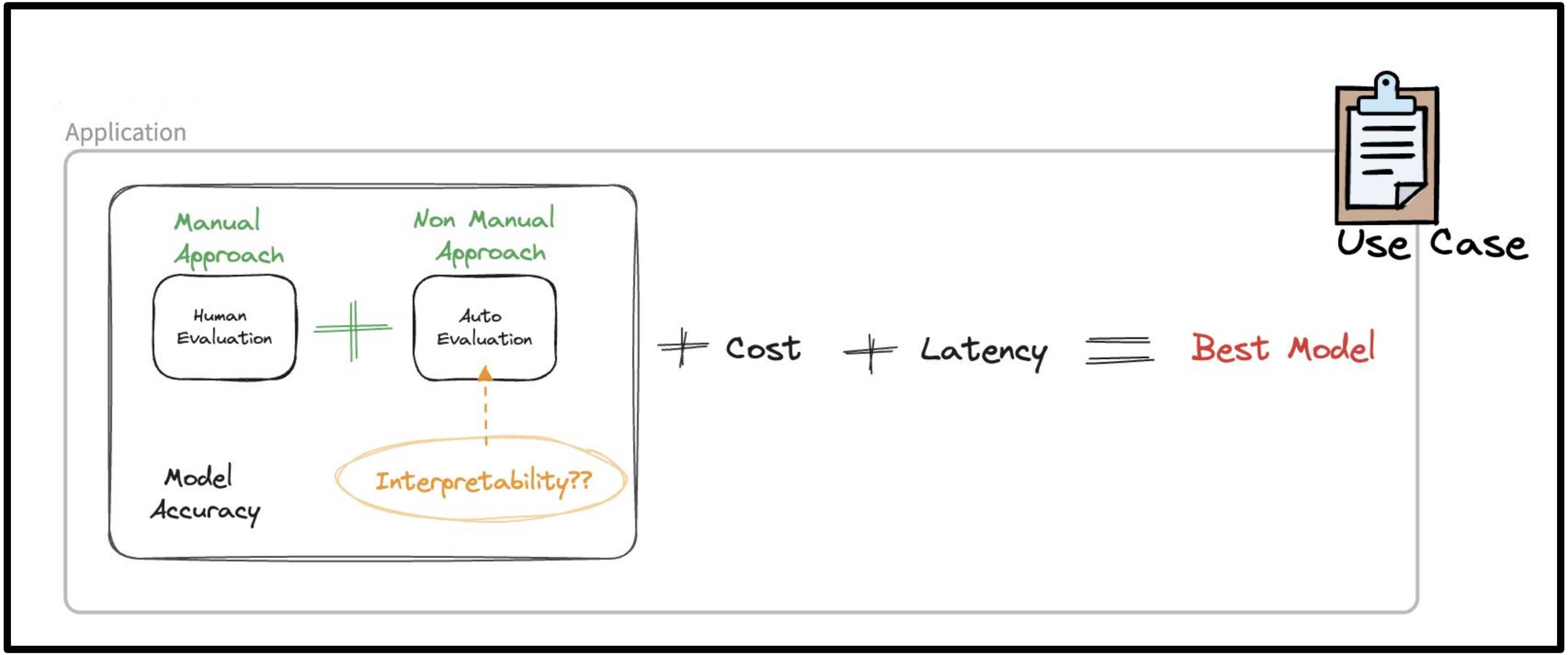
- **Instruction Fine Tune:**
  - Start from 1-10k training examples
  - PEFT
- **Continue Pre-Training**
  - Starts from around 100m-1bn tokens
  - PEFT
- **Pre-Train LLM from scratch:**
  - Requires carefully crafted and very huge(1T) training datasets

Complexity	Cost	Training Loop
Distributed GPU setup required High	Depending on model size Medium to High	Occasional
Distributed GPU setup required High	Depending on model size Medium to High	Occasional
Distributed GPU setup required High	100K - 2.5 Mil $ Very high	Rare

Your Best LLM

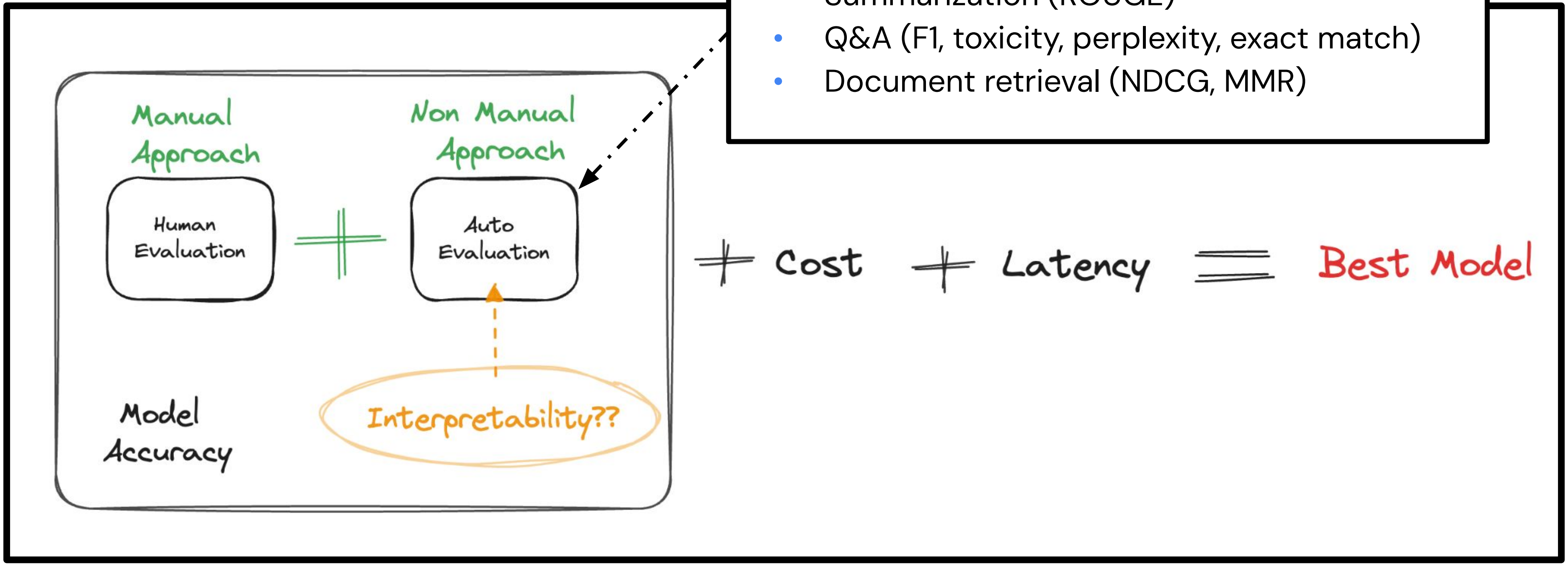


Your Best LLM



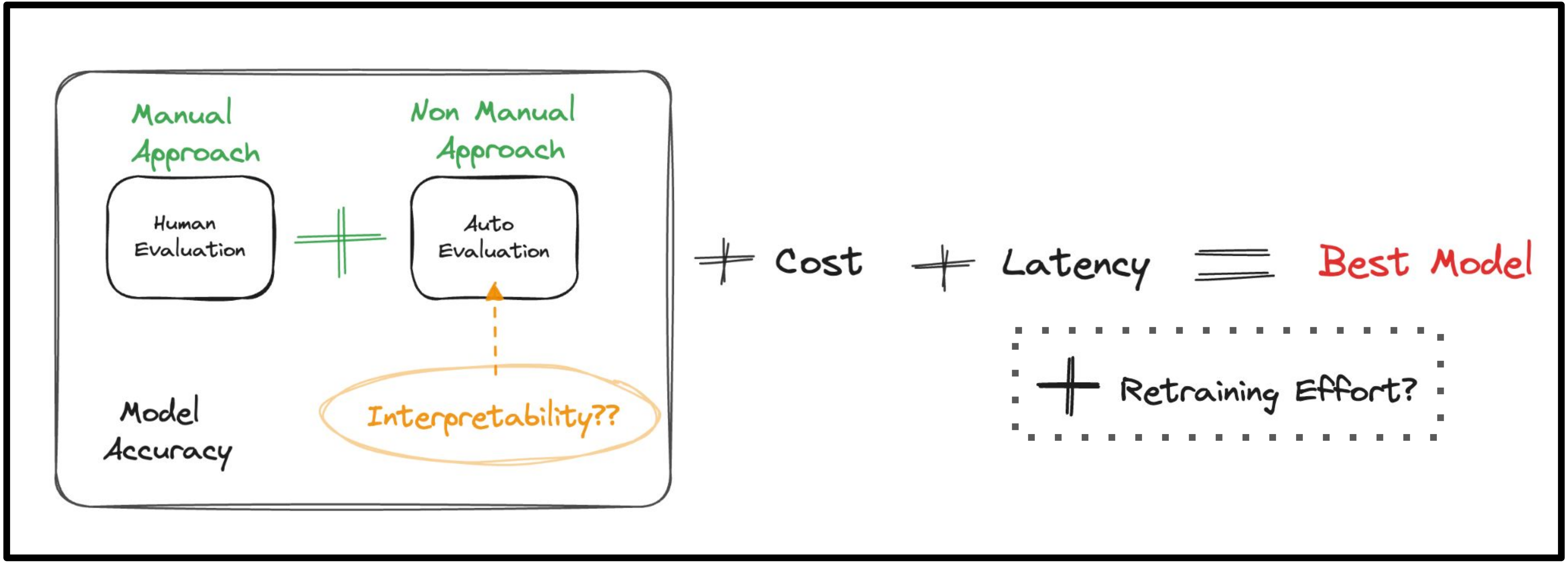
Evaluate LLMs – Auto Evaluation

- Translation (BLEU),
- Summarization (ROUGE)
- Q&A (F1, toxicity, perplexity, exact match)
- Document retrieval (NDCG, MMR)



Examples here

Your Best LLM



Agenda

- New Wave of Deep Learning
- Customisation Phases of GenAI
- Build a quick RAG application with your own data
- Live demo
- Fine Tuning Concepts
- Live demo
- 5 mins on Pretraining
- Do you really want to discuss cost?
- **Summary – Call to Action**

Summary

- New wave of LLMs
- Customisation Phases of LLMs
  - Prompt Engineering
  - RAG
  - Fine Tuning
  - Pretraining
- Aspects of cost

An Overview of Common LLMs

Use case	Quality-optimized	Balanced	Speed-optimized	Notes
Text generation following instructions	Mixtral-8x7B-Instruct-v0.1 MPT-30B-Instruct † Llama-2-70b-chat-hf	Mistral-7B-Instruct-v0.2 MPT-7B-Instruct † MPT-7b-8k-instruct Llama-2-7b-chat-hf Llama-2-13b-chat-hf	phi-2	† Supervised fine-tuning using databricks-dolly-15k dataset
Text embeddings (English only)	e5-mistral-7b-instruct (7B)	Bge-large-en-v1.5 (0.3B) e5-large-v2 (0.3B)	bge-base-en-v1.5 (0.4B) bge-small-en-v1.5 (0.1B) e5-base-v2 (0.1B)	
Transcription (speech to text)		whisper-large-v3 (1.6B)	distil-large-v2 (0.7B)	
Image generation		stable-diffusion-xl		
Code generation	CodeLlama-70b-hf CodeLlama-70b-Instruct-hf CodeLlama-70b-Python-hf (Python optimized) CodeLlama-34b-hf CodeLlama-34b-Instruct-hf CodeLlama-34b-Python-hf (Python optimized)	CodeLlama-13b-hf CodeLlama-13b-Instruct-hf CodeLlama-13b-Python-hf (Python optimized) CodeLlama-7b-hf CodeLlama-7b-Instruct-hf CodeLlama-7b-Python-hf (Python optimized)		Code LLMs usually need fine-tuning to follow instructions and work on application-specific code

Thank you so much!

Maria Zervou

Sr. Specialist Solutions Architect

Code & Slides