to Fine-tune *or* Prompt *or* RAG

Using LLMs in business

Sandeep Arora

www.trilyen.com

LLMs in Business

LLMs are trained on internet scale data & have knowledge of the world. But these LLMs do not understand your business or domain data.

Using LLMs to solve your own business problems can be challenging. Take for example a Law Firm which wants to create a Knowledge extraction chatbot using its knowledge and data. This can be hard despite available tools and frameworks. There is no one-stop shop for this. Here is a framework on how to adapt LLMs to solve business problems you care about.

Law Firm

International Law firm *(multiple countries, jurisdictions, business domains)* wants to use LLMs to prepare draft case documents, compare previous cases, new employee Q&A

LLM

Foundation Model

- **Understand language**
- **Have knowledge of the world**
- **Have reasoning capability**

- **Lack law & legal domain knowledge**

Framework to extend model knowledge and adapt to your task

LLM

How to unlock the full potential of LLMs to solve business problems

Track 1

Use model as-is.
Model is not modified.

Prompt Engineering

Few-shot
Zero-Shot
Chain-of-Thought

Used for simple use cases like extractive Q+A or rapid prototyping

Framework to extend model knowledge and adapt to your task

LLM
How to unlock the full potential of LLMs to solve business problems

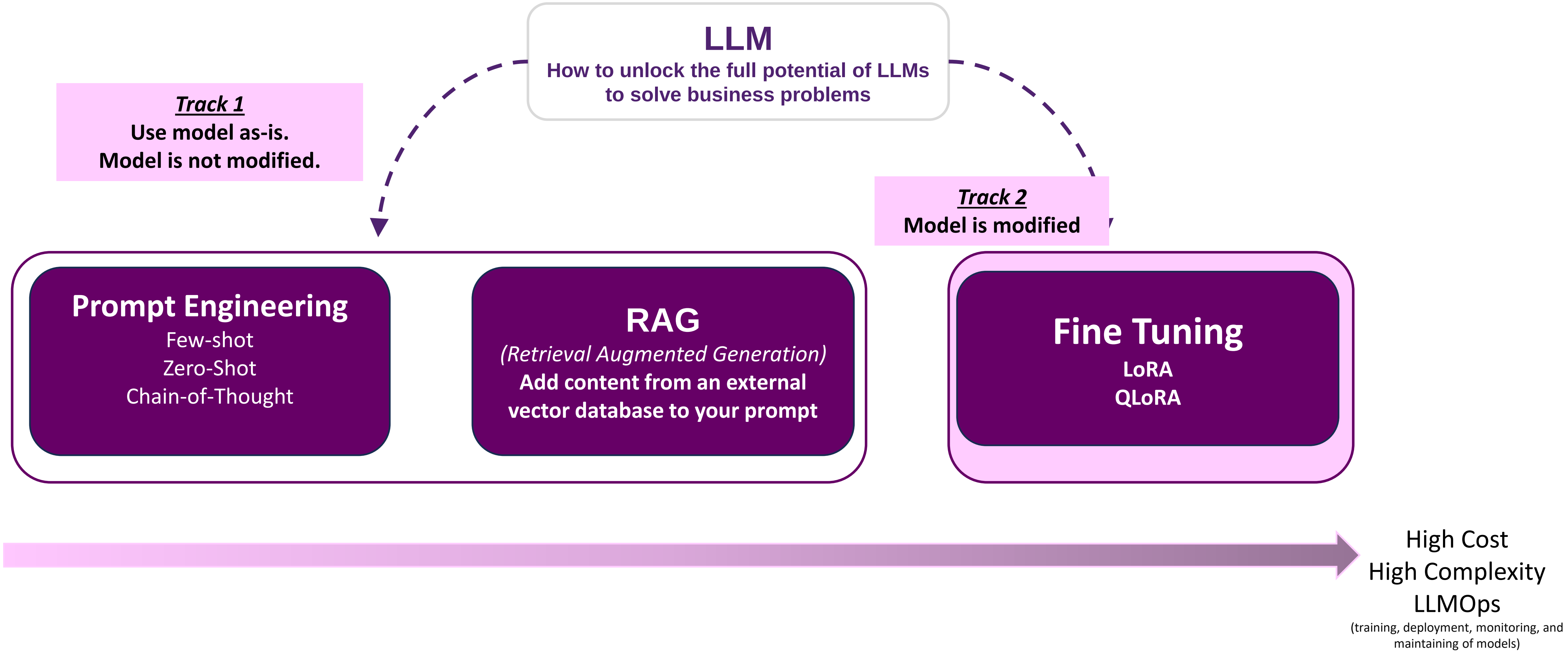
<u>Track 1</u>
Use model as-is.
Model is not modified.

Prompt Engineering
Few-shot
Zero-Shot
Chain-of-Thought

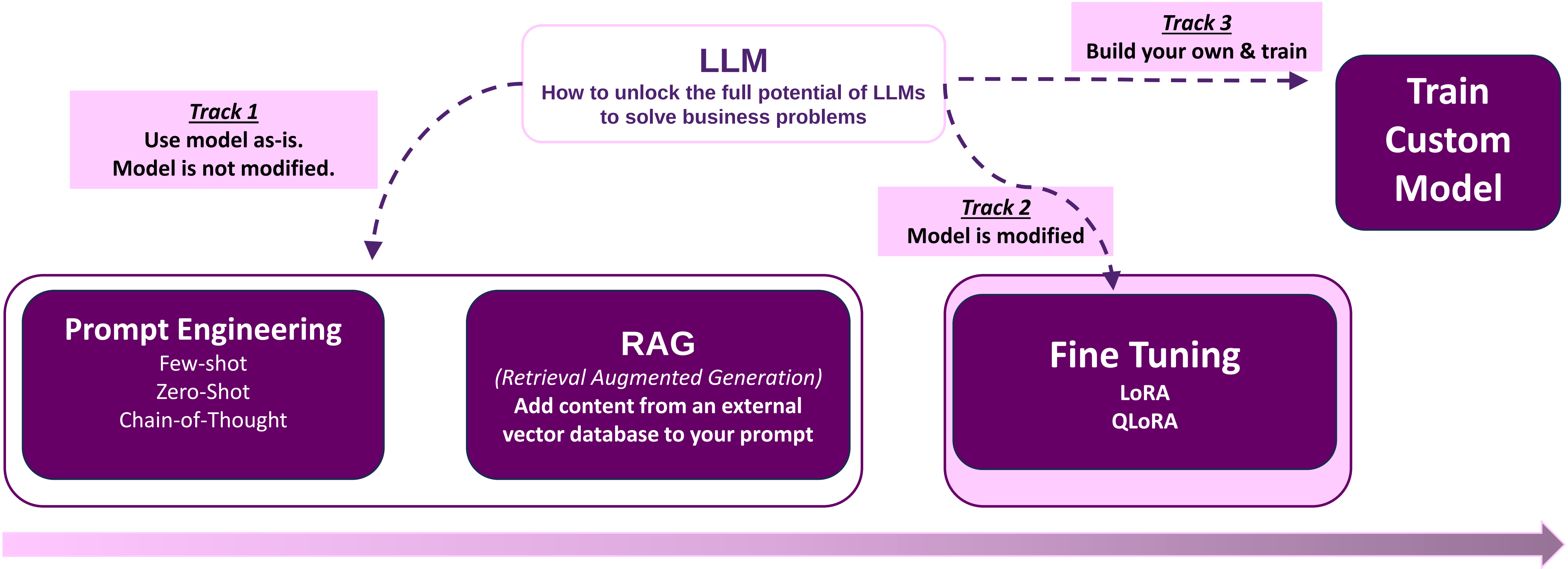
RAG
(Retrieval Augmented Generation)
Add content from an external vector database to your prompt

High Cost
High Complexity

Framework to extend model knowledge and adapt to your task



Framework to extend model knowledge and adapt to your task



High Cost
High Complexity
LLMOps needed
(training, deployment, monitoring, and maintaining of models)

By Sandeep Arora
www.trilyen.com

(Track 1) Prompt Engineering

Prompt engineering is the process of carefully designing and refining the input (the "prompt") given to an AI model to achieve the desired output.

What is a Prompt? A prompt is a combination of requirements, instructions, guardrails combined with data to be acted upon by the LLM.

Anatomy of a Prompt

Style & Tone	Guardrails
Requirement	Instructions
Output Format	Examples

Types of Prompts

Zero-Shot

The model predicts the answer given only a natural language prompt with no examples

<u>Use cases</u>

Language Translation
Content moderation
Sentiment analysis

One-Shot

Is a technique where a model leverages a single example to understand and execute a task

<u>Use cases</u>

Automated content generation
Document summarization
Email drafting

Few-Shot

Add few examples to the prompt. The model will adapt results based with in-context learning

<u>Use cases</u>

Customer support automation
Content generation
Data cleaning
JSON extraction

CoT(Chain-of-Thought)

Guides a model to solve complex problems by explicitly prompting it to follow a step-by-step reasoning process.

<u>Use Cases</u>

Mathematical problem solving
Data analysis

Few-shot prompt

SA **You**

You are a bank customer service bot.

Style & Tone

Your task is to assess customer intent and categorize customer inquiry after into one of the following predefined categories:

Requirement

change pin
lost card
exchange rate
country support
charge dispute

If the customer inquiry doesn't fit into any of the above categories, classify it as: customer service.

Instructions

You will only respond with the predefined category. Do not provide explanations or notes.

Output Format

Here are some examples:

Inquiry: When will my credit card arrive? I'm going on a vacation and will not be able to check my mails for few days in first week of May

Examples

Category: card arrival
Inquiry: I am planning an international trip to Mexico. Does the card work there ?Category: Country support Inquiry: I forget pin of my credit card. Can you help me reset it ?
Category: Change pin
For the below question coming from a customer what is the service category "My wallet got stolen. It was in my bag at a conference. It had my credit card. What should i do?"

‹ 8 / 8 ›

ChatGPT

lost card

Output Format

Chain-of-thought prompt

SA **You**

Karen went to the grocery store. She saw mangoes were on Sale and she bought a dozen. She already had 5 mangoes at home. Later that evening her father came to her home and gave her half-a-dozen mangoes. How many mangoes does she have in total. Don't show any code. Lets work this step-by-step

‹ 2 / 2 ›

ChatGPT

Sure, let's calculate how many mangoes Karen has now, step by step.

1. Karen initially had 5 mangoes.
2. She bought a dozen mangoes. A dozen equals 12.
3. Her father gave her half a dozen mangoes later, which equals 6.

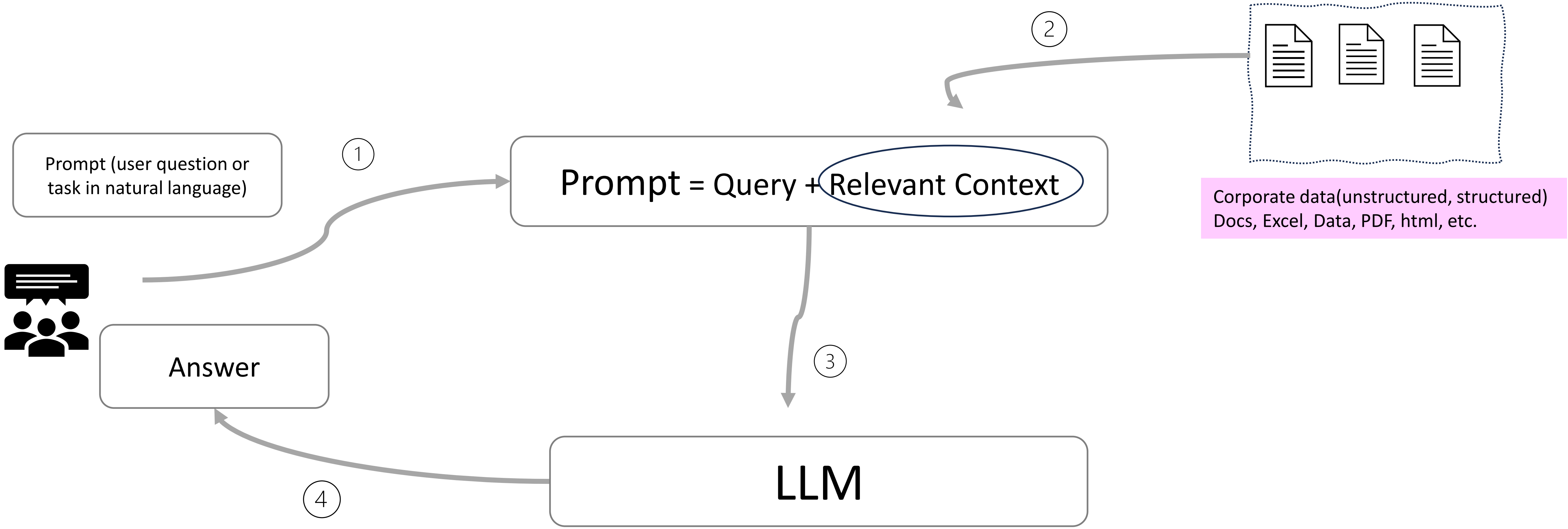
Now, let's add all these together:

- Start with the 5 mangoes she already had.
- Add the 12 mangoes she bought.
- Add the 6 mangoes her father gave her.

That makes a total of 23 mangoes.

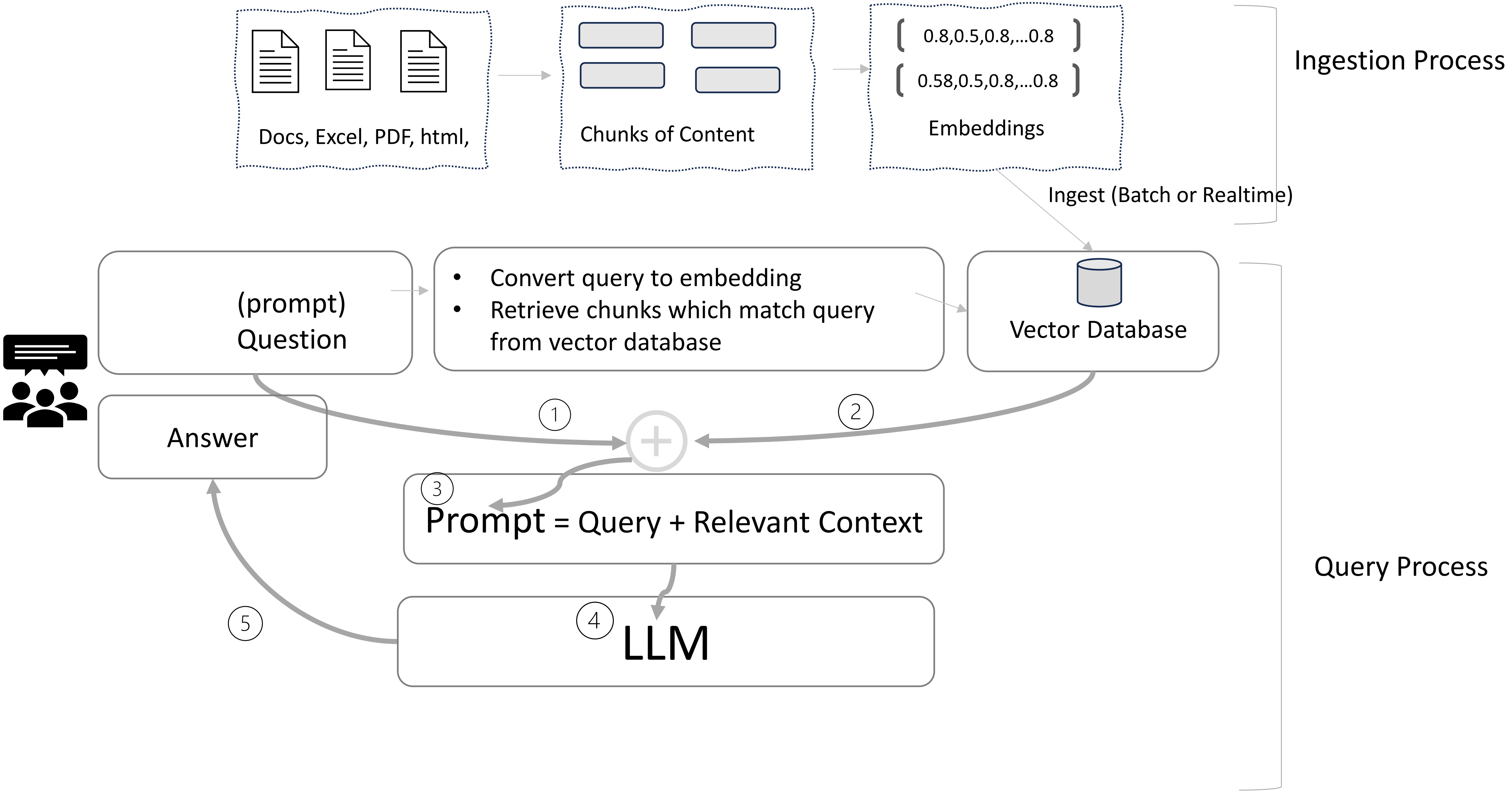
RAG (Retrieval Augmented Generation)

In simple words, RAG is a technique where you combine external content like (business knowledge/data) with a prompt before sending it to the LLM. The LLM then acts & reasons on that to provide you with results. Essentially you are using the AI power of a LLM on your custom knowledge/data.



Watch my video on RAG for more details and a demo

RAG Application Architecture



Framework to extend model knowledge and adapt to your task

LLM

How to unlock the full potential of LLMs to solve business problems

Track 1
Use model as-is.
Model is not modified.

Track 2
Model is modified

Track 3
Build your own & train

Prompt Engineering
Few-shot
Zero-Shot
Chain-of-Thought

RAG
(Retrieval Augmented Generation)
Add content from an external vector database to your prompt

Fine Tuning
LoRA
QLoRA

Custom Model
Train

High Cost
High Complexity
LLMOps
(training, deployment, monitoring, a
maintaining of models)

(track 2): Fine Tuning

Fine tuning is a method to train a foundation model with your own business data or domain knowledge. During the training process, the weights are adjusted. After fine-tuning, you can use the custom model it to build applications that are specific to your organization and use cases.

For example, **Boeing** wants to build an AI Assistant which helps train new employees on Aeronautical Engineering.

LLM
Foundation Model

*Has **NO** knowledge about Aeronautical Engineering*

Fine Tuning
(PEFT, LoRA, QLoRA)

(Aircraft Engineering)
Domain specific data

LLM
Custom Model

Now *Has knowledge about Aircraft Engineering*

This is a new model. Weights have been adjusted. Can be used to build domain specific AI applications.

Aeronautical engineering is a field of engineering that focuses on designing, developing, testing and producing aircraft.

FineTuning

Challenges of Fine-Tuning

Fine-tuning a large foundation model changes the model's weights and can be computationally intensive, expensive and time consuming, making it out of reach for many business.

PEFT (parameter efficient fine-tuning)

Addresses the challenges of full fine-tuning

Parameter Efficient Fine-Tuning (PEFT) refers to a range of techniques used to adapt large pre-trained models to specific tasks with **minimal updates** to the model parameters. The goal is to retain the benefits of the pre-trained model while customizing it for a particular dataset or task without the need for extensive re-training, which can be costly in terms of computational resources and time. Here we will talk about two such techniques:

1. LoRA (Low-Rank Adaptation)
2. QLoRA (Quantized Low-Rank Adaptation) are two such techniques.

LoRA: Low-Rank Adaptation

Is a type of efficient fine-tuning technique

Problem statement:

The default (fine-tuning: full weight) for LLMs tends to be expensive, compute-intensive & slow. This becomes a major obstacle for companies to adapt these models for their business use.

How does LoRA it solve it :

LoRA reduces the number of trainable parameters for downstream tasks by freezing the weights of the model and inserting a smaller number of new weights into the model. This makes training with LoRA faster, memory-efficient and less costly. All while maintaining the quality of the model outputs..

LoRA (Freeze weights & insert small number of trainable weights)

Preprocessor: "gemma_causal_lm_preprocessor_1"

Tokenizer (type)	Vocab #
gemma_tokenizer (GemmaTokenizer)	256,000

Model: "gemma_causal_lm_1"

Layer (type)	Output Shape	Param #	Connected to
padding_mask (InputLayer)	(None, None)	0	-
token_ids (InputLayer)	(None, None)	0	-
gemma_backbone (GemmaBackbone)	(None, None, 2048)	2,506,172,416	padding_mask[0][0], token_ids[0][0]
token_embedding (ReversibleEmbedding)	(None, None, 256000)	524,288,000	gemma_backbone[0][0]

Total params: 2,506,172,416 (4.67 GB)
Trainable params: 2,506,172,416 (4.67 GB)
Non-trainable params: 0 (0.00 B)

- Freeze weights
- Insert small number of trainable weights

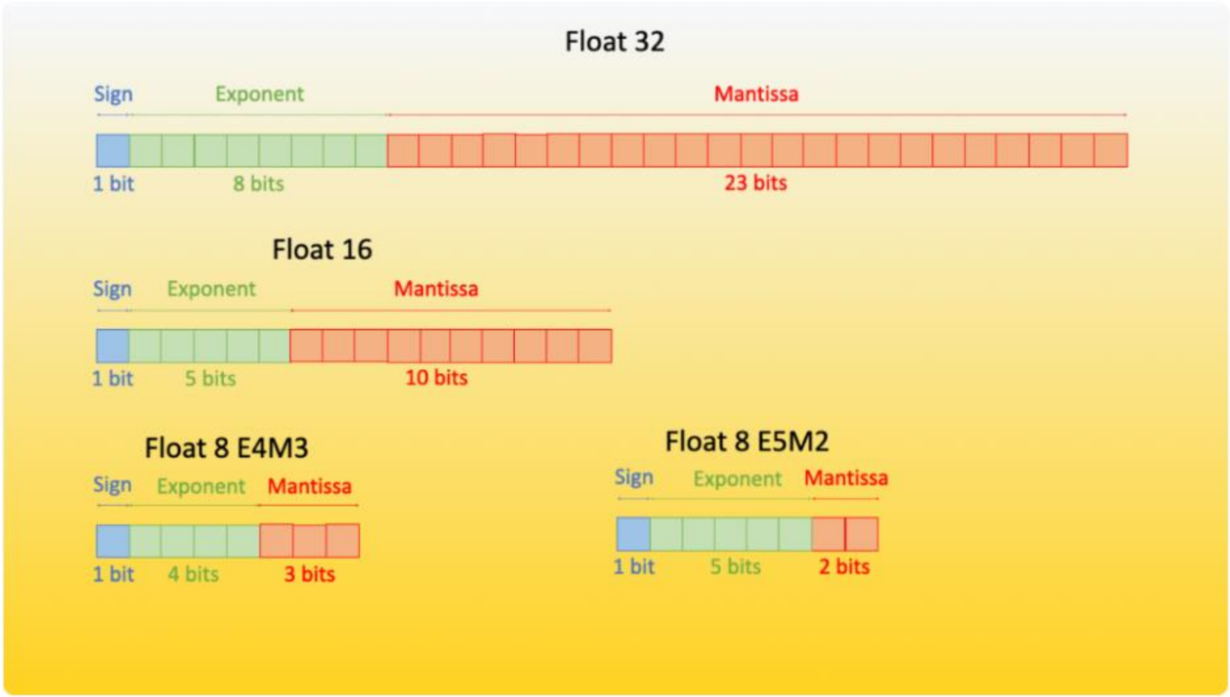
Model: "gemma_causal_lm_1"

Layer (type)	Output Shape	Param #	Connected to
padding_mask (InputLayer)	(None, None)	0	-
token_ids (InputLayer)	(None, None)	0	-
gemma_backbone (GemmaBackbone)	(None, None, 2048)	2,507,536,384	padding_mask[0][0], token_ids[0][0]
token_embedding (ReversibleEmbedding)	(None, None, 256000)	524,288,000	gemma_backbone[0][0]

Total params: 2,507,536,384 (4.67 GB)
Trainable params: 1,363,968 (2.60 MB)
Non-trainable params: 2,506,172,416 (4.67

QLoRA (Quantized Low-Rank Adaption)

Building on LoRA, QLoRA incorporates quantization into the adaptation process. It uses techniques like 4-bit quantization to further reduce the memory and computational requirements. This quantization applies not just to the low-rank matrices but potentially also to other parts of the model, making the storage and processing of model weights more efficient



Overview of Floating Point 8 (FP8) format. Source: Original content from sgugger

huggingface.co/blog/4bit-transformers-bitsandbytes

(track 3): Custom Model

LLM

How to unlock the full potential of LLMs to solve business problems

Track 1

Use model as-is.

Model is not modified.

Track 2

Model is modified

Track 3

Build your own & train

Custom Model

Train

Prompt Engineering

Few-shot

Zero-Shot

Chain-of-Thought

RAG

(Retrieval Augmented Generation)

Add content from an external vector database to your prompt

Fine Tuning

LoRA

QLoRA

High Cost

High Complexity

LLMOps

(training, deployment, monitoring, a

maintaining of models)