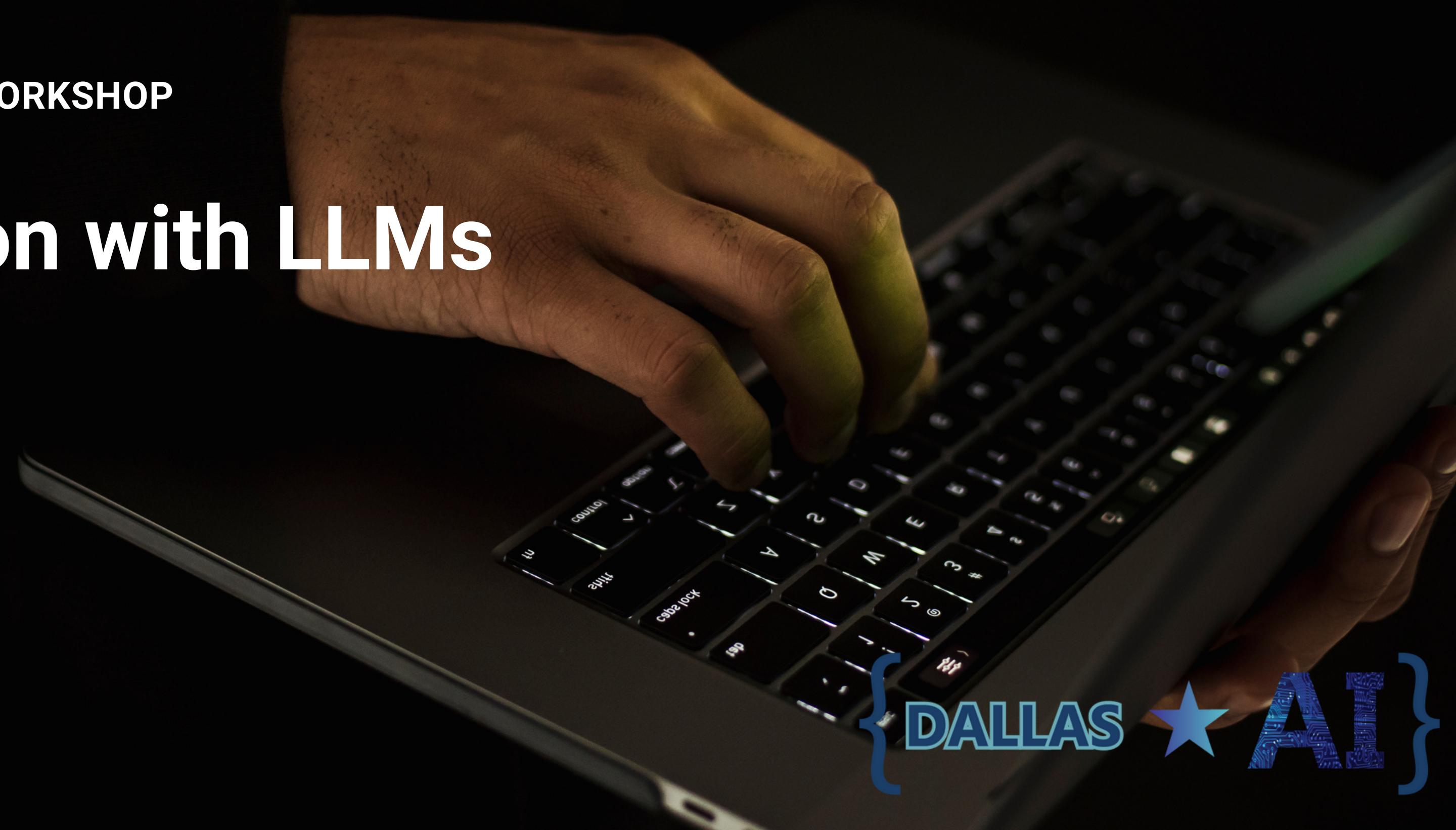


GENERATIVE AI WORKSHOP

Hands-on with LLMs

For developers

Dec 16 2023



{ DALLAS ★ AI }

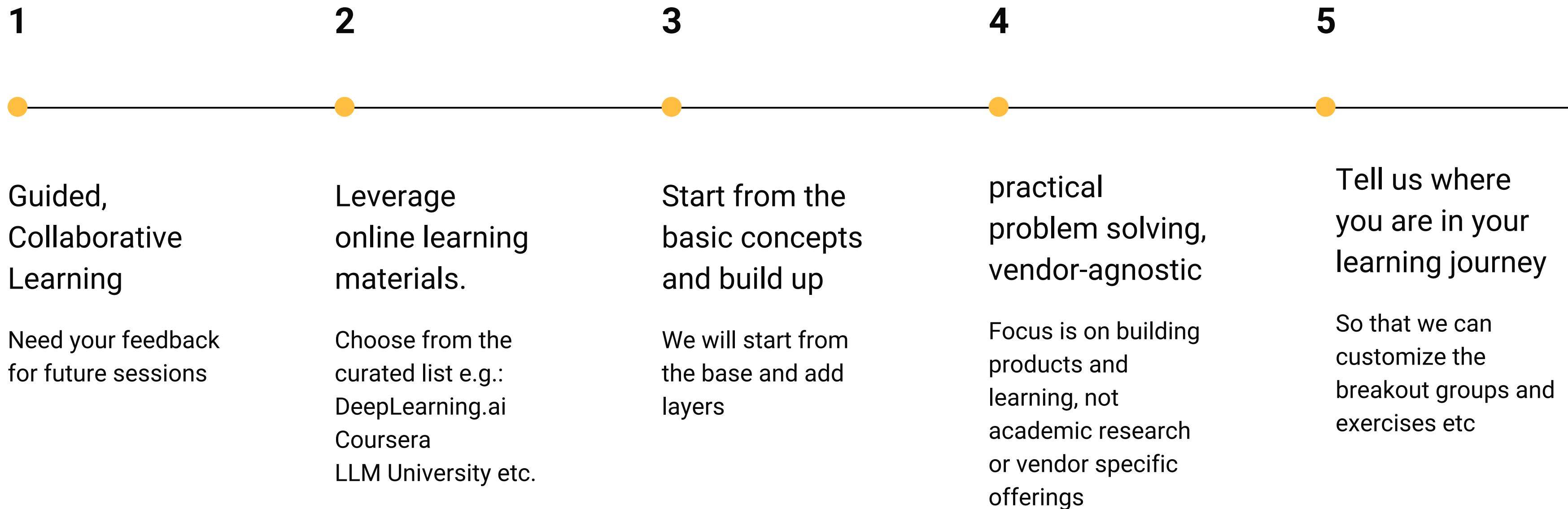


CONTENT

Today's Topics

- 01 Overview of Generative AI & Foundation Models
- 02 Getting started with LLMs - Types and Design Patterns
- 03 LLMs for Enterprises: Performance, Cost + Safety, Security, Trust, Architecture and Workflow
- 04 Customizing Approaches for LLMs, When/How to Use
- 05 Tools and Trends (Functions/Libraries, Vector databases)
- 06 Semantic Search with LLMs. Retrieval Augmented Generation
- 07 Resources

About the developer session #1 (LLMs)

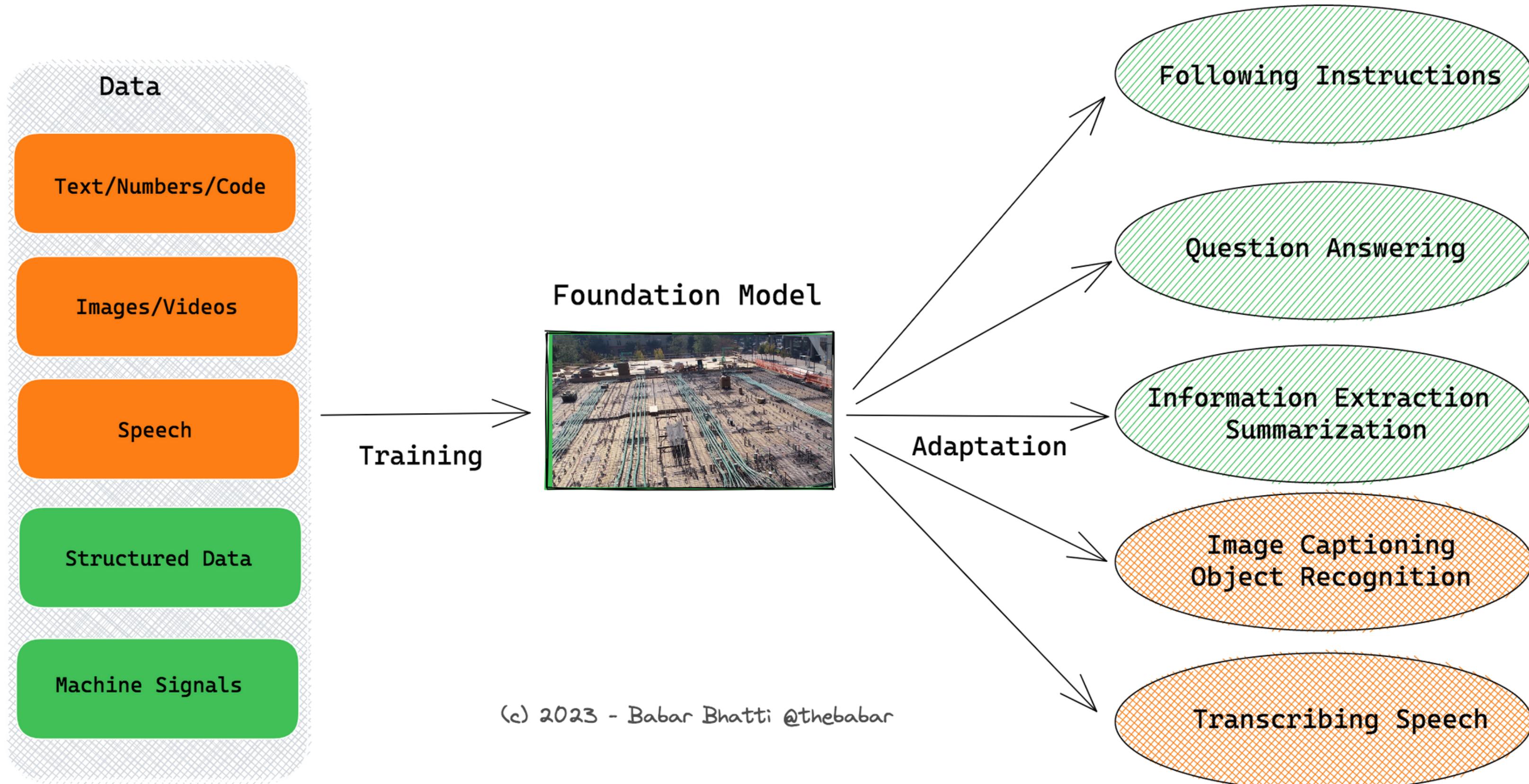


LLMs are _____
they are not _____

1

Generative AI & Foundation Models

Power of Foundation Models



Foundation Models (FMs)

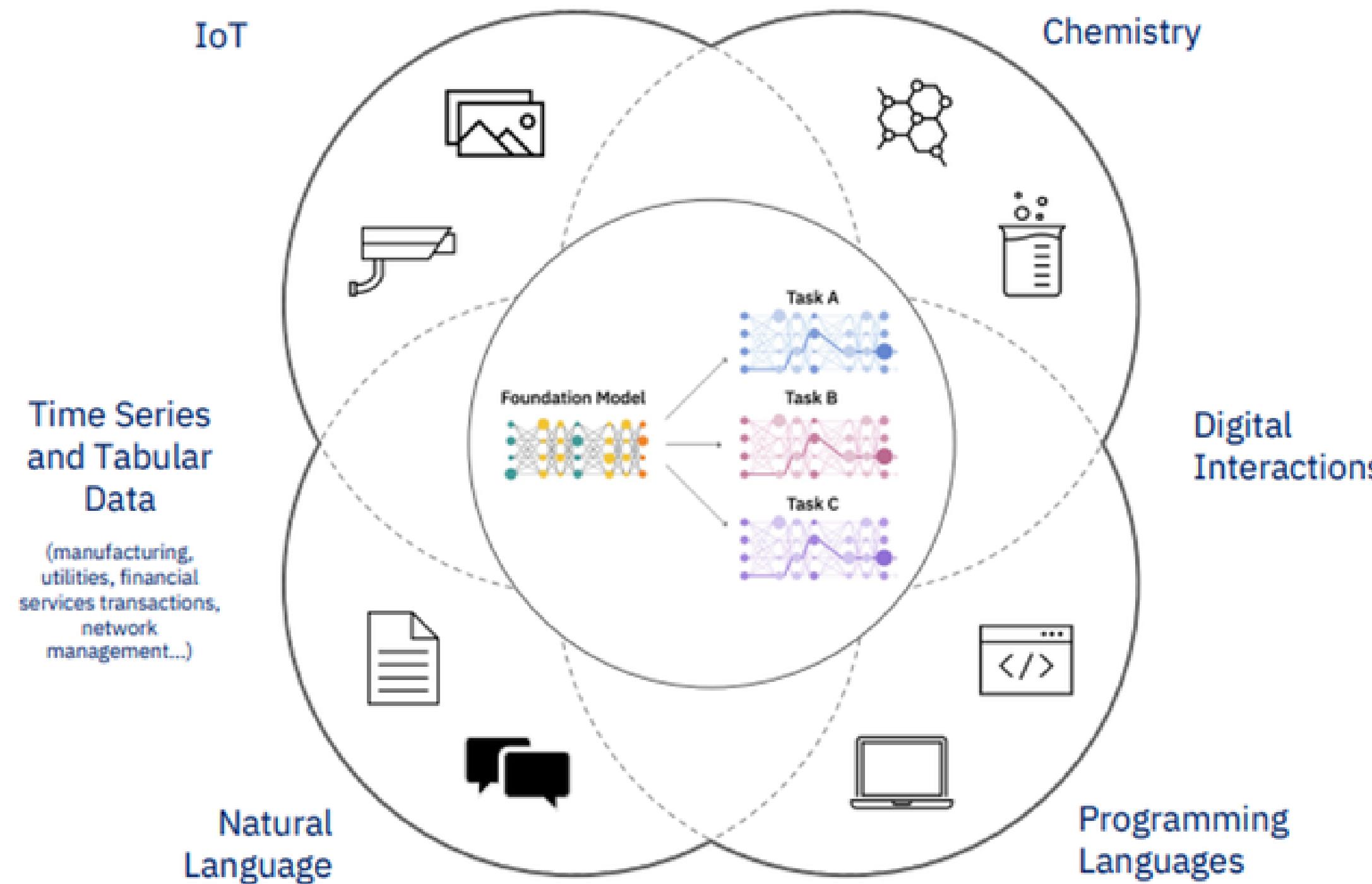
- * Pretrained
- * Generalized
- * Adaptable
- * Large
- * Self-supervised
- + For Generative & non-generative AI

Large Language Models (LLMs)

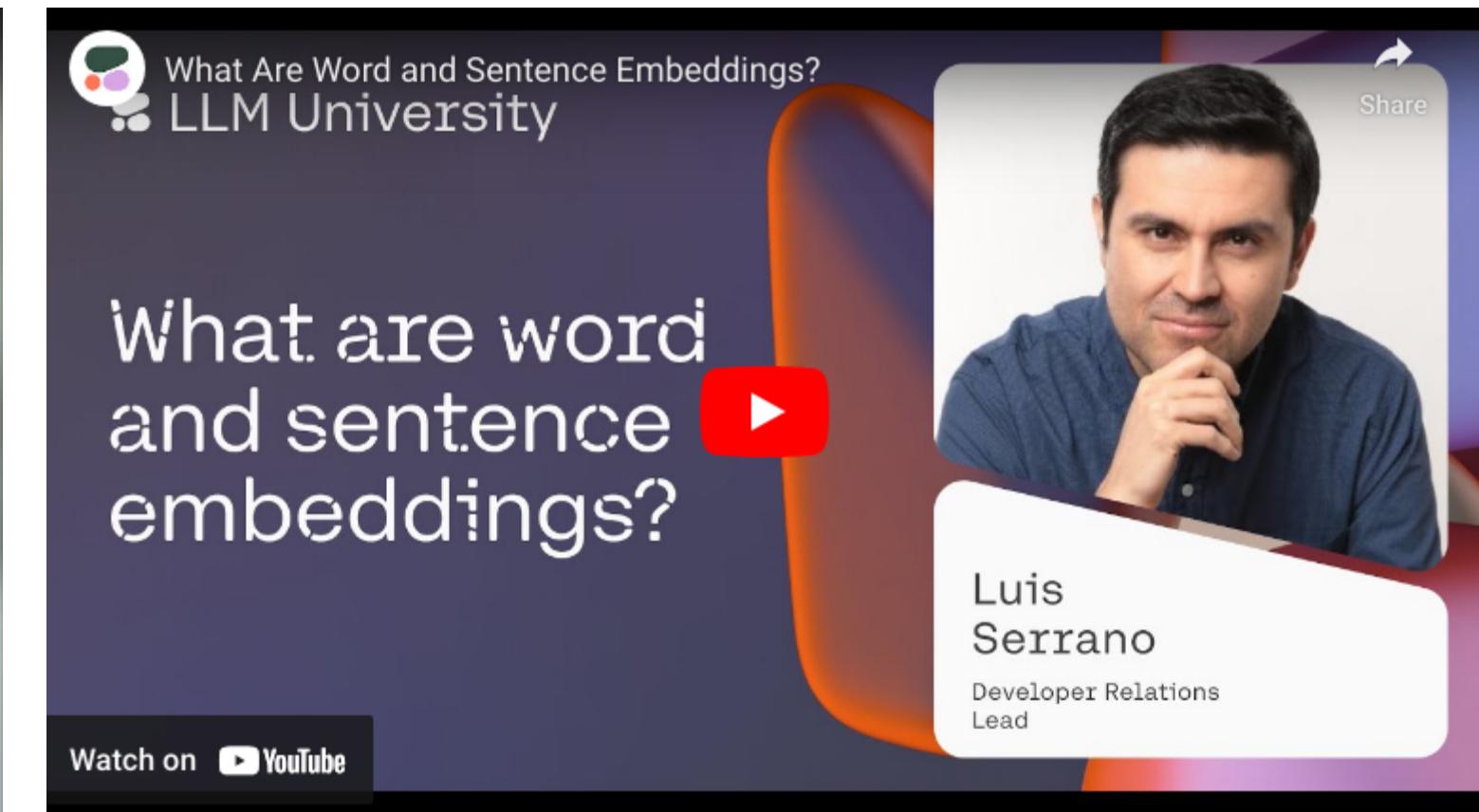
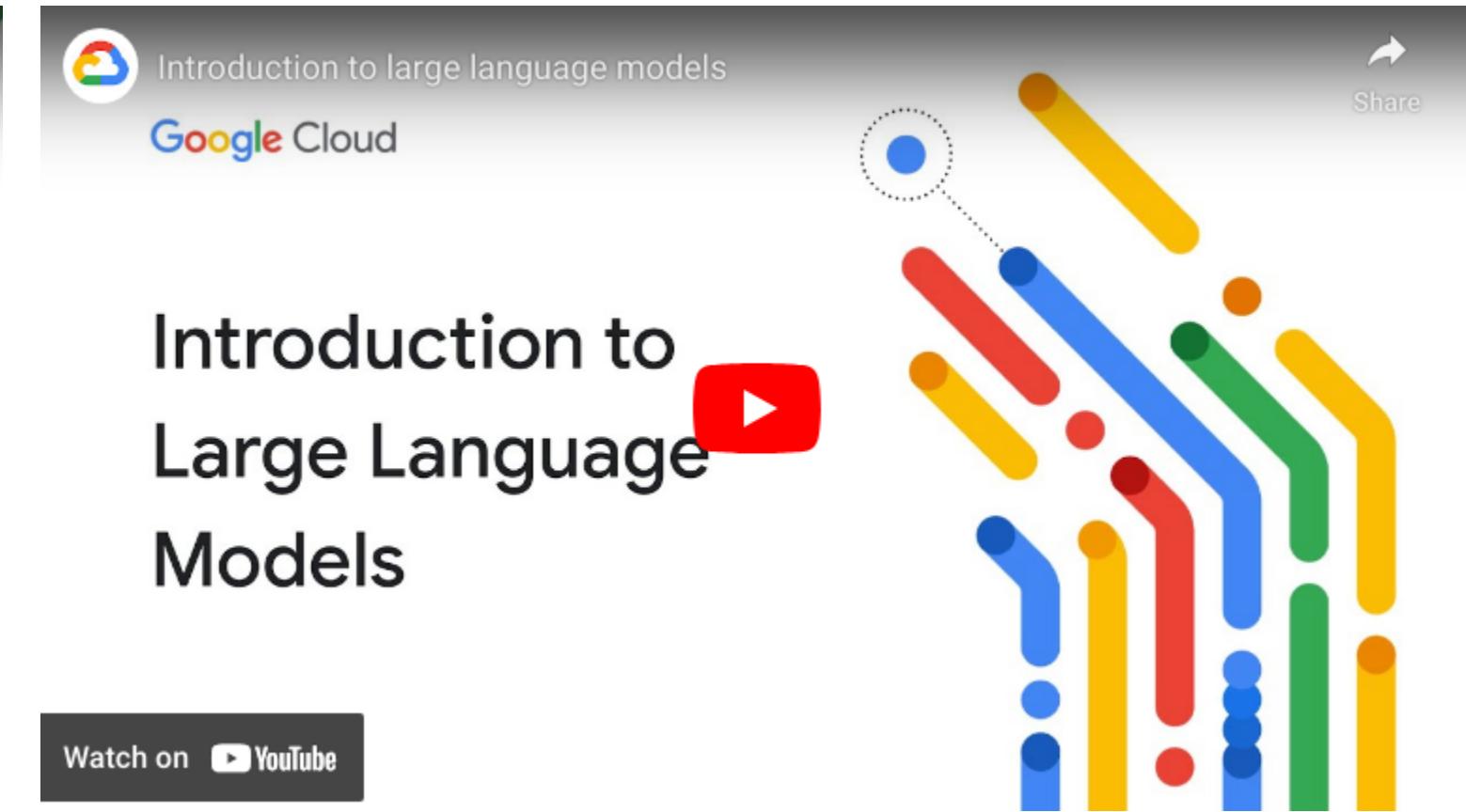
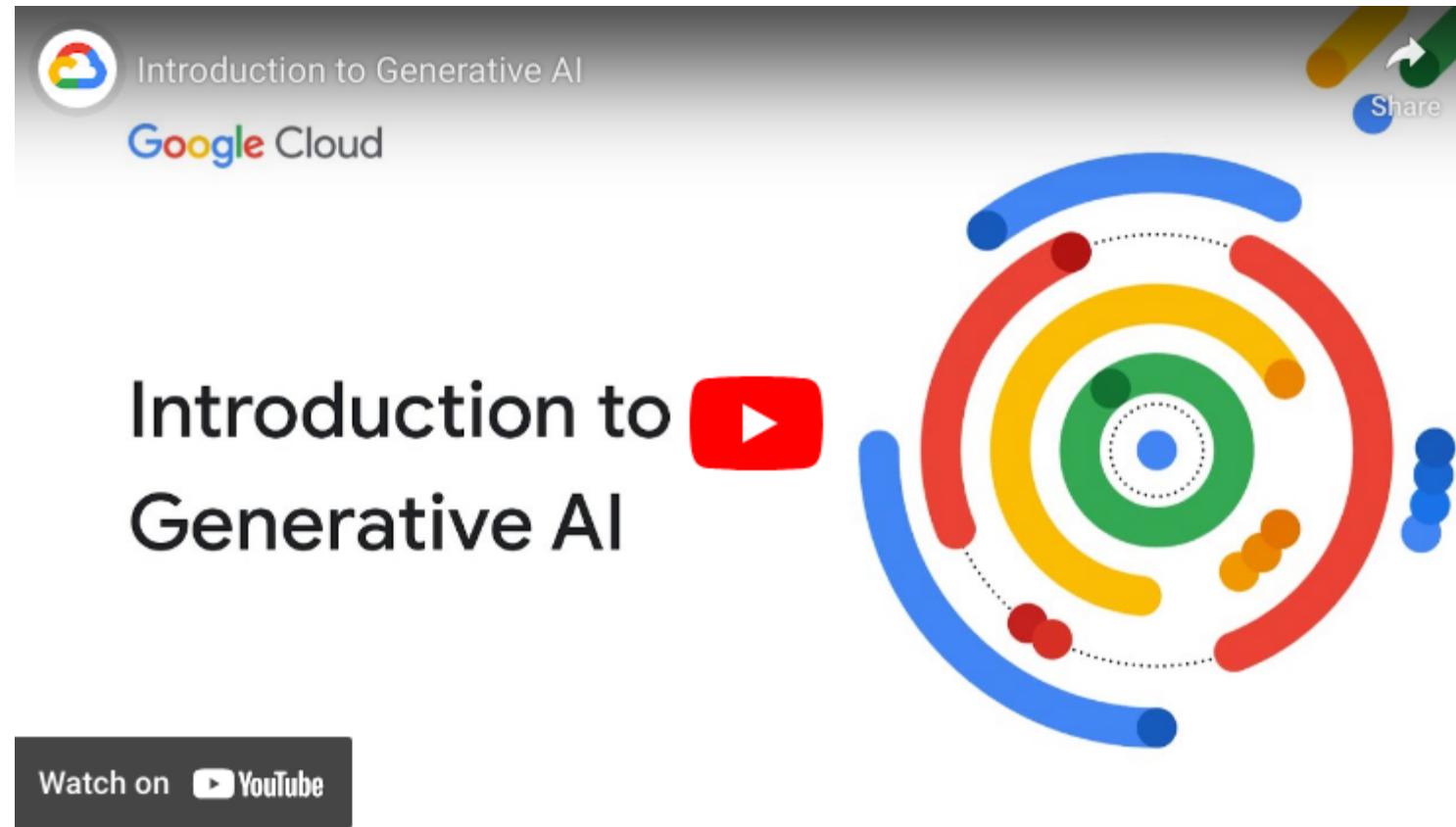
Chat Models (eg ChatGPT)
tuned for
following instructions

FMs are models trained on broad data (using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

Foundation Models Apply to Many Domains



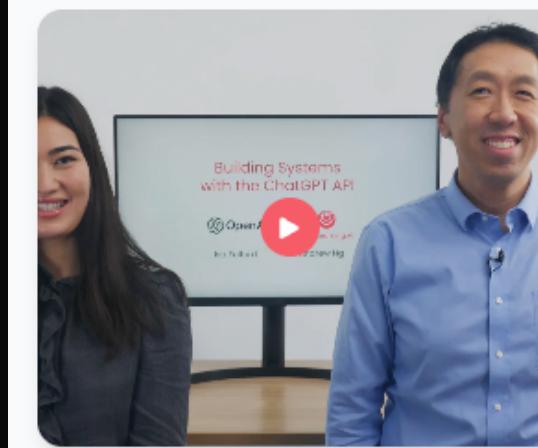
Intro Video





SCOPE OF OUR SESSIONS

LLM COURSES



Building Systems with the ChatGPT API

Level up your use of LLMs. Learn to break down complex tasks, automate workflows, chain LLM calls, and get better outputs.

Prerequisite recommendation: Basic Python

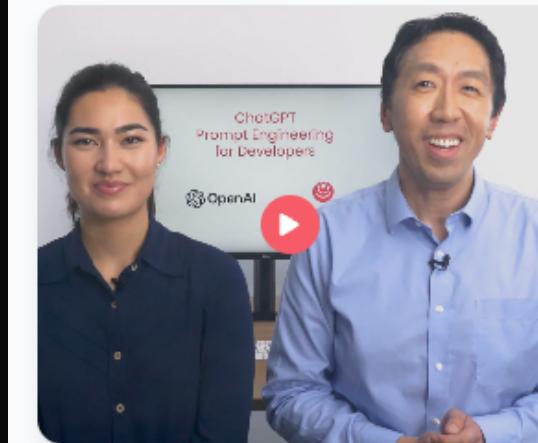
Beginner to Advanced

1 Hour

Free for a limited time

[Enroll For Free](#)

[Learn more](#)



ChatGPT Prompt Engineering for Developers

Go beyond the chat box. Use API access to leverage LLMs into your own applications, and learn to build a custom chatbot.

Prerequisite recommendation: Basic Python

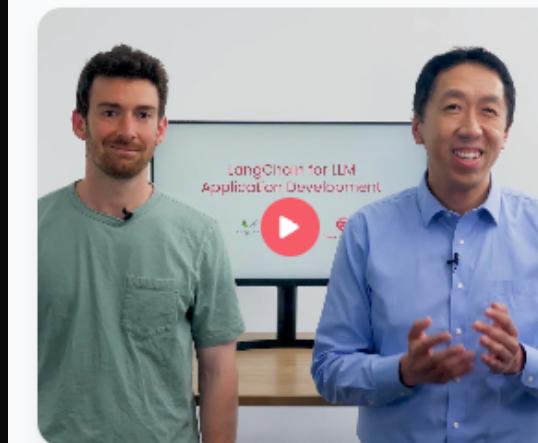
Beginner to Advanced

1 Hour

Free for a limited time

[Enroll For Free](#)

[Learn more](#)



LangChain for LLM Application Development

The framework to take LLMs out of the box. Learn to use LangChain to call LLMs into new environments, and use memories, chains, and agents to take on new and complex tasks.

Prerequisite recommendation: Basic Python

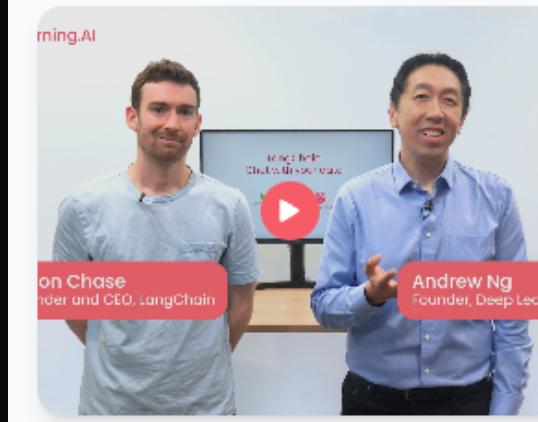
Beginner

1 Hour

Free for a limited time

[Enroll For Free](#)

[Learn more](#)



LangChain: Chat with Your Data

Create a chatbot to interface with your private data and documents using LangChain.

Prerequisite recommendation: Basic Python

Beginner

1 Hour

Free for a limited time

[Enroll For Free](#)

[Learn more](#)

Course Content At a Glance

Building Systems with the ChatGPT API

- Introduction
- Language Models, the Chat Format and Prompts
- Classification
- Moderation
- Chain of Thought Reasoning
- Chaining Prompts
- Check Outputs
- Evaluation
- Evaluation Part I
- Evaluation Part II
- Summary
- Course Feedback
- Community

ChatGPT Prompt Engineering for Developers

- Introduction
- Guidelines
- Iterative
- Summarizing
- Inferring
- Transforming
- Expanding
- Chatbot
- Conclusion
- Course Feedback
- Community

LangChain for LLM Application Development

- Introduction
- Models, Prompts and parsers
- Memory
- Chains
- Question and Answer
- Evaluation
- Agents
- Conclusion
- Course Feedback
- Community

LangChain Chat with Your Data

- Introduction
- Document Loading
- Document Splitting
- Vectorstores and Embedding
- Retrieval
- Question Answering
- Chat
- Conclusion
- Course Feedback
- Community

2

Large Language Models 101



Looking for specific info?

Search in reviews, Q&A...

Docs AI Ask a question

These answers are generated using artificial intelligence. This is an experimental technology, and information may occasionally be incorrect or misleading.

Customer reviews

★★★★★ 4.4 out of 5

41,661 global ratings

Customers say

Customers like the ease of use, quality, value, and size of the air mattress. They mention that it's convenient for when guests stay over, it holds up night long, and is worth the price.

AI-generated from the text of customer reviews

Quality Comfort Ease of use

Size Value Inflation

Performance Leakage



How can I view and send receipts for payments in the Stripe Dashboard?

To view and send receipts for payments in the Stripe Dashboard, you can follow these steps:

1. Log in to the Stripe Dashboard.
2. Go to the Payments section of the Dashboard.
3. Locate the payment for which you want to view or send a receipt.
4. Click on the payment to access its page.
5. In the receipt history section, you can manually send a receipt by providing an email address or resend a receipt to a different email address. You can also view a history of the last 10 receipts for that payment.

Additionally, you can customize the appearance and content of your receipts by adjusting the settings in the Stripe Dashboard. You can specify the language to use, customize your logo and colors, and include public information such as your phone number or website address. You can even add custom text to the receipt description using the description attribute on the PaymentIntent.

For more information on how to customize receipts and other receipt features, you can refer to the [Email receipts](#) documentation.

Please note that receipts for payments created using test API keys are not sent automatically in test mode. Instead, you can view or manually send a receipt using the [test Dashboard](#).

Was this useful?



BA

How can accountants and controllers use ChatGPT to help with their work?



Accountants and controllers can use ChatGPT to enhance their work in various ways. Here are some practical applications:

1. Data Analysis:

- ChatGPT can assist in data analysis by answering questions about financial data, helping identify trends or anomalies in financial reports, and offering insights into the data.

2. Financial Modeling:

- Accountants can use ChatGPT to assist in creating and refining financial models. ChatGPT can help with formulae, assumptions, and provide guidance on building complex models.

3. Accounting Standards:

- ChatGPT can provide information and guidance on accounting standards, helping accountants and controllers stay up-to-date with regulations and compliance requirements.

4. Research:

- Accountants can use ChatGPT to research topics related to accounting, taxation, or finance. It can provide information on tax laws, industry benchmarks, and best practices.

5. Documentation:

- ChatGPT can help in drafting financial reports, memos, or other documents. It can provide suggestions for wording, formatting, and ensuring clarity in communication.

6. Financial Forecasting:

- Controllers can use ChatGPT to aid in financial forecasting by providing historical data and asking for predictions or insights regarding future financial performance.

7. Cost Analysis:

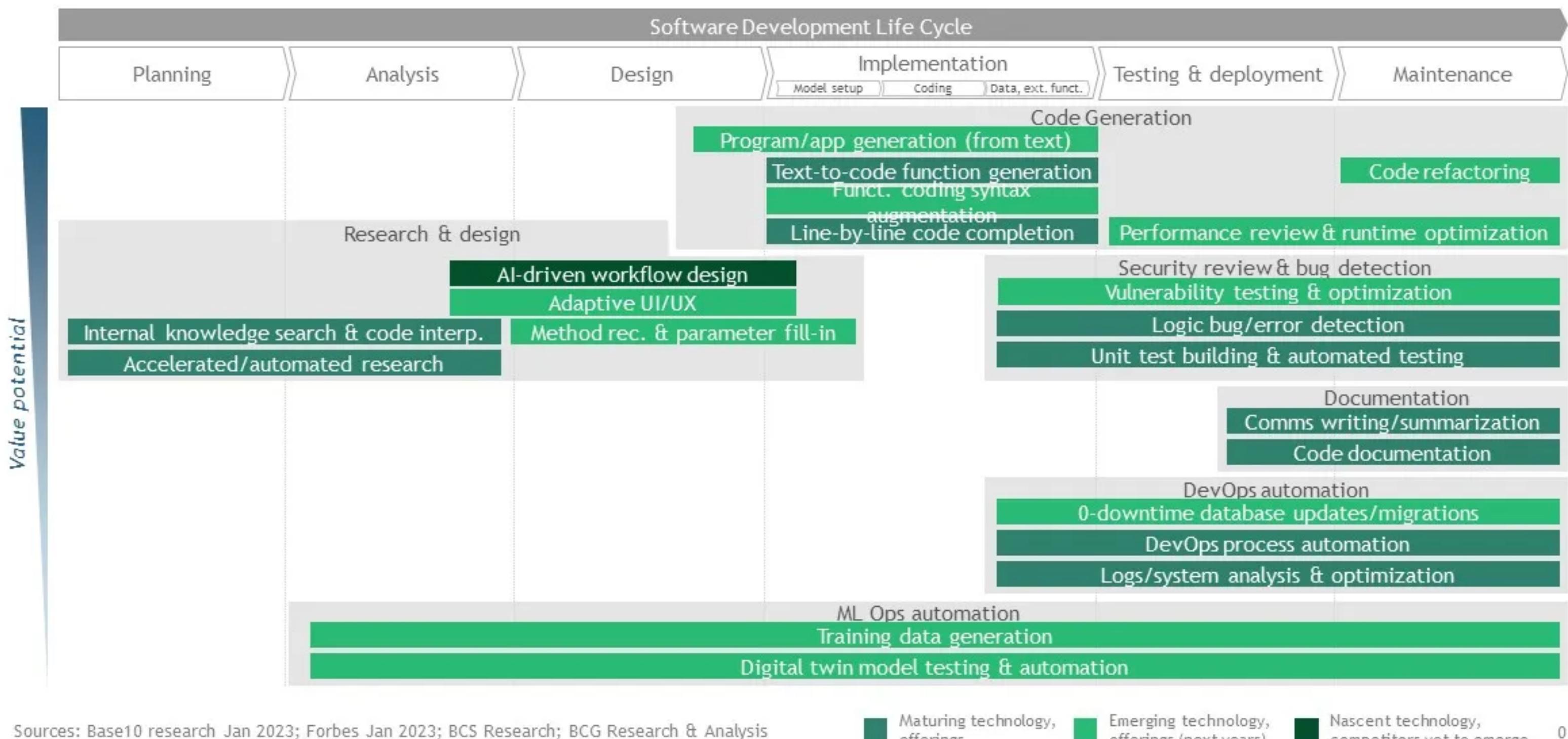
- Accountants can seek cost analysis assistance from ChatGPT, whether it's for cost allocation, cost reduction strategies, or comparing expenses across different time periods.

8. Audit Support:

- During audits, ChatGPT can help by summarizing relevant data, explaining transactions, and providing insights to auditors to streamline the process.

9. Troubleshooting and Problem-Solving:

Emerging tools for ~50% of potential gen AI use cases in software development, other use cases expected to be addressed in next 3-5 years



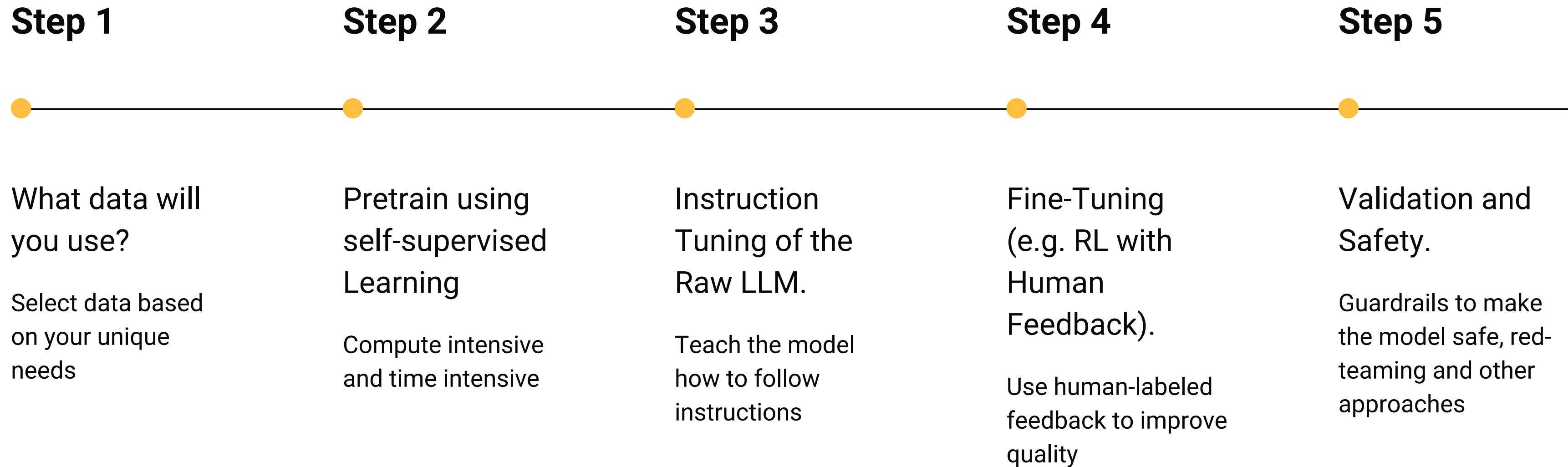


#2

GETTING STARTED WITH LLMs

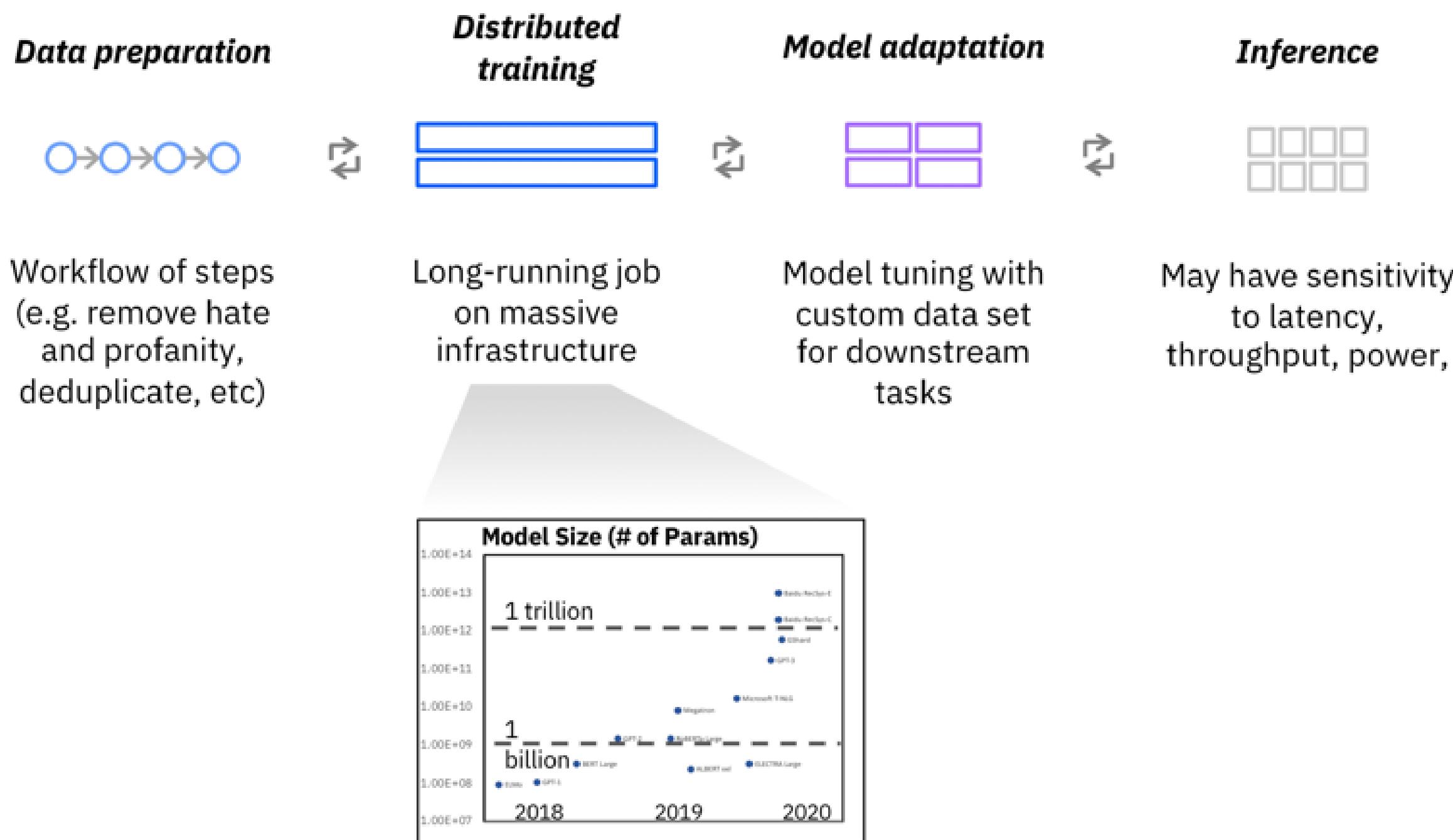
- 01 What types of model types (open source vs proprietary/closed) are available??
- 02 Training LLMs*
- 03 Prompt Engineering
- 04 Safety for LLMs

Building LLMs

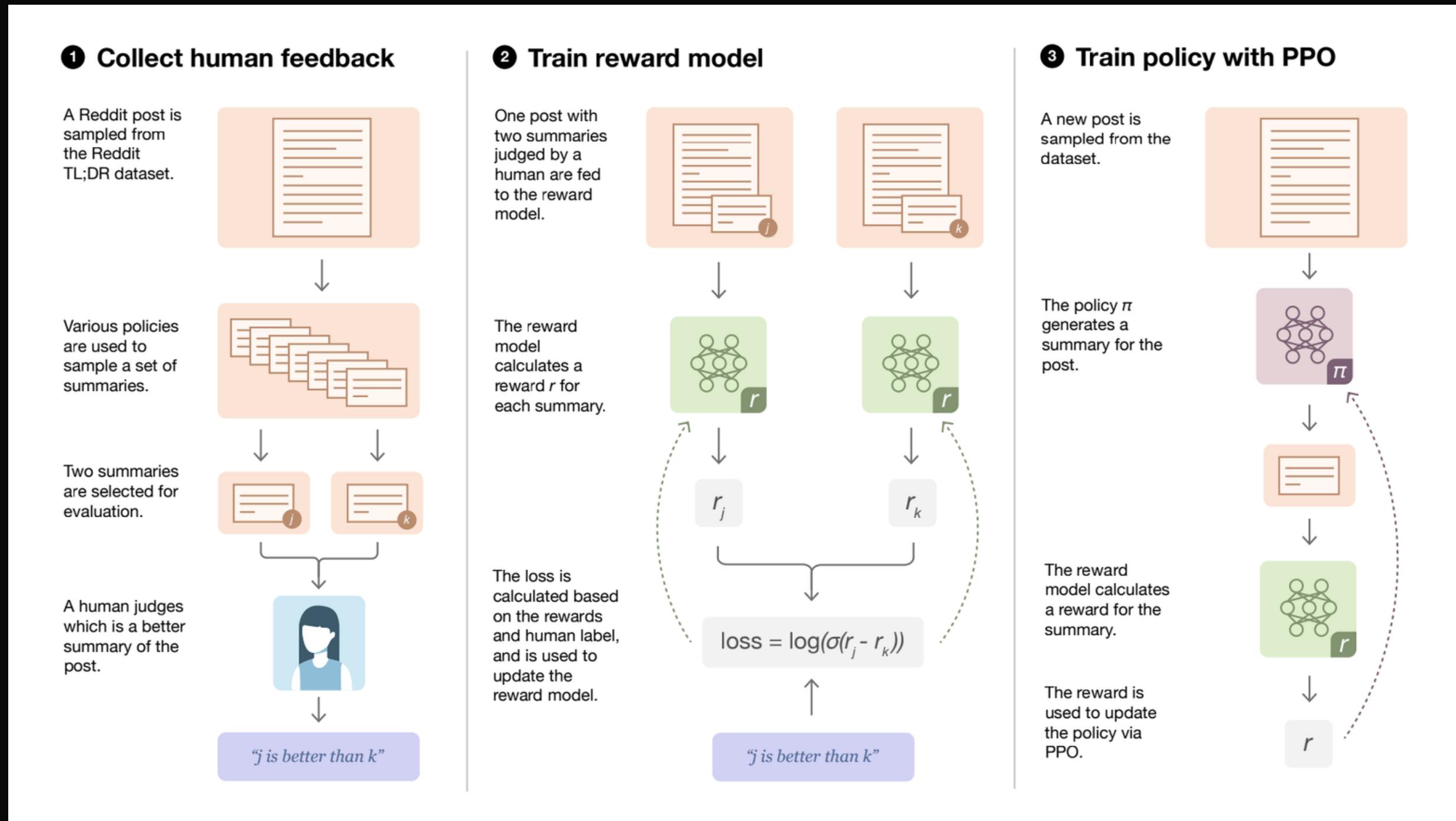


Workflow for Training LLMs

The workflow



Training LLMs Requires AI+Humans

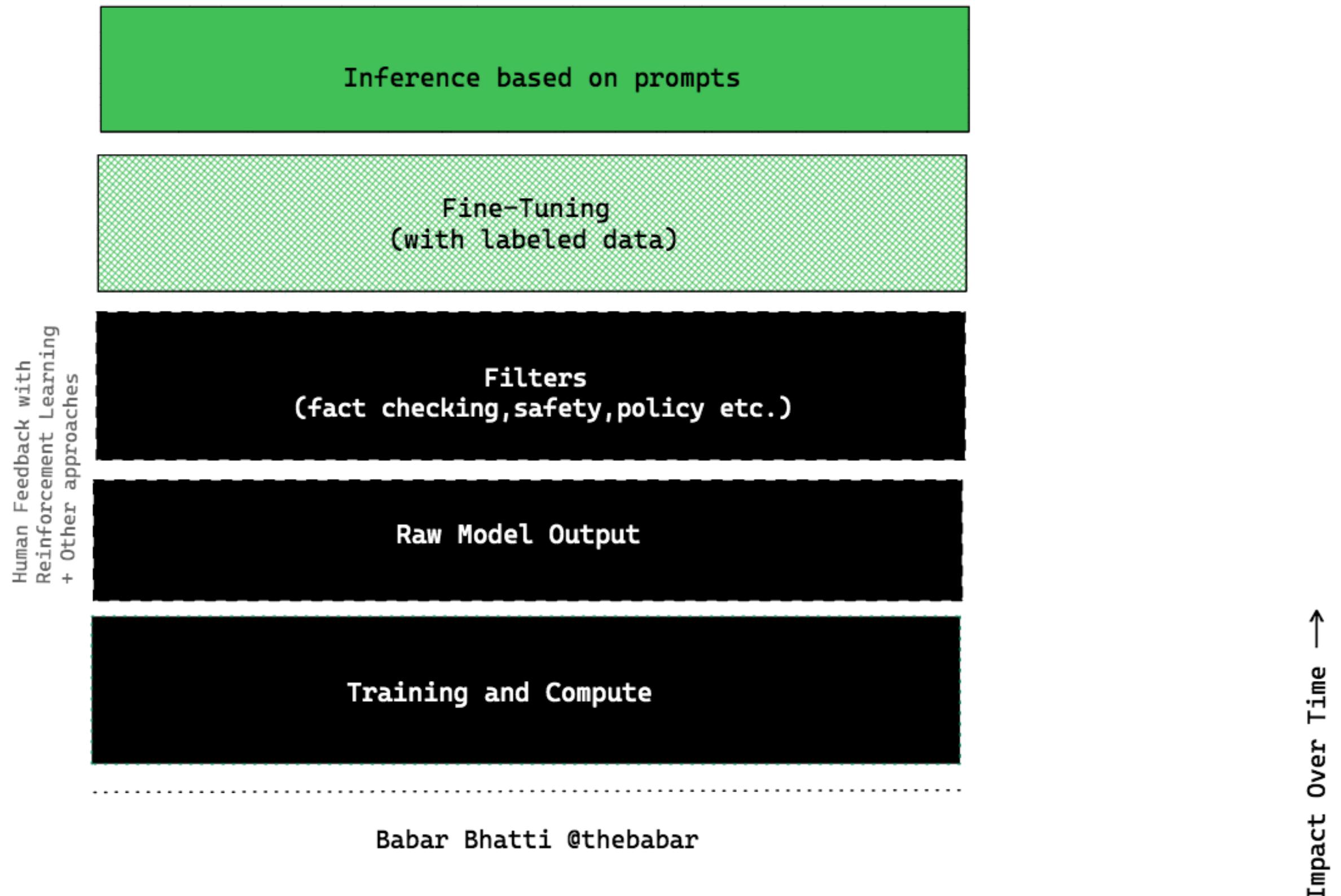


Prompt Engineering Methods & Skills

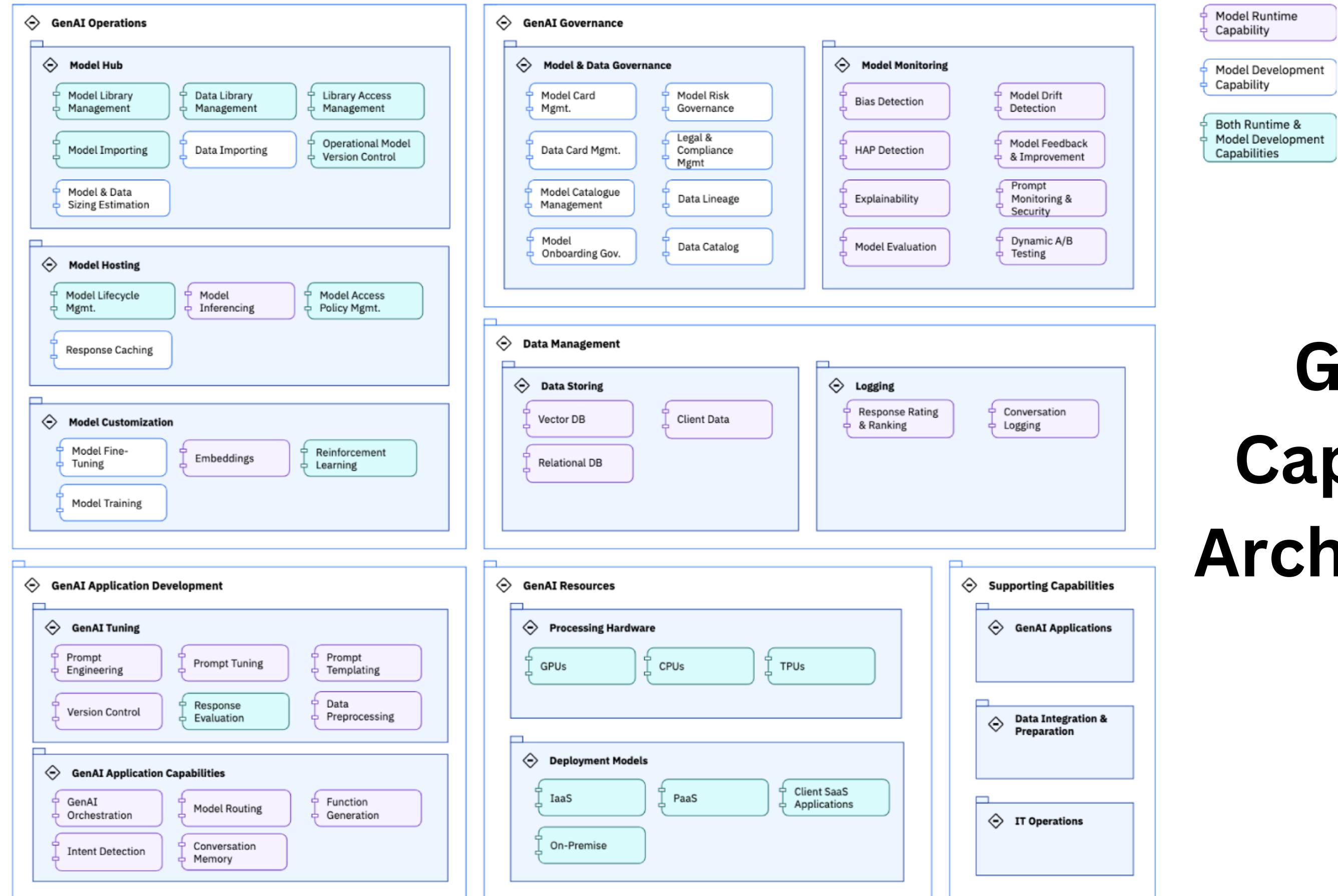
	Technical Skills	Soft Skills	Tools
Core	Prompt Formulation/Chaining	Writing/Communication Business Domain Knowledge	Prompt Management
Valuable	Advanced Prompting Methods Prompt Monitoring/Relevance Scoring	Creativity Reasoning Product Sense Thinking End-To-End Collaboration	Prompt Engineering/Prompt Infrastructure Search/Indexing/Vector Databases
Specialized	Semantic Search Knowledge Engineering Adversarial Prompting Prompt Optimization	Architecture User Empathy Design Thinking Persuasion	Automation/Workflow Platforms Symbolic AI Plugins Data Labeling

LLM Stack

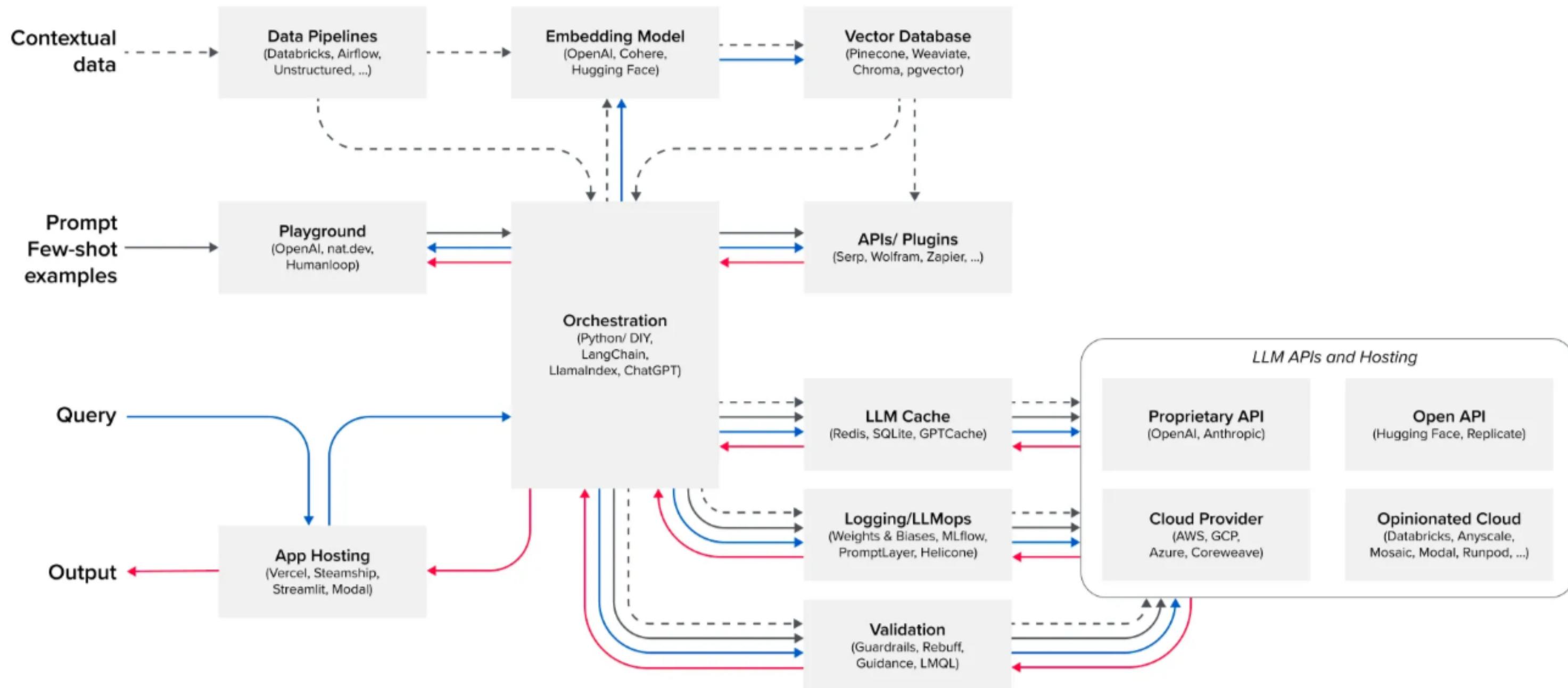
Layers of Large Language Model



Gen AI Capability Architecture



Emerging LLM App Stack



LEGEND

Gray boxes show key components of the stack, with leading tools/systems listed

Arrows show the flow of data through the stack

- - - → Contextual data provided by app developers to condition LLM outputs
- Prompts and few-shot examples that are sent to the LLM
- Queries submitted by users
- Output returned to users

HuggingFace.co

Open LLM Leaderboard

Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4

Discover amazing ML apps made by the community

huggingface

The 🎉 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🎉 Submit a model for automated evaluation on the 🎉 GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!

LLM Benchmark Metrics through time About Submit here!

Search for your model (separate multiple queries with `;`) and press ENTER...

Select columns to show

Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande
<input checked="" type="checkbox"/>					
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>					
<input type="checkbox"/>					

Model types

pretrained	fine-tuned	instruction-tuned	RL-tuned	?
<input checked="" type="checkbox"/>				

Precision

float16	bfloat16	8bit	4bit	GPTQ	?
<input checked="" type="checkbox"/>					

Model sizes (in billions of parameters)

?	~1.5	~3	~7	~13	~35	~60	~70+
<input checked="" type="checkbox"/>							

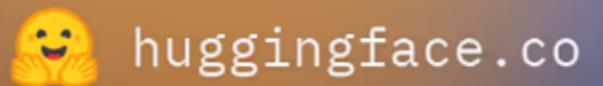
Show private/deleted models

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
□	upstage/SOLAR-10.7B-Instruct-v1.0	74.2	71.08	88.16	66.21	71.43	83.58	64.75
◆	fbigit/una-xaberius-34b-v1beta has been flagged! See discussion #444	74.18	70.39	86.77	78.15	61.45	84.93	63.38
?	viethq188/LeoScorpius-7B-Chat-DPO	73.92	70.48	87.97	65.08	68.83	82.08	69.07
◆	ValiantLabs/ShiningValiant	73.78	71.33	90.96	71.21	70.29	84.21	54.66
◆	sequelbox/SunsetBoulevard	73.78	71.33	90.96	71.21	70.29	84.21	54.66
□	violet/violet_merged_dpo_7B_v3	73.68	72.61	87.7	63.51	69.07	82.32	66.87
?	GreenNode/GreenNodeLM-7B-violet	73.68	72.61	87.7	63.51	69.07	82.32	66.87
□	Owen/Owen-72B	73.6	65.19	85.94	77.37	60.19	82.48	70.43
○	SUSTech/SUS-Chat-34B	73.22	66.3	83.91	76.41	57.04	83.5	72.18
◆	violet/violet_marcoroni-go-bruins-merge-7B	72.81	70.05	87.17	65.17	61.42	81.45	71.57
○	mistralai/Mixtral-8x7B-Instruct-v0.1	72.62	70.22	87.63	71.16	64.58	81.37	60.73
◆	AIDC-ai-business/Marcoroni-7B-v3	72.53	69.45	86.78	65	60.4	81.45	72.1

Model FactSheets

meta-llama

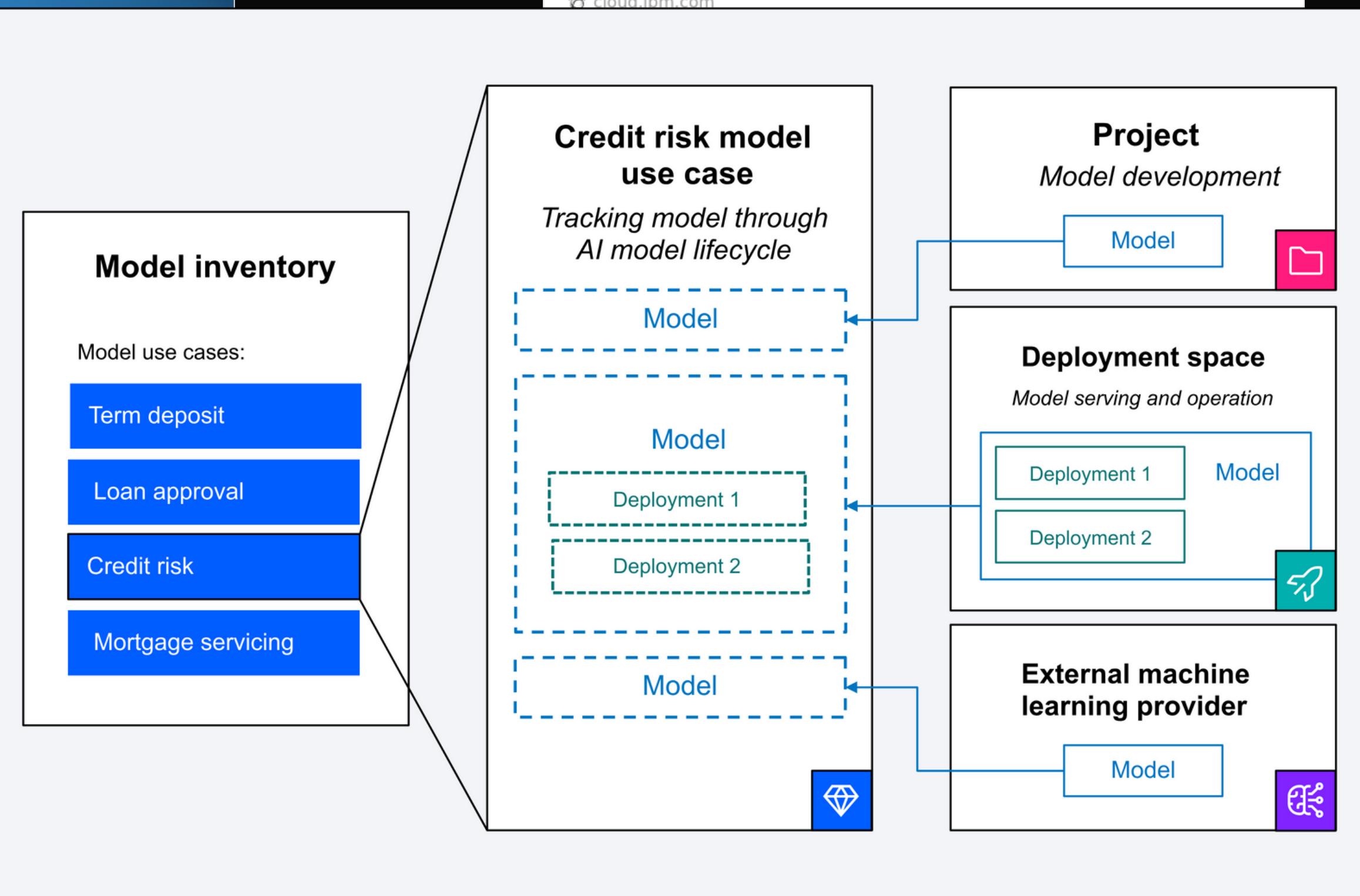
/Llama-2-7b-chat-hf



meta-llama/Llama-2-7b-chat-hf · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open

huggingface



Using AI Factsheets for AI Governance | IBM Cloud Pak for Data as a Service

Track a machine learning model from request to production by gathering metadata and facts about the model lifecycle using AI Factsheets. Use the detailed information in the factsheets to keep stakeholders informed and to...

[cloud.ibm.com](#)

Risks of LLM

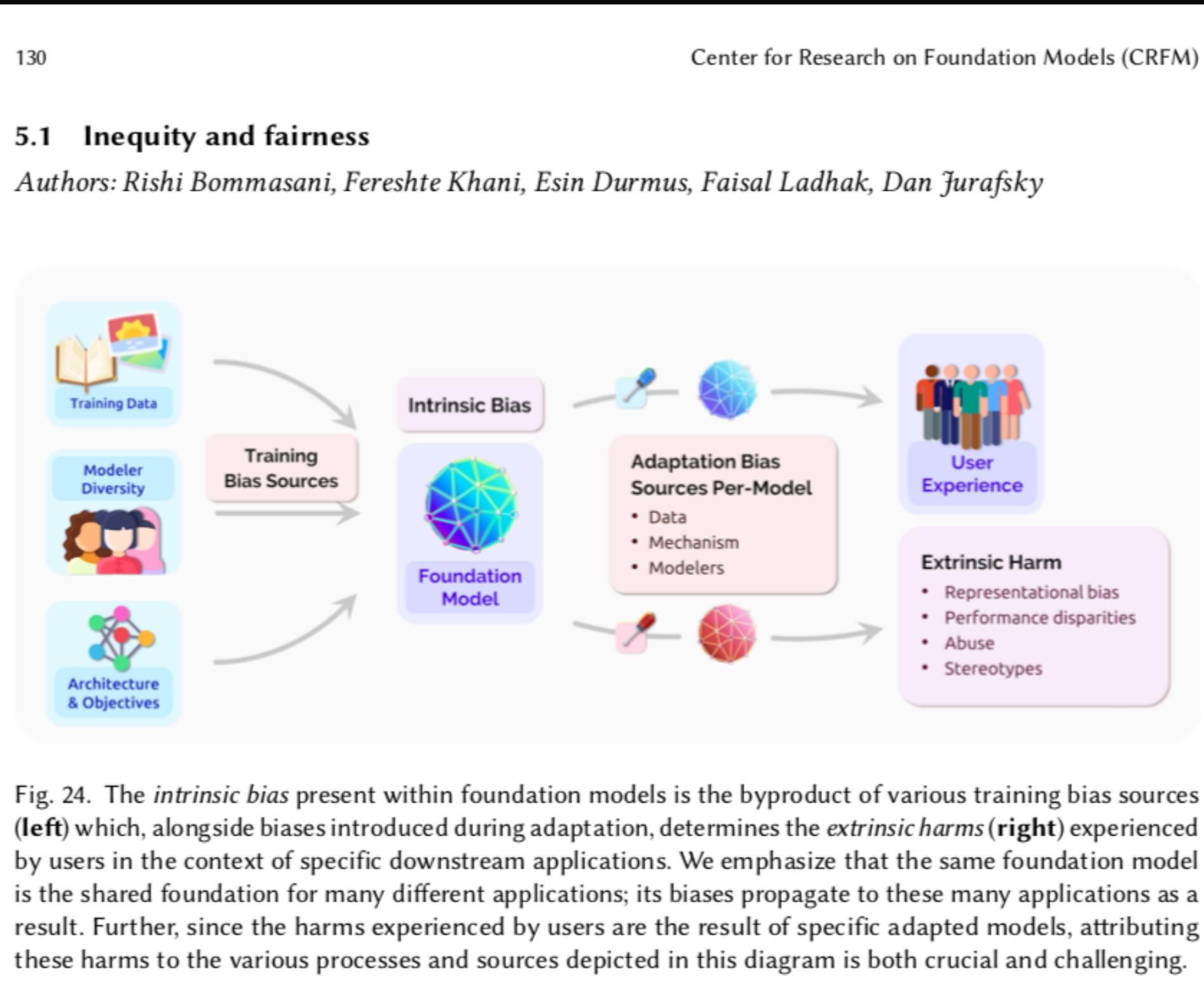


Fig. 24. The *intrinsic bias* present within foundation models is the byproduct of various training bias sources (**left**) which, alongside biases introduced during adaptation, determines the *extrinsic harms* (**right**) experienced by users in the context of specific downstream applications. We emphasize that the same foundation model is the shared foundation for many different applications; its biases propagate to these many applications as a result. Further, since the harms experienced by users are the result of specific adapted models, attributing these harms to the various processes and sources depicted in this diagram is both crucial and challenging.

garymarcus.substack.com

The Road to AI We Can Trust

Scientists, governments, and corporations urgently need to work together to mitigate AI risk

"It is hard to see how you can prevent the bad actors from using it for bad things" -- but we must try

GARY MARCUS MAY 1, 2023

17 11 Share ...

'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead

For half a century, Geoffrey Hinton nurtured the technology at the heart of chatbots like ChatGPT. Now he worries it will cause serious harm.

New York Times, May 1, 2023

FORTUNE RANKINGS MAGAZINE NEWSLETTERS PODCASTS MORE SEARCH

TECH · A.I.
A robot's \$100 billion error: Alphabet shuns tank after its ChatGPT rival makes a mistake in its very first ad

BY TRISTAN BOVE
February 8, 2023 at 12:42 PM PST

CIO JOURNAL

Generative AI Pilots Have Companies Reaching for the Guardrails

As JPMorgan Chase and others seek out a safe way to experiment with OpenAI's ChatGPT and other advanced AI tools, vendors like Nvidia offer protection against the tech's potential downsides

By Belle Lin [Follow](#)

May 19, 2023 5:42 pm ET



OpenAI's ChatGPT can assume a lot of tasks that aid companies, but it poses potential problems as well.
PHOTO: RICHARD A. BROOKS/AGENCE FRANCE PRESSE/GETTY IMAGES

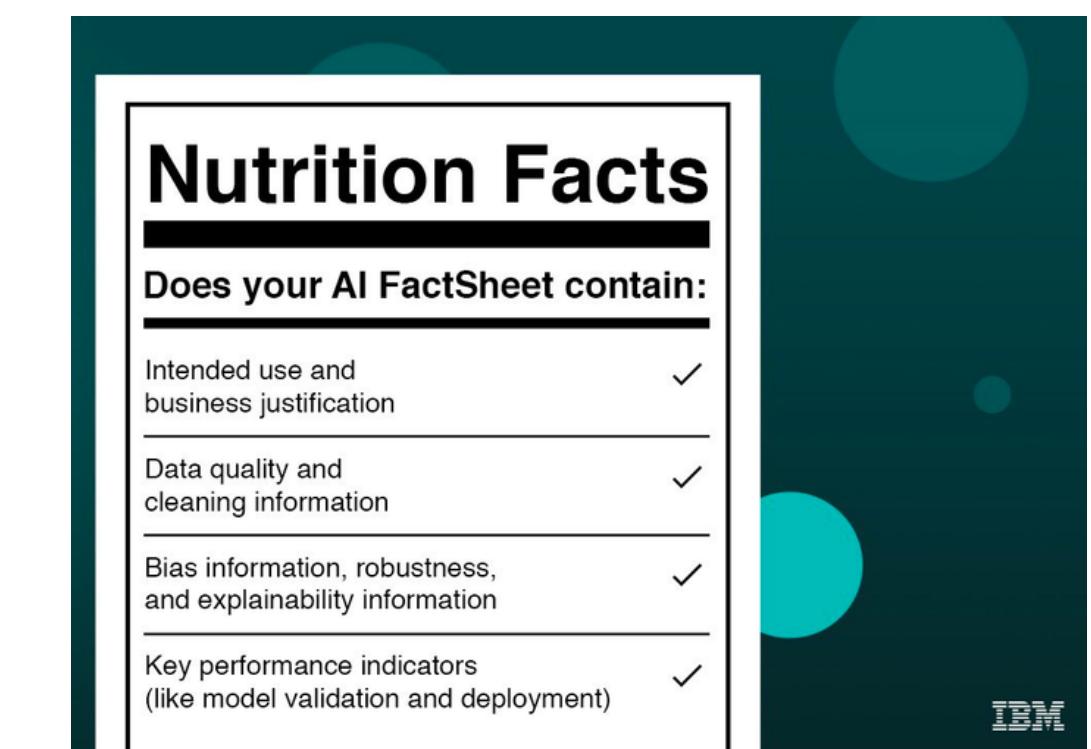
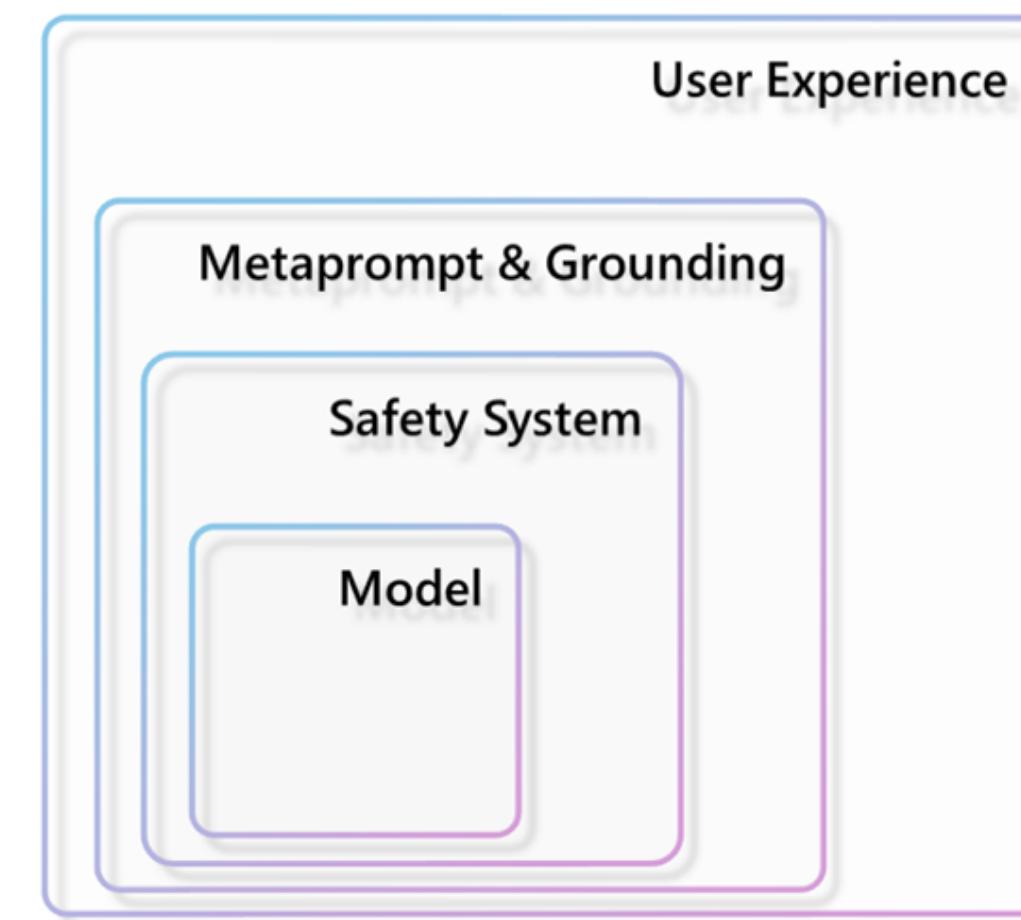
Hoping to take advantage of artificial intelligence technologies like ChatGPT, but without all of its risks, businesses are turning to in-house solutions and a growing ecosystem of vendors that aim to protect sensitive company data that could be fed into new AI tools.

Generative AI has the potential to completely upend office work: holding the promise of doing everything from helping developers write code, search and extract insight from email and other business documents, and engage with customers. But for corporate technology leaders, the urgent need is to weigh those tremendous capabilities against the technology's downsides—the potential to expose proprietary data to competitors, disclosing customer information like email addresses, and opening up new avenues for cyberattacks.

Safety and Governance for LLMs

Deploying AI requires careful management to prevent unintentional but significant damage, not only to brand reputation but, more important, to workers, individuals, and society as a whole.

Mitigation Layers



Framework to Assess Feasibility of Large Language Models For Business

Is the potential harm within acceptable risk range?

If Unsure, Exit

Model results need to be explainable?

If Yes, Exit



Can the AI meet legal or compliance requirements, if any?

If No, Exit

Requires accurate or factually correct information?

If Yes, Exit

Are there sufficient filters and guardrails in place (e.g. human reviewers)?

If No, Exit

3

Large Language Models for Business Applications



TOPIC 3

LLMs for Business

- 01 3 Ways to Use LLMs: (UI / via APIs, customization, build your own)
- 02 Prompting LLMs, Types of Prompts
- 03 Short Course: ChatGPT Prompt Engineering
- 04 Automation and workflow for LLMs: Langchains
- 05 How can LLMs be customized / tuned for domain / proprietary use case? Large vs small models? Costs!

Types of LLMs: Open vs Closed

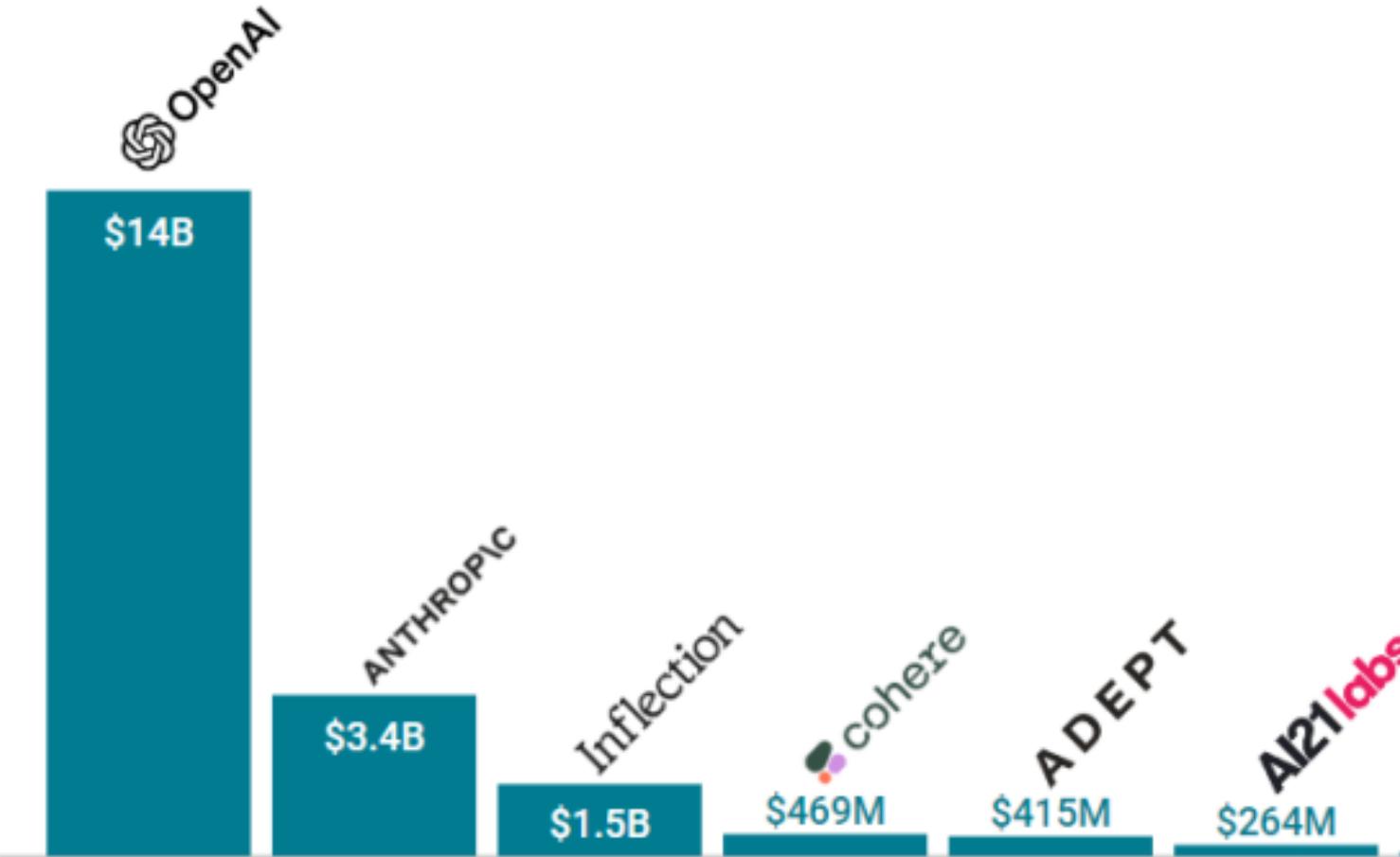


The private market is split into open vs. closed

Disclosed equity funding to LLM developers (as of 10/27/2023)

■ Private company ■ Acquired company

Closed-source LLMs



*Some developers may offer open-source versions of their models
but keep their core models proprietary

Open-source LLMs



*Excludes open-source developers that have not raised equity funding

One Model or Many?

Does One Large Model Rule Them All?

Predictions on the Future AI Ecosystem

Maithra Raghu (Samaya AI), Matei Zaharia (Databricks), Eric Schmidt (Schmidt Futures)

April 4, 2023

We believe the contrary!

- There will be many entities contributing to the advance of the AI ecosystem.
- And numerous, high-utility AI systems will emerge, distinct from (single) general AI models.
- These AI systems will be complex in structure, powered by multiple AI models, APIs, etc, and will spur new technical AI developments.
- Well defined, high-value workflows will primarily be addressed by specialized AI systems not general purpose AI models.

High-value workflows will be primarily addressed by specialized AI systems not general purpose AI models

Raghu et al., "Does One Large Model Rule Them All," 2023

"Companies with large and unique data stores will see clear advantages to training their own models as moats."

Konstantine Buhler, "Generative AI Trends," Forbes, 2023

Source: MosaicML Presentation at EmTech May 2, 2023

CLOSED / PROPRIETARY FOUNDATION MODELS

EASIER TO START, COSTLY TO SCALE

CLOSED SOURCE ALL PURPOSE MODELS

API ACCESS

LIMITED TO NO VISIBILITY TO DATA

PRIVACY CONCERNS

NO ACCESS TO WEIGHTS

NO CONTROL OVER MODELS

PROMPT ENGINEERING

PROMPT TUNING

EASIER TO START

COSTLY TO SCALE



co:here



USE APIs DIRECTLY OR WITH LANGCHAIN FOR MORE CONTROL

CURATED ENTERPRISE MODELS (GENERAL / DOMAIN SPECIFIC)

API ACCESS (ON-PREM possible)
→ VETTED DATA, SECURITY

LIMITED CONTROL OVER MODELS

COST VARIES



Generative AI on
Google Cloud

OPEN SOURCE FOUNDATION MODELS

MANY MODELS AVAILABLE, QUALITY VARIES

OPEN SOURCE MODELS
PROMPT ENGINEERING & TUNING

ACCESS TO CODE & MODEL WEIGHTS
VISIBILITY INTO DATA (WORK NEEDED)
FULLY CUSTOMIZABLE
LICENSE TERMS VARY



GOOGLE FLAN-T5



FINE-TUNE OPEN SOURCE MODELS
FOR ENTERPRISE REQUIREMENTS

ACCESS TO CODE & MODEL WEIGHTS
TUNE WITH PRIVATE DATA
ADAPT FOR USE CASE

REQUIRES TRAINING EXPERTISE
LIFECYCLE MANAGEMENT

co:here



MORE CONTROL BUT REQUIRES MOST WORK

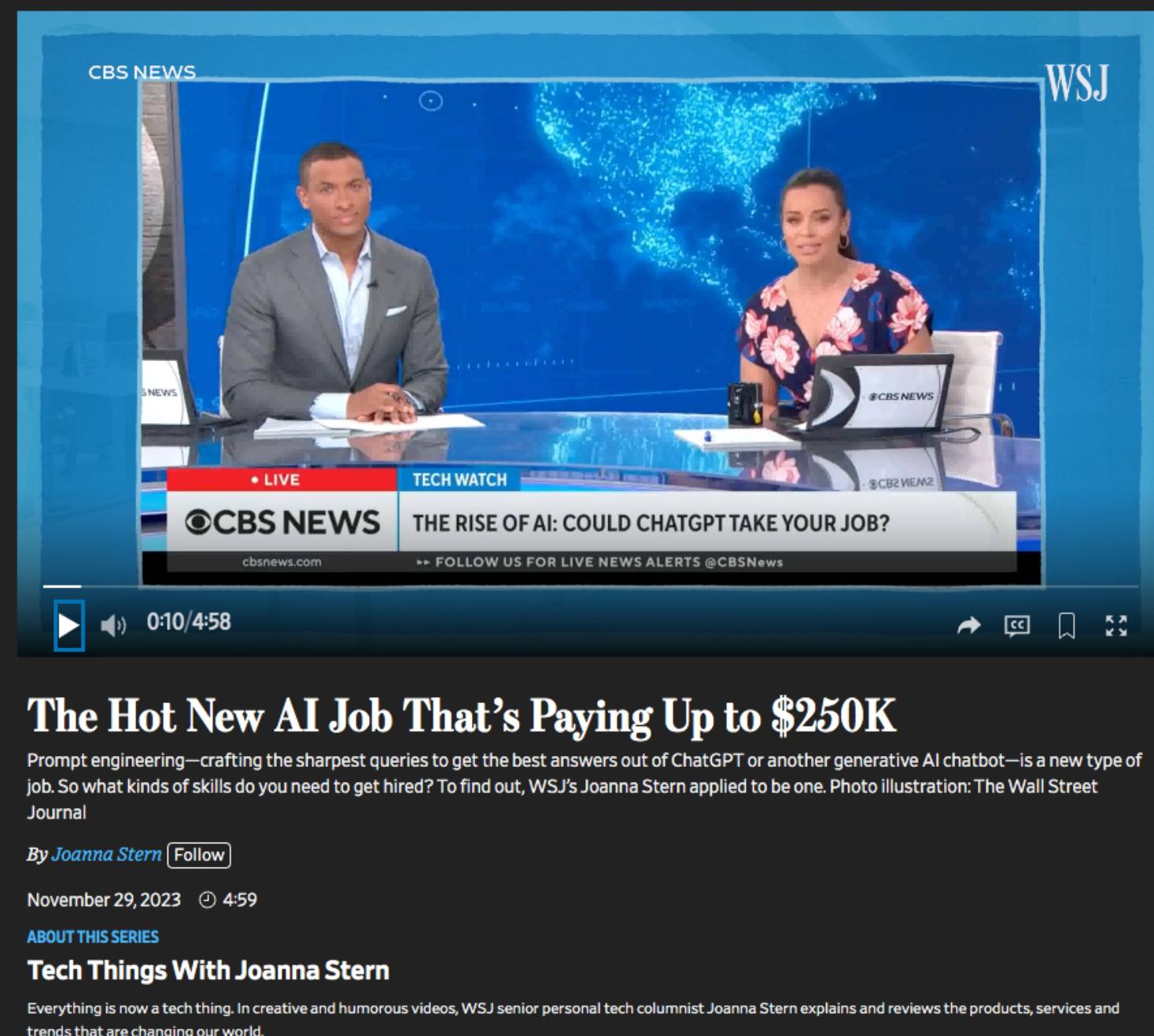
@thebabar 2023



4

Using and Adapting LLMs

Prompt Engineering



The Hot New AI Job That's Paying Up to \$250K

Prompt engineering—crafting the sharpest queries to get the best answers out of ChatGPT or another generative AI chatbot—is a new type of job. So what kinds of skills do you need to get hired? To find out, WSJ's Joanna Stern applied to be one. Photo illustration: The Wall Street Journal

By [Joanna Stern](#) [Follow](#)

November 29, 2023 ④ 4:59

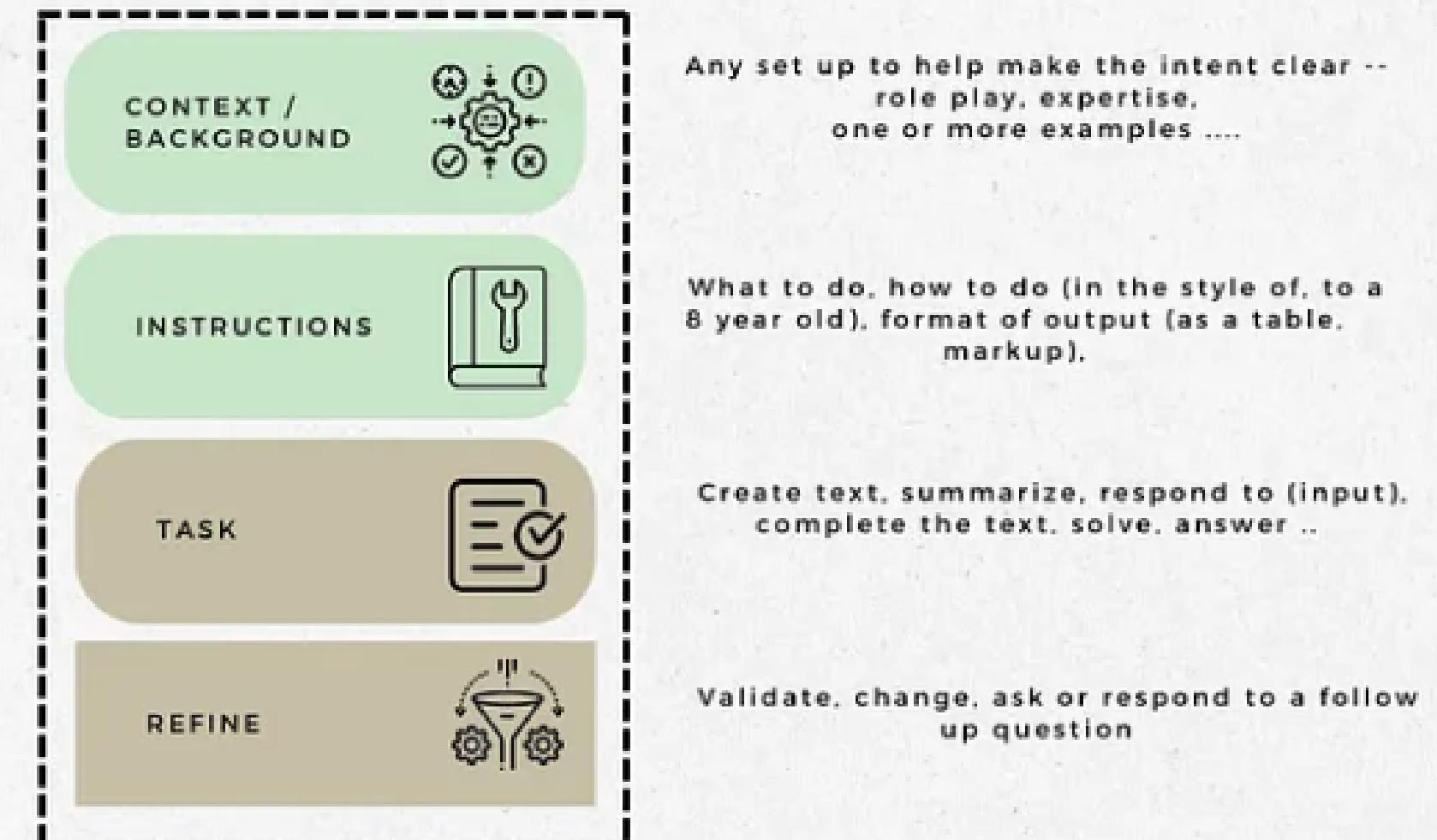
[ABOUT THIS SERIES](#)

Tech Things With Joanna Stern

Everything is now a tech thing. In creative and humorous videos, WSJ senior personal tech columnist Joanna Stern explains and reviews the products, services and trends that are changing our world.

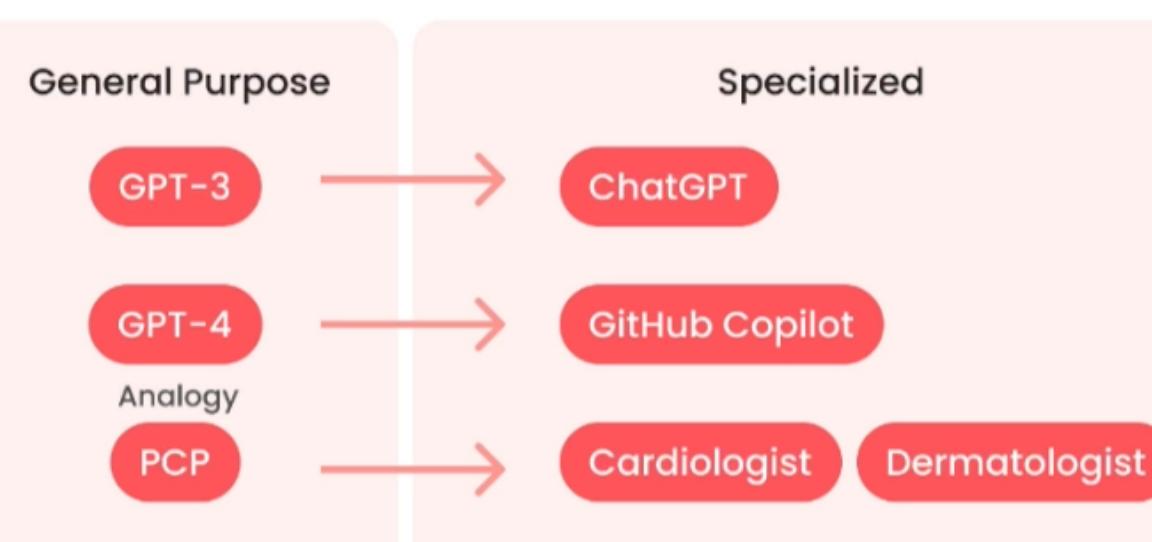
PROMPT CRAFTING

ANATOMY OF A PROMPT



Anatomy of a Prompt — Illustration by [Babar M Bhatti](#)

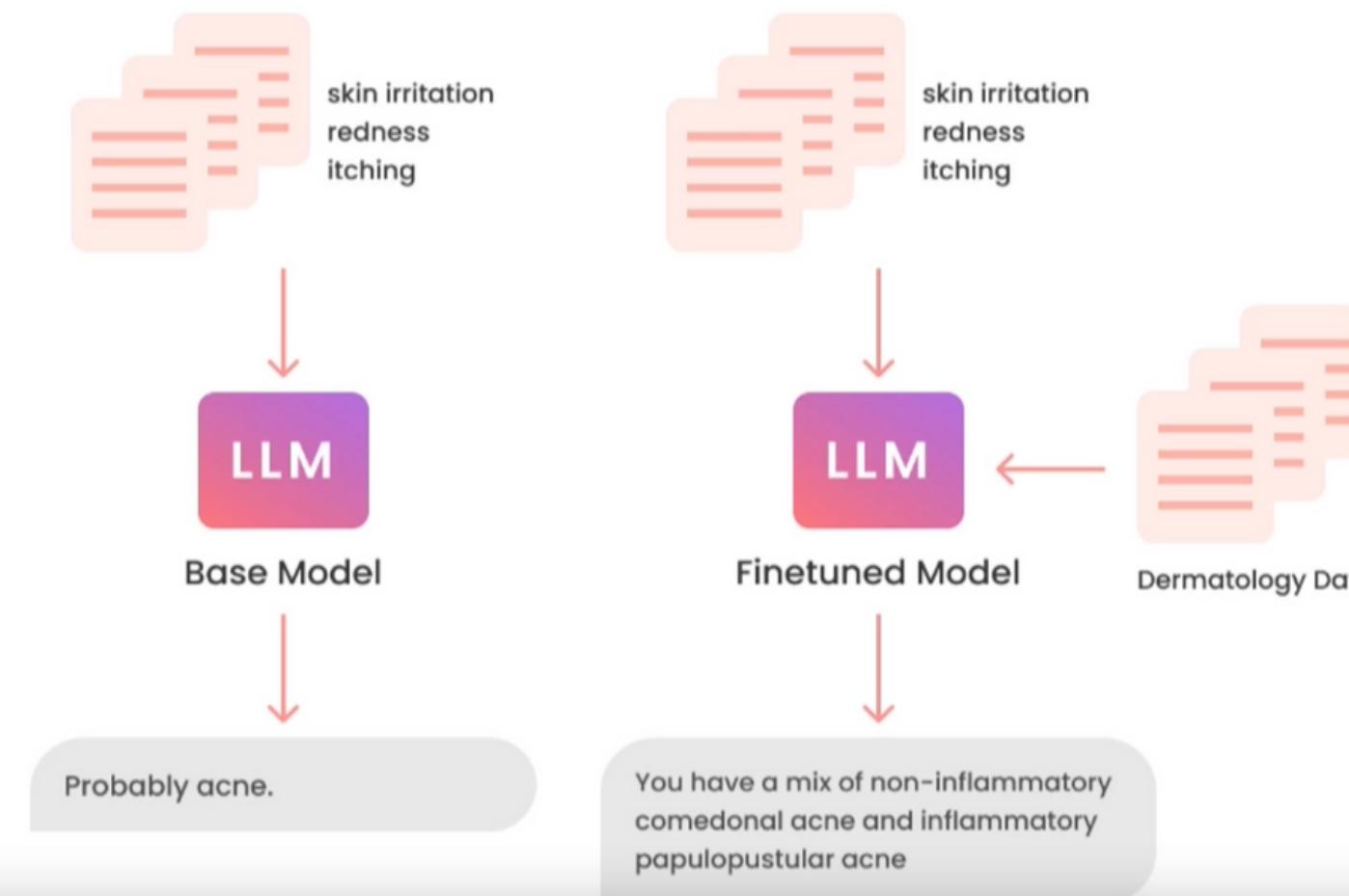
But first: what is finetuning?



- Steers the model to more consistent outputs
- Reduces hallucinations
- Customizes the model to a specific use case
- Process is similar to the model's earlier training

What does finetuning do for the model?

- Lets you put *more* data into the model than what fits into the prompt
- Gets the model to *learn* the data, rather than just get access to it



Parameter Efficient Fine-Tuning

PEFT

😊 PEFT (Parameter-Efficient Fine-Tuning) is a library for efficiently adapting large pretrained models to various downstream applications without fine-tuning all of a model's parameters because it is prohibitively costly. PEFT methods only fine-tune a small number of (extra) model parameters - significantly decreasing computational and storage costs - while yielding performance comparable to a fully fine-tuned model. This makes it more accessible to train and store large language models (LLMs) on consumer hardware.

PEFT is integrated with the Transformers, Diffusers, and Accelerate libraries to provide a faster and easier way to load, train, and use large models for inference.

Get started

Start here if you're new to 😊 PEFT to get an overview of the library's main features, and how to train a model with a PEFT method.

How-to guides

Practical guides demonstrating how to apply various PEFT methods across different types of tasks like image classification, causal language modeling, automatic speech recognition, and more. Learn how to use 😊 PEFT with the DeepSpeed and Fully Sharded Data Parallel scripts.

Conceptual guides

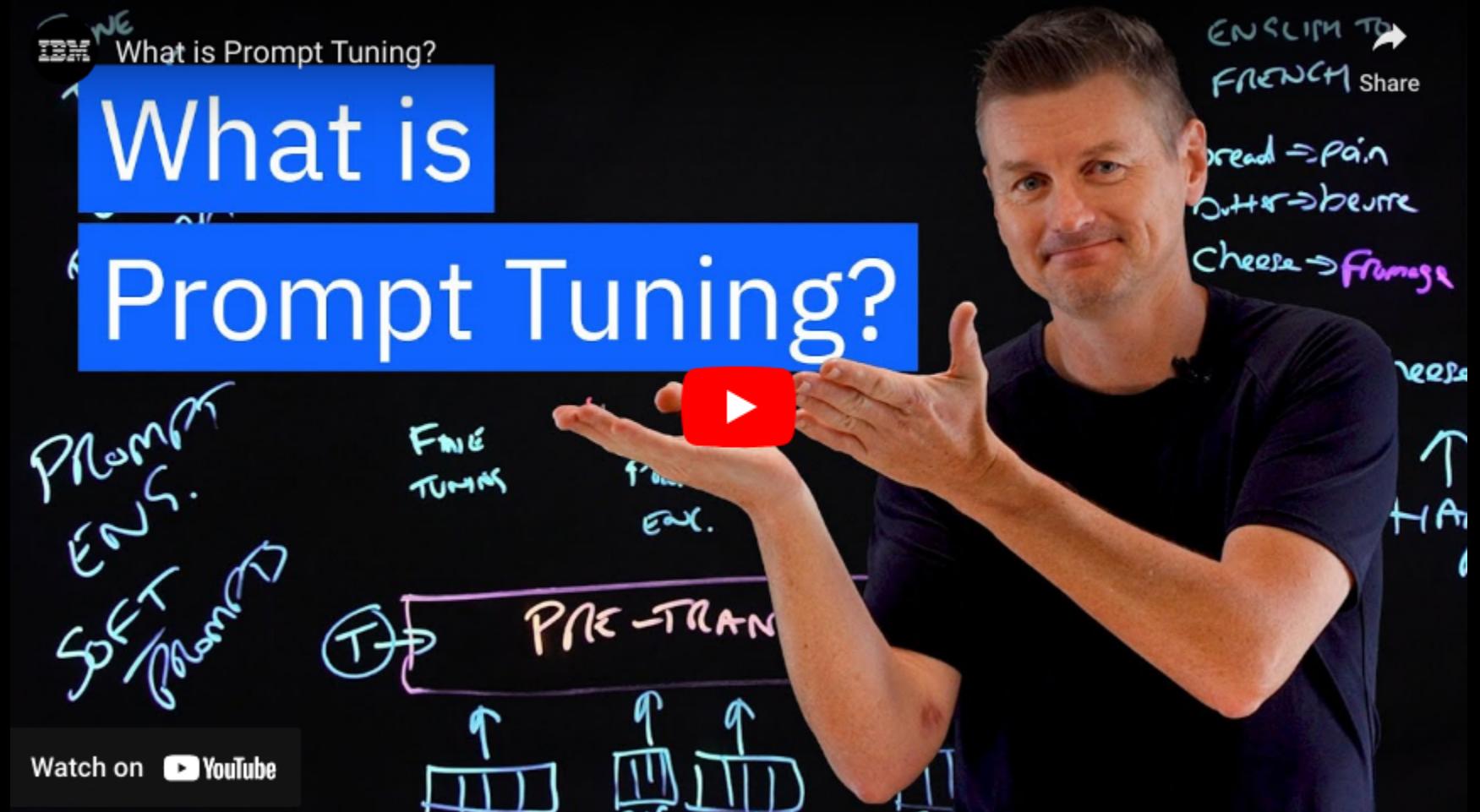
Get a better theoretical understanding of how LoRA and various soft prompting methods help reduce the number

Reference

Technical descriptions of how 😊 PEFT classes and methods work

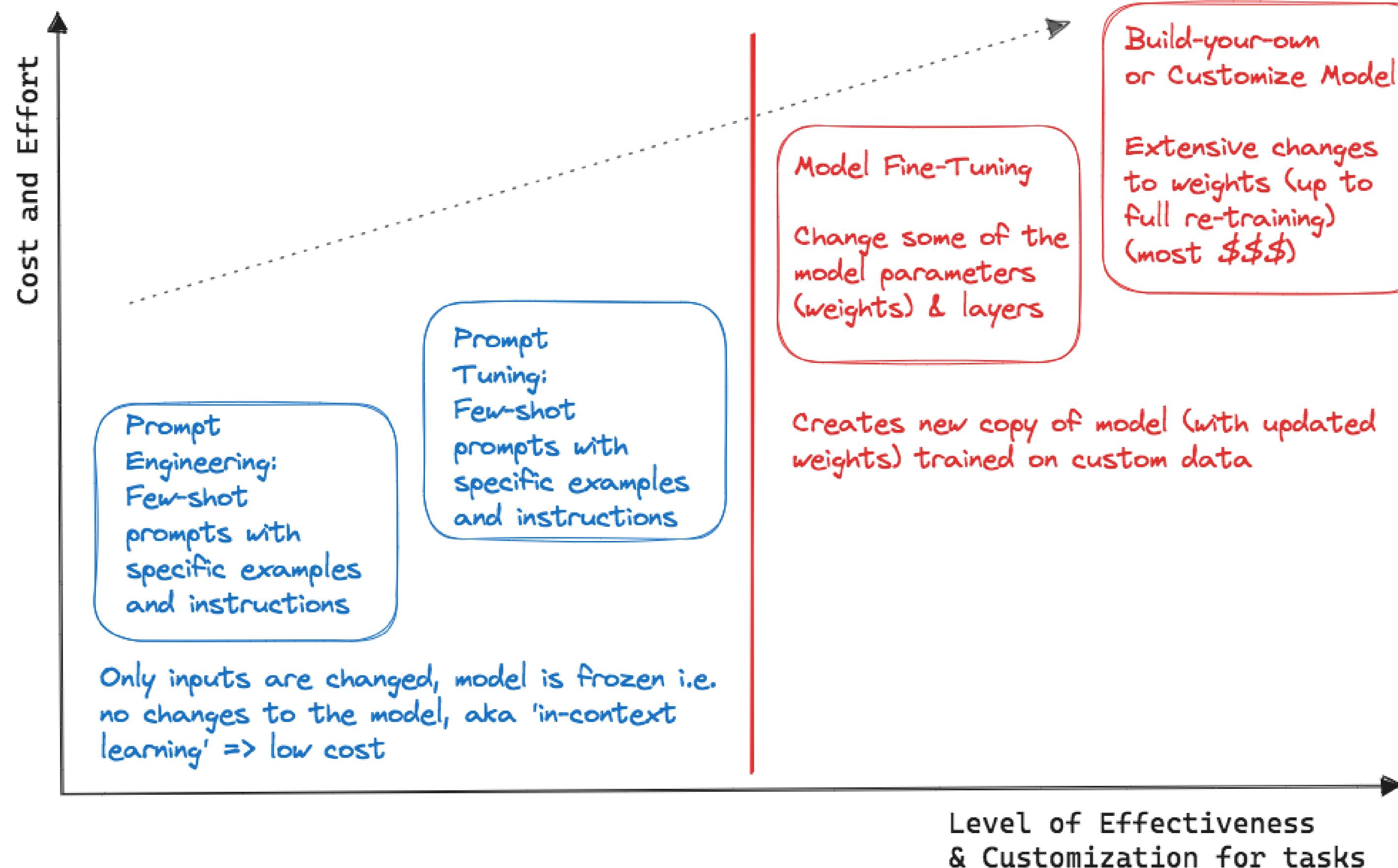
Prompt Engineering vs. Finetuning

	Prompting	Finetuning
Pros	<ul style="list-style-type: none">• No data to get started• Smaller upfront cost• No technical knowledge needed• Connect data through retrieval (RAG)	<ul style="list-style-type: none">• Nearly unlimited data fits• Learn new information• Correct incorrect information• Less cost afterwards if smaller model• Use RAG too
Cons	<ul style="list-style-type: none">• Much less data fits• Forgets data• Hallucinations• RAG misses, or gets incorrect data	<ul style="list-style-type: none">• More high-quality data• Upfront compute cost• Needs some technical knowledge, esp. data
	Generic, side projects, prototypes	Domain-specific, enterprise, production usage, ...privacy!



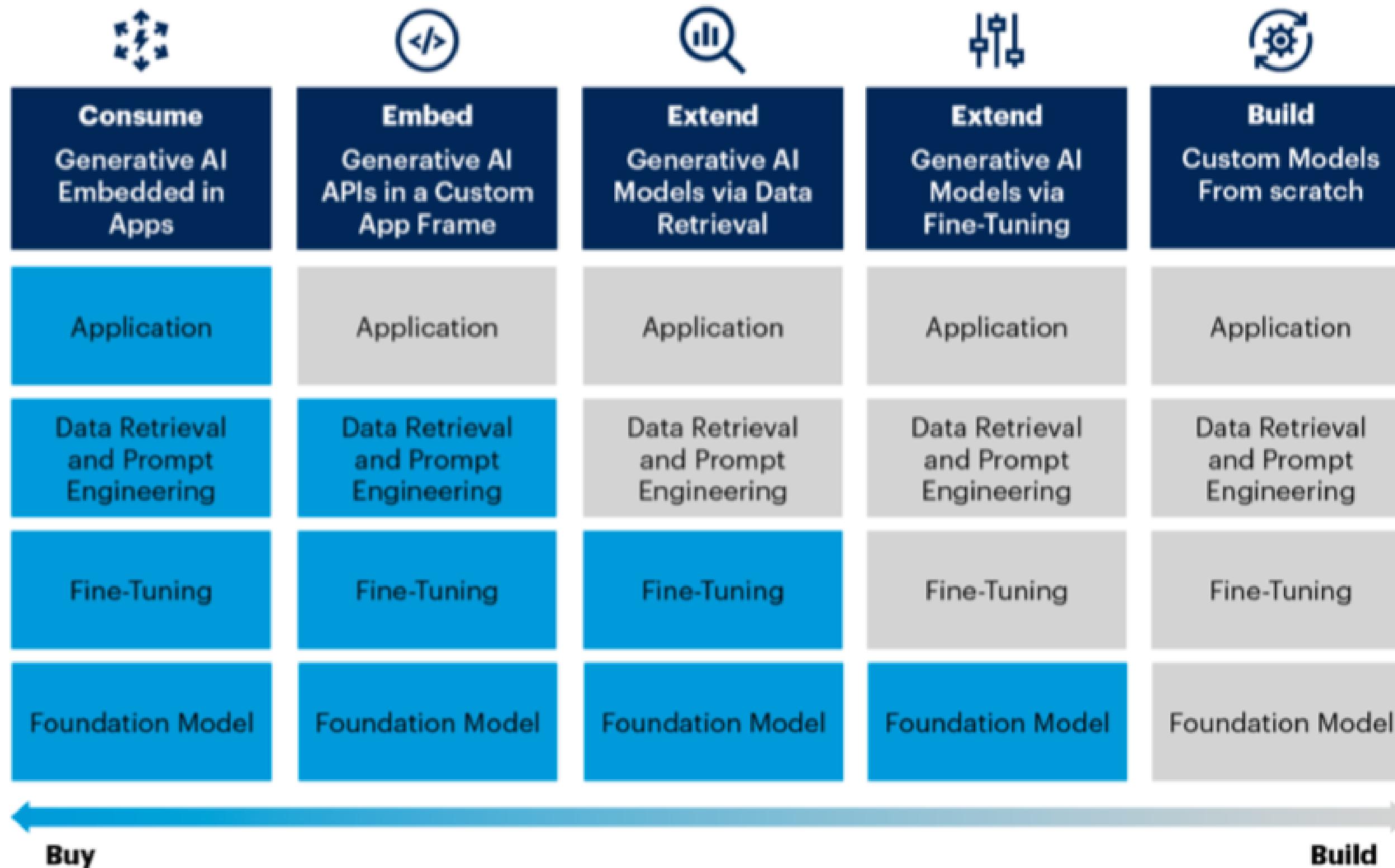
Prompt tuning is an efficient, low-cost way of adapting an AI foundation model to new downstream tasks without retraining the model and updating its weights.

Adapting/Tuning Foundation Models for Business/Private Use Cases



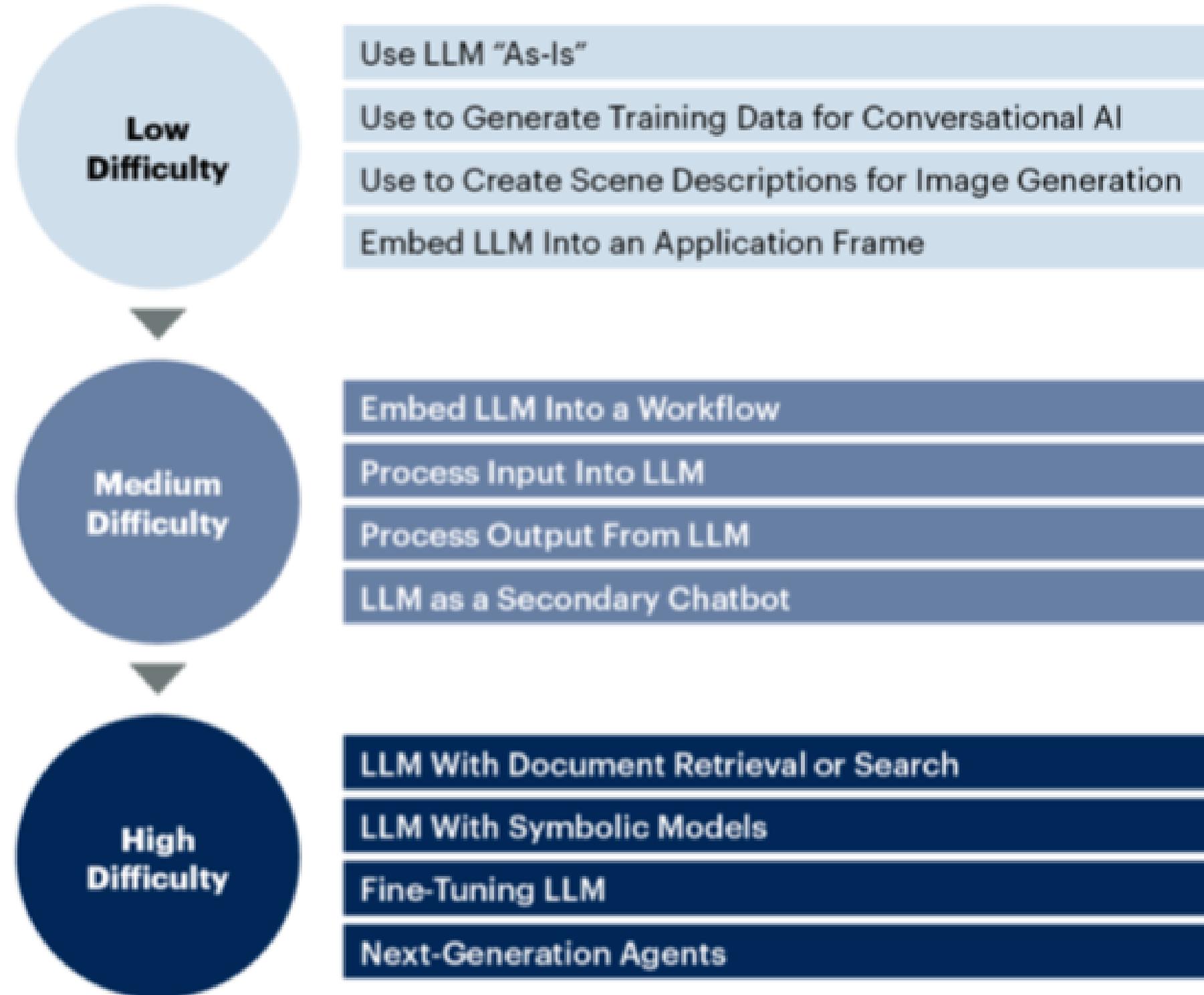
Generative AI Deployment Approaches

■ Provider-Managed ■ Self-Managed



Source: Gartner
794559_C

Design Patterns for Large Language Models



5

Tools and Trends



TRENDS

Building LLM Applications

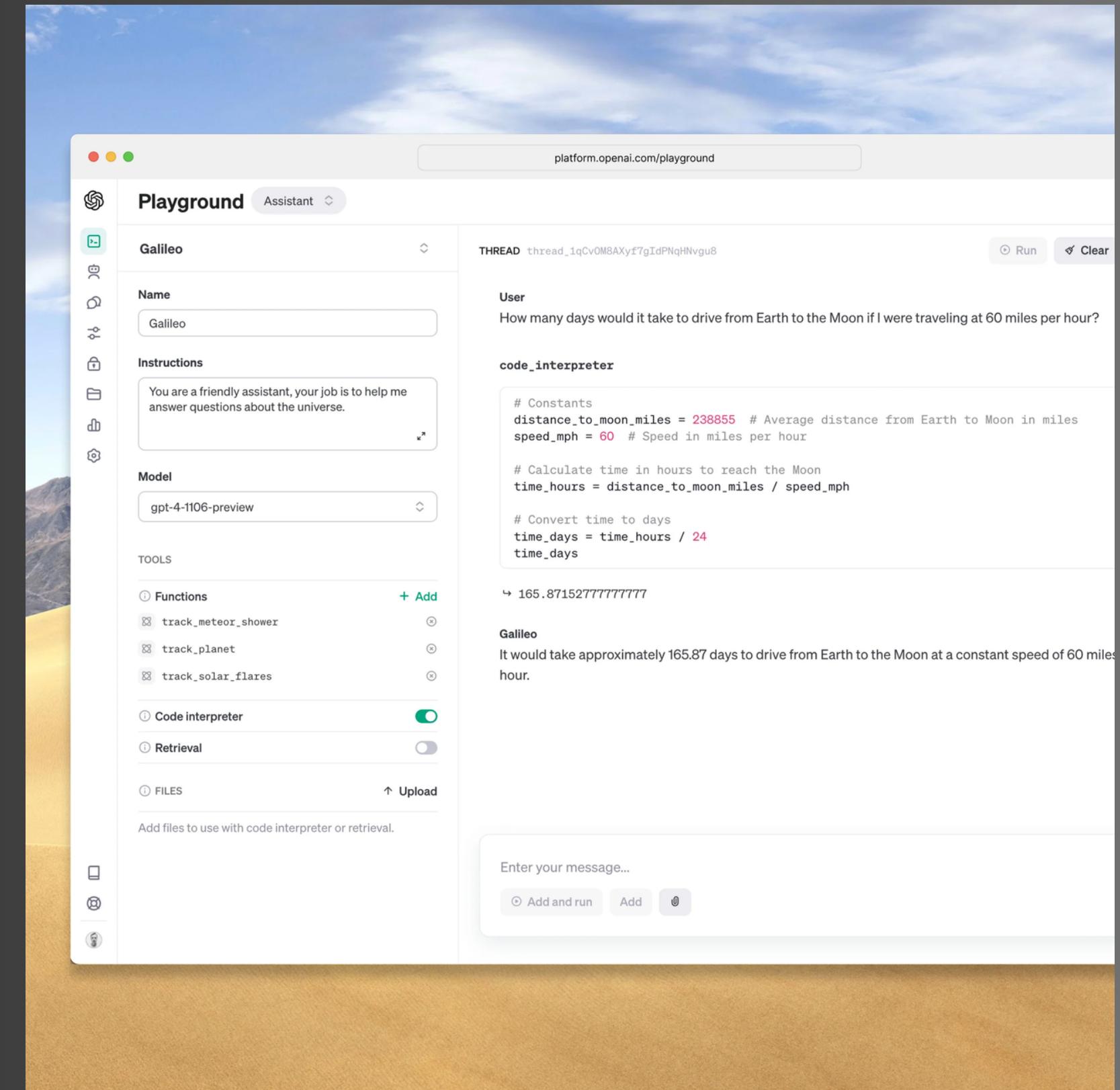
01	Are you a consumer or a creator?
02	Vector Databases
03	Frameworks
04	Open Source
05	Low-code Offerings (e.g. GPT Builder, Assistants)
06	Governance

Are you a Consumer OR Creator?

Three Choices for building LLM Applications:

1. Commercial vendor eg OpenAI
2. Use framework e.g. langchain
3. Your own

[Sample discussion on Reddit](#)



The screenshot shows the OpenAI Playground interface. On the left, there's a sidebar with icons for file management, a code interpreter, and retrieval. The main area is titled "Playground" and shows a session named "Galileo". The "Instructions" section contains the text: "You are a friendly assistant, your job is to help me answer questions about the universe." The "Model" section is set to "gpt-4-1106-preview". The "TOOLS" section includes "Functions" (track_meteor_shower, track_planet, track_solar_flares), "Code interpreter" (which is turned on), and "Retrieval". The "FILES" section has an "Upload" button. In the center, a "THREAD" window is open with the URL "platform.openai.com/playground". It shows a user query: "How many days would it take to drive from Earth to the Moon if I were traveling at 60 miles per hour?". Below the query is a "code_interpreter" block with the following Python code:

```
# Constants
distance_to_moon_miles = 238855 # Average distance from Earth to Moon in miles
speed_mph = 60 # Speed in miles per hour

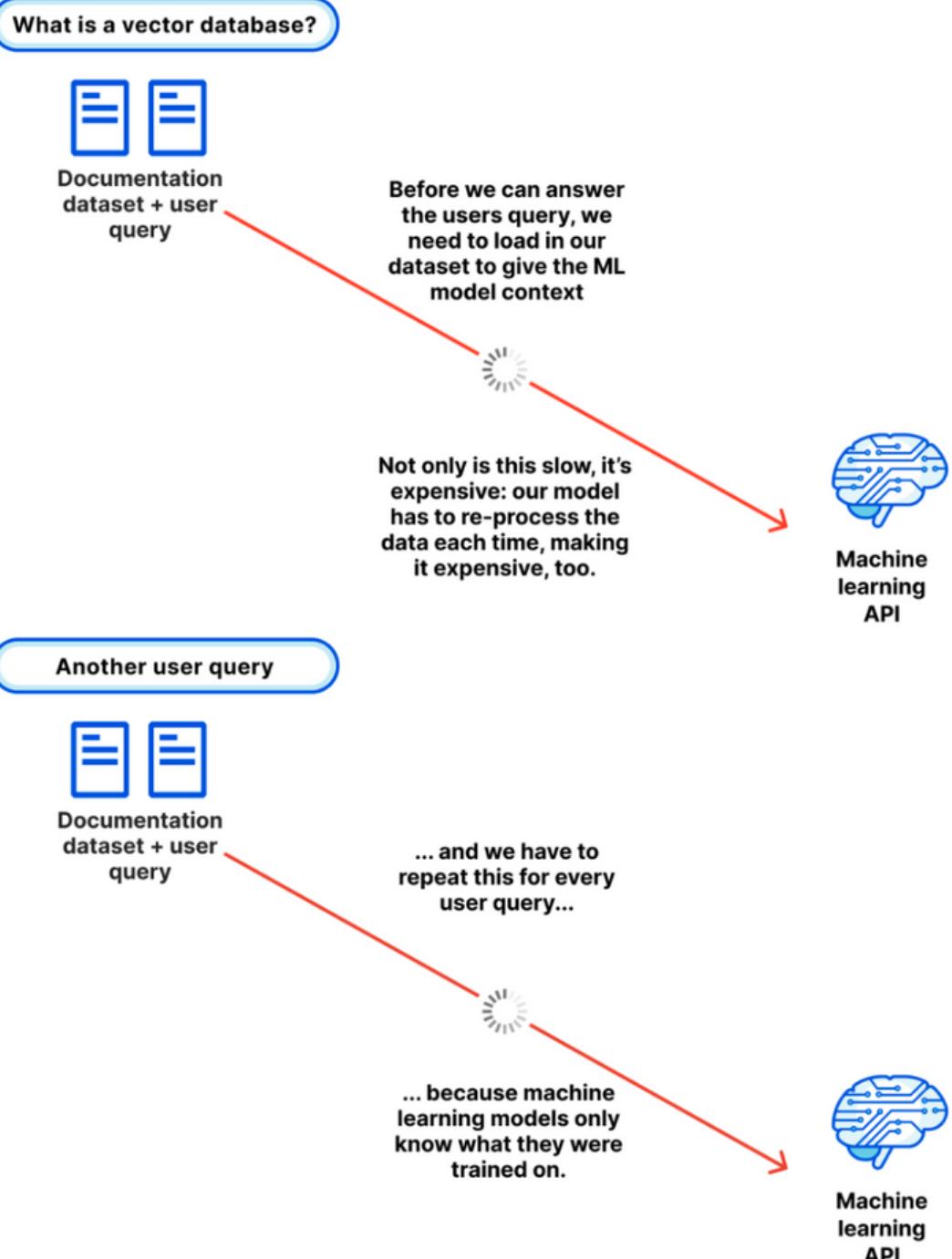
# Calculate time in hours to reach the Moon
time_hours = distance_to_moon_miles / speed_mph

# Convert time to days
time_days = time_hours / 24
time_days
```

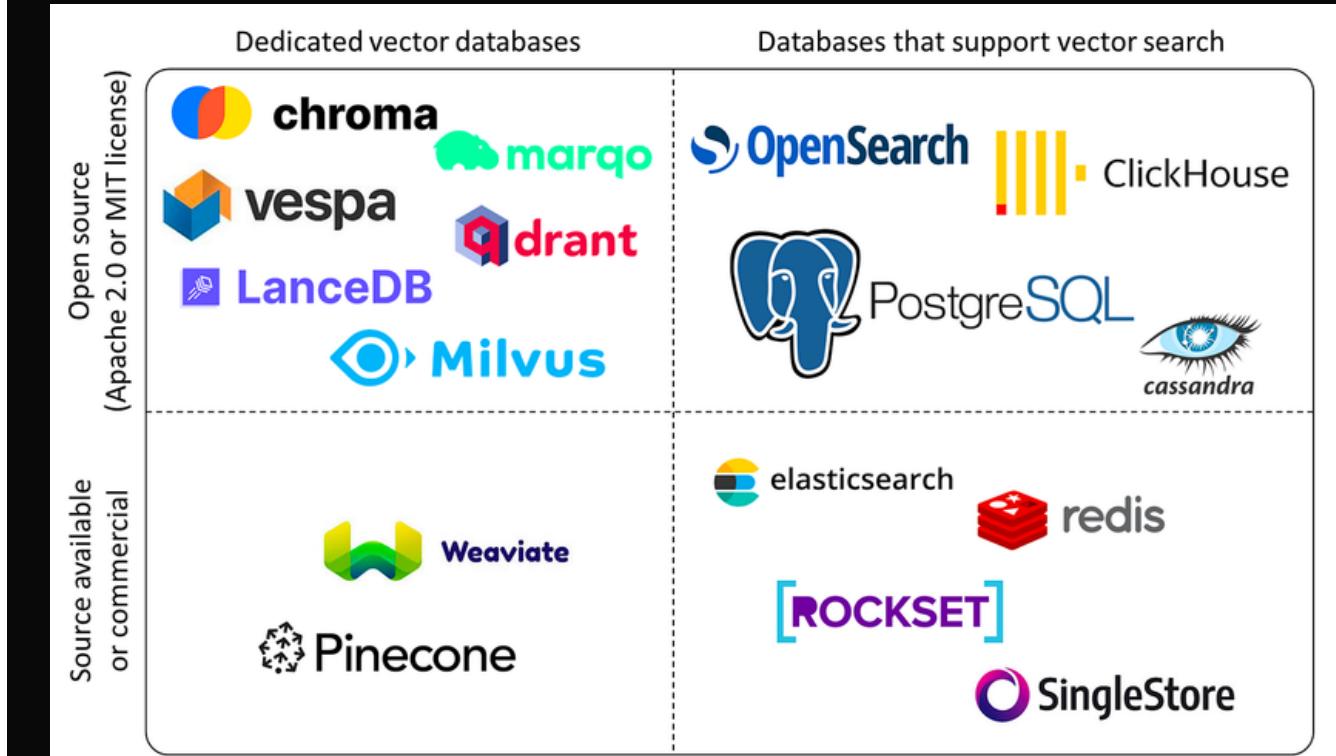
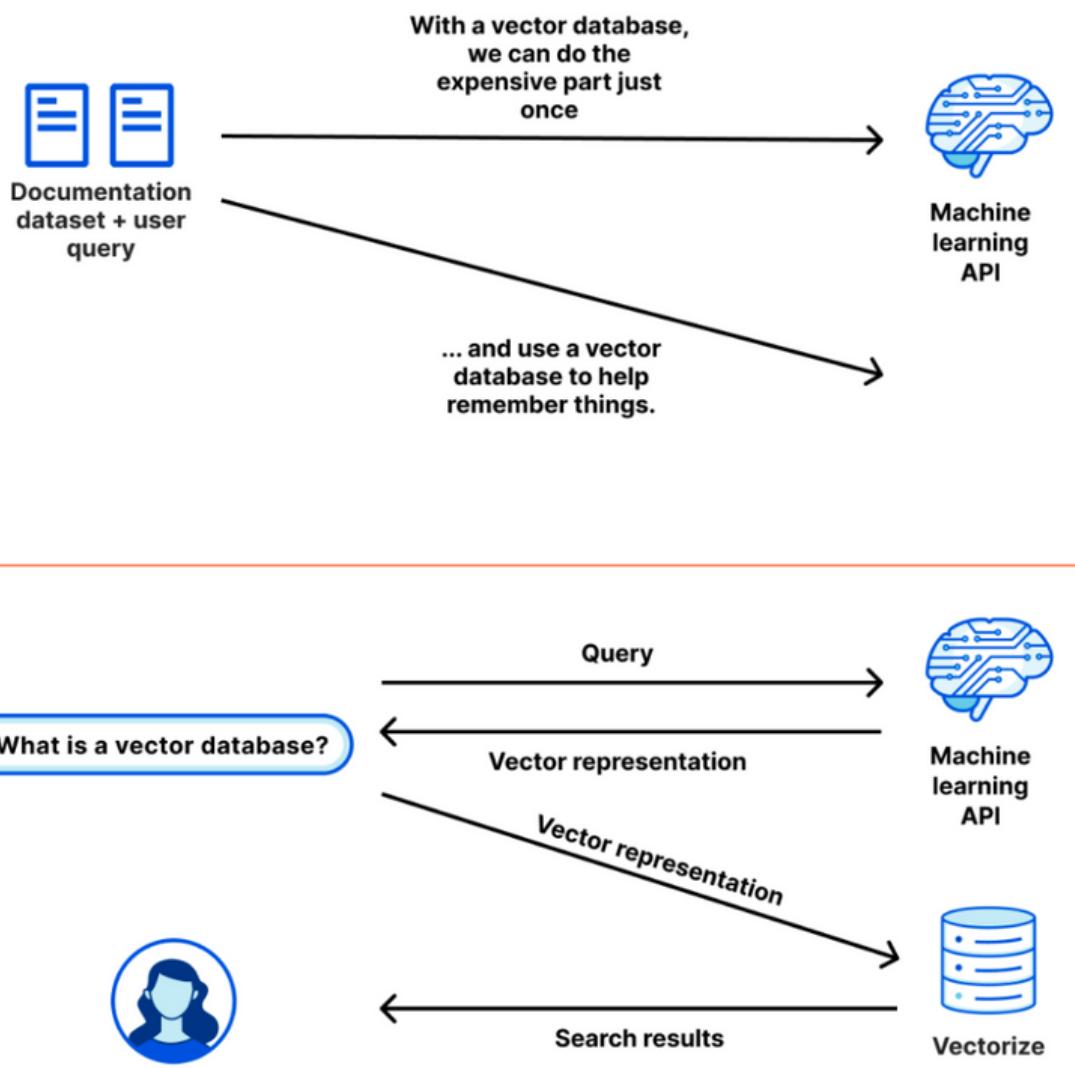
The response below the code is: "It would take approximately 165.87 days to drive from Earth to the Moon at a constant speed of 60 miles hour." At the bottom of the thread, there's an input field labeled "Enter your message..." and buttons for "Add and run", "Add", and a count of "0".

Vector Databases

No vector database



Vector database



6

Retrieval Augmented Generation

Chat with your Data

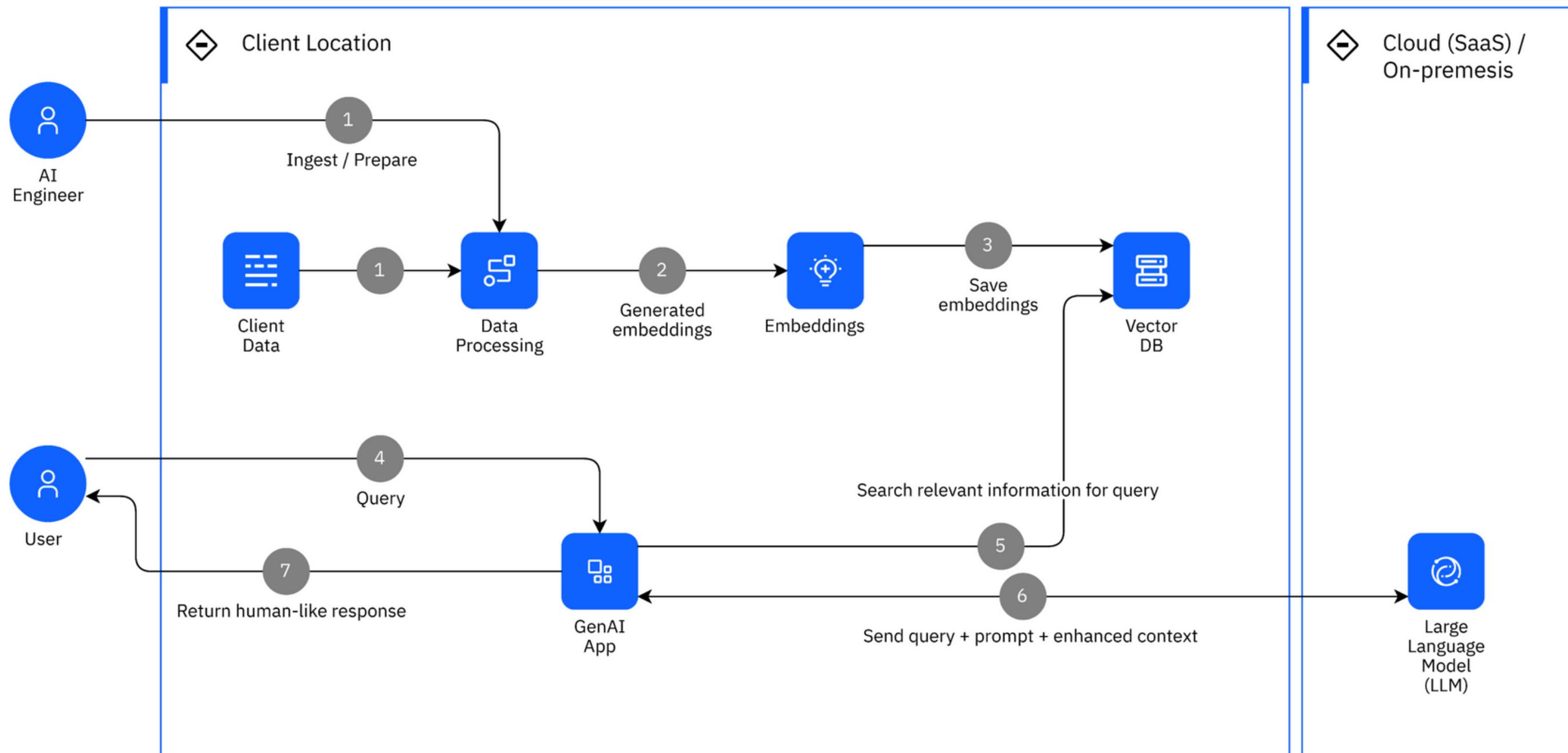
(Use LLMs for Enterprise Language Use Cases)

R(retrieval) A(ugmented) G(eneration)

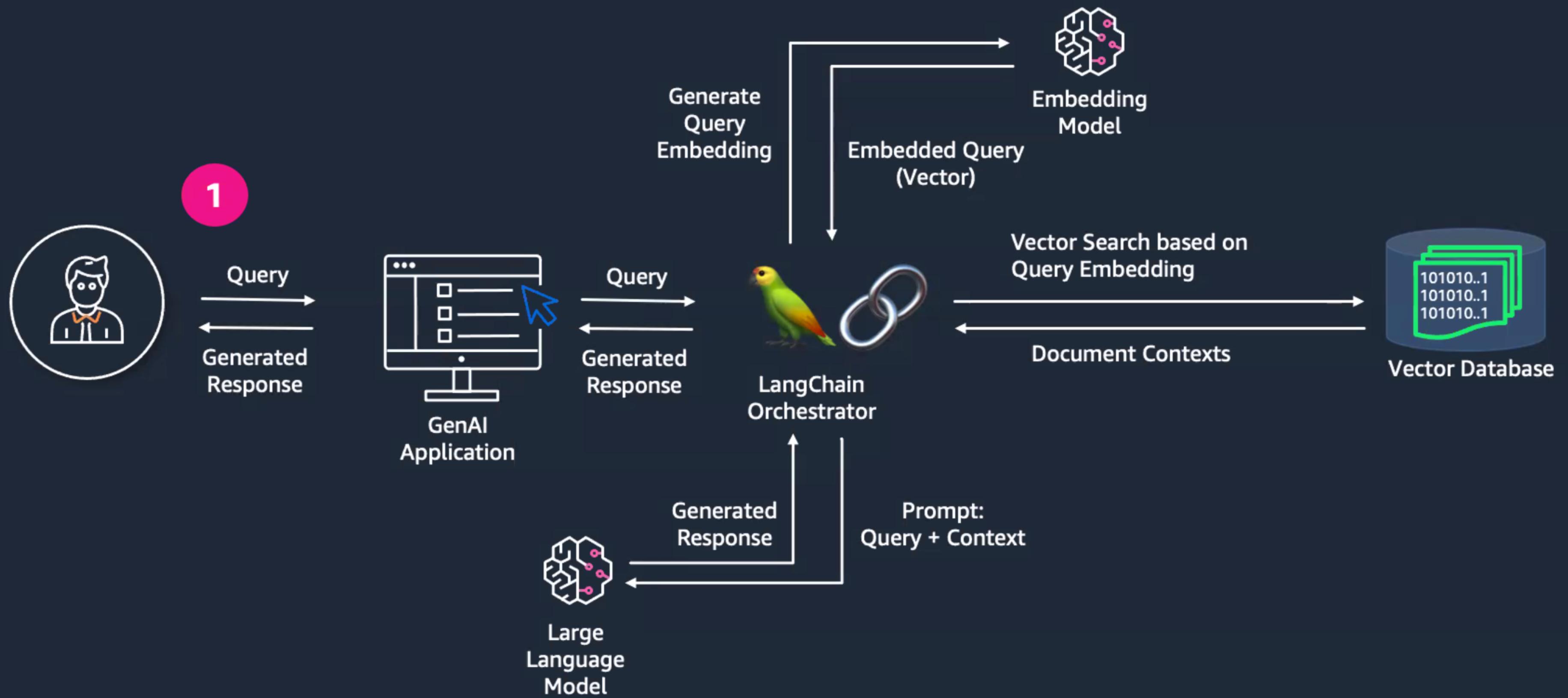
RAG has 3 Key Components:

1. Combines NL Understanding and NL Generation
2. Comprehends, Retrieves, Generates
3. Uses vector databases for embeddings

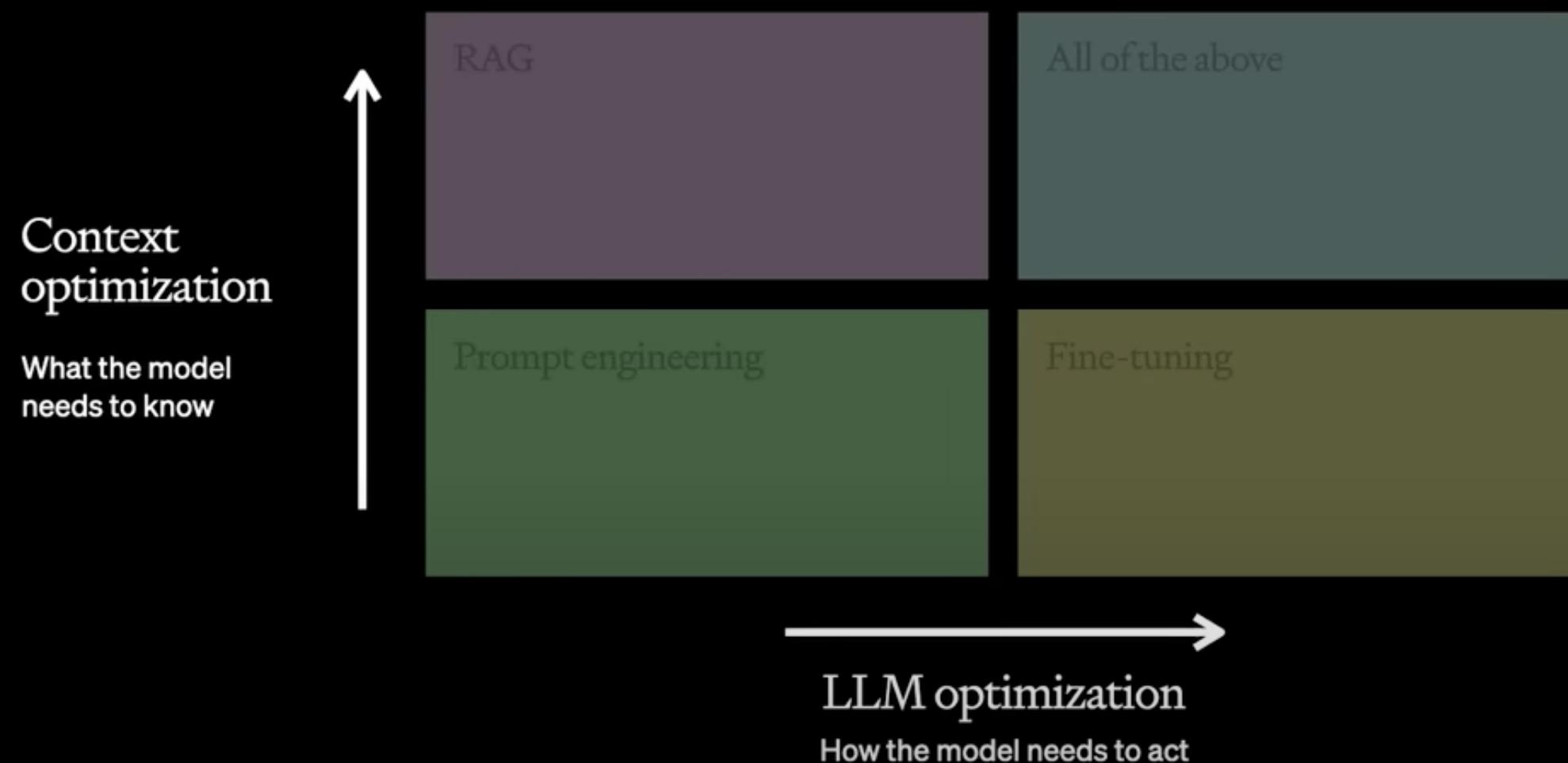
RAG Conceptual Architecture (IBM)



Semantic Search / RAG



The optimization flow



When to use RAG

	RAG	Finetuning
External knowledge req'd?	✓	✗
Changing model behaviour req'd?	✗	✓
Minimise hallucinations?	✓	✗
Training data available?	✗	✓
Is data (mostly) dynamic?	✓	✗
Interpretability req'd?	✓	✗

When to use RAG

Getting in Touch

AI+babar.bhatti@dallas-ai.org
[linkedin.com/in/bbhatti](https://www.linkedin.com/in/bbhatti)