

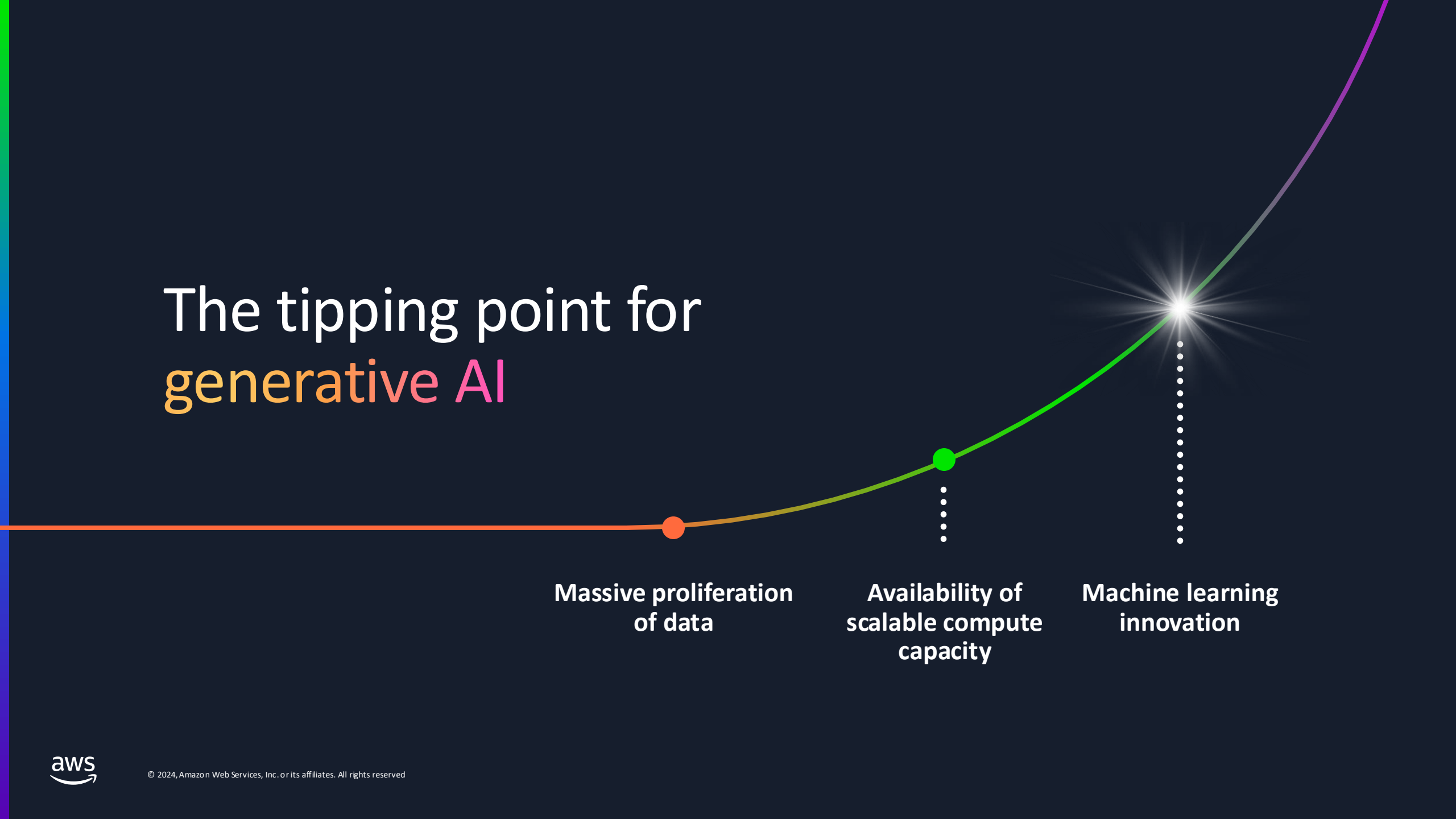


# Amazon Bedrock

**Michael Lin**

Sr. Solutions Architect  
Amazon Web Services

# The tipping point for generative AI

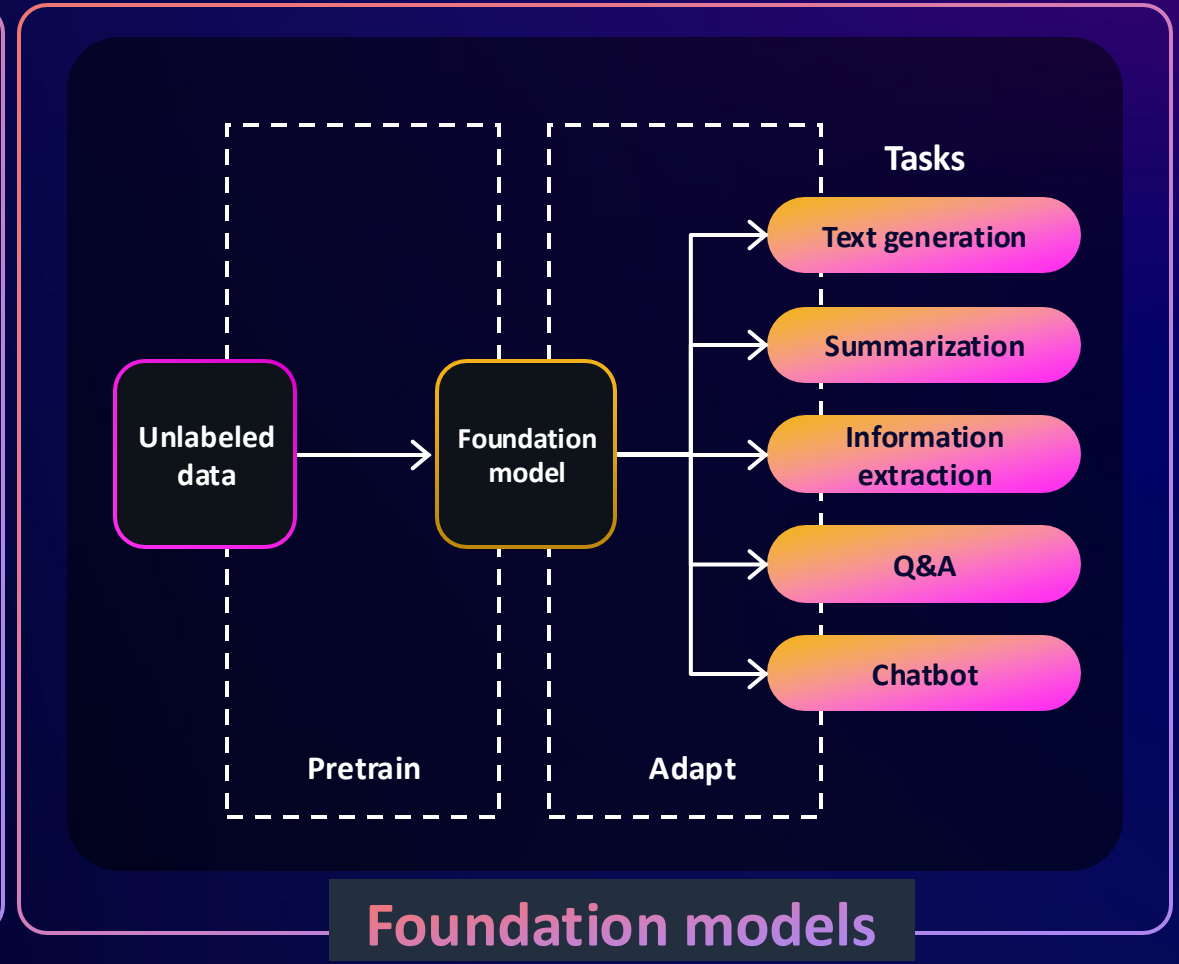
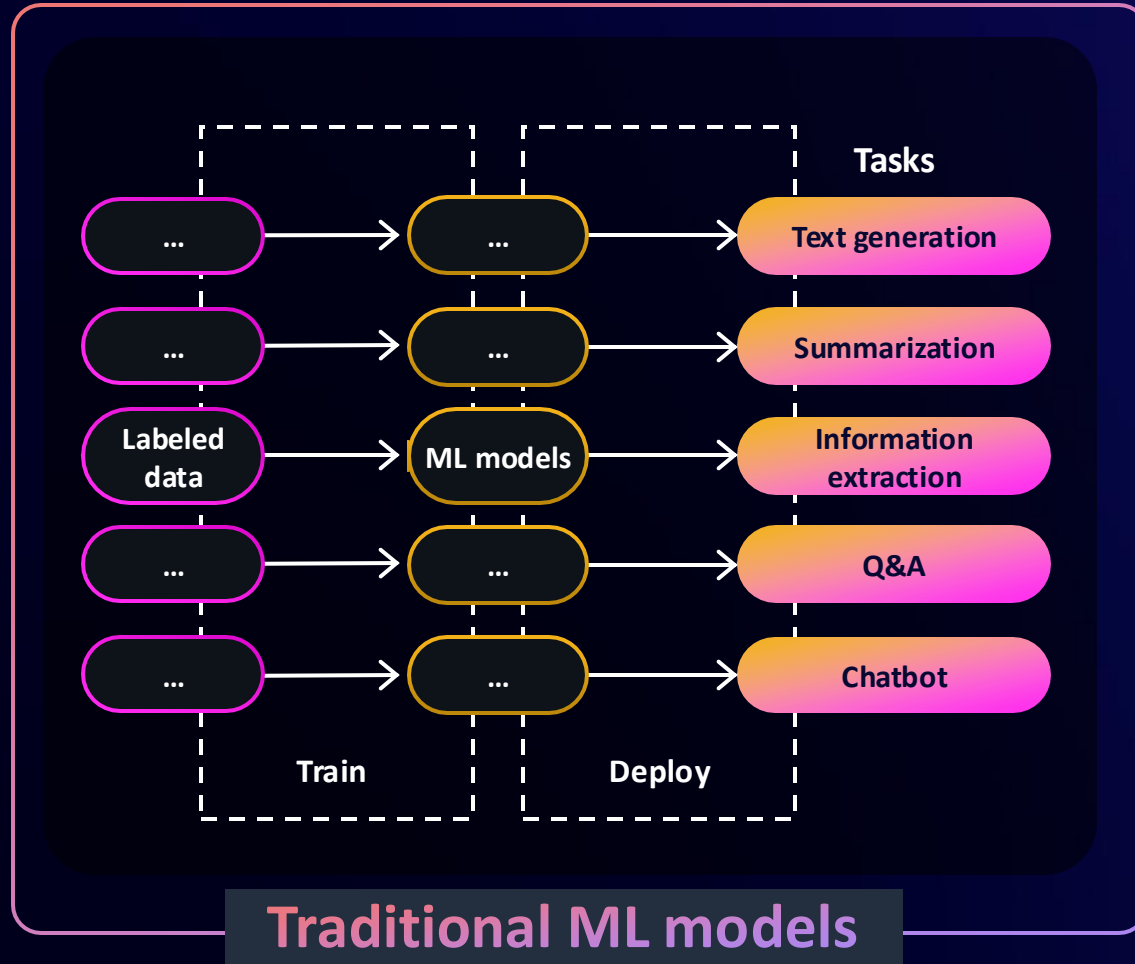


Massive proliferation  
of data

Availability of  
scalable compute  
capacity

Machine learning  
innovation

# Generative AI is powered by foundation models (FMs)





## Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and safety

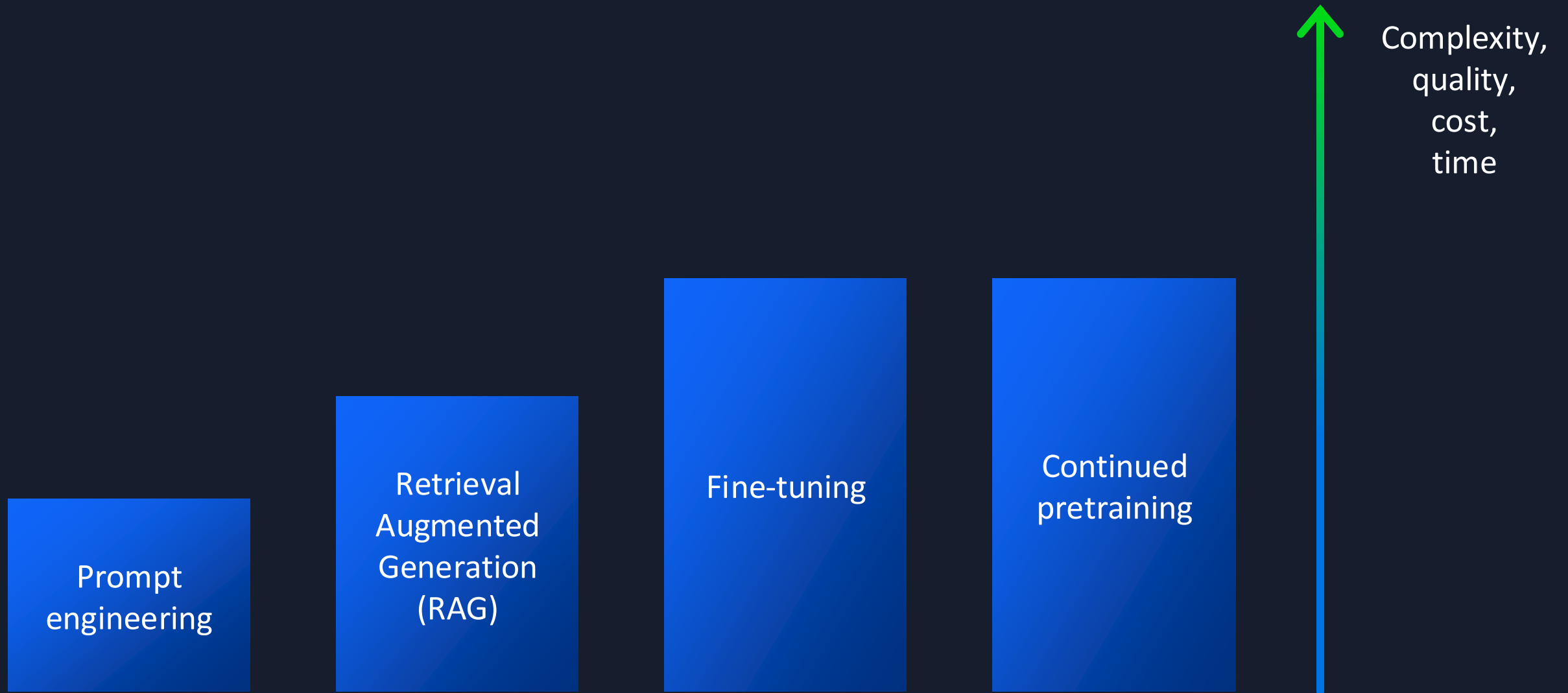
# Amazon Bedrock

## BROAD CHOICE OF MODELS

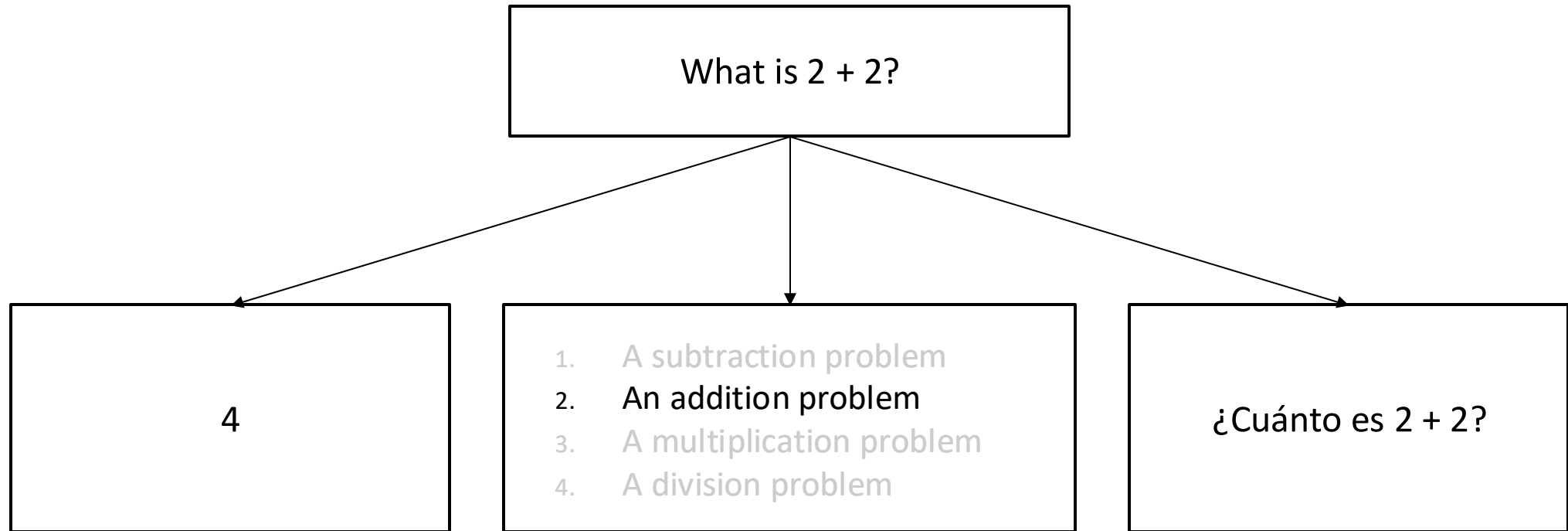
AI21labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI	stability.ai
Contextual answers, summarization, paraphrasing	Text summarization, generation, Q&A, search, image generation	Summarization, complex reasoning, writing, coding	Text generation, search, classification	Q&A and reading comprehension	Text summarization, text classification, text completion, code generation, Q&A	High-quality images and art
Jamba-Instruct	Amazon Titan Text Premier	Claude 3.5 Sonnet	Command	Llama 3 8B	Mistral Small	Stable Diffusion XL1.0
Jurassic-2 Ultra	Amazon Titan Text Lite	Claude 3 Opus	Command Light	Llama 3 70B	Mistral Large	Stable Diffusion XL 0.8
Jurassic-2 Mid	Amazon Titan Text Express	Claude 3 Sonnet	Embed English	Llama 2 13B	Mistral 7B	
	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed Multilingual	Llama 2 70B	Mixtral 8x7B	
	Amazon Titan Text Embeddings V2	Claude 2.1	Command R+			
	Amazon Titan Multimodal Embeddings	Claude 2	Command R			
	Amazon Titan Image Generator	Claude Instant				



# Common approaches for customizing FMs



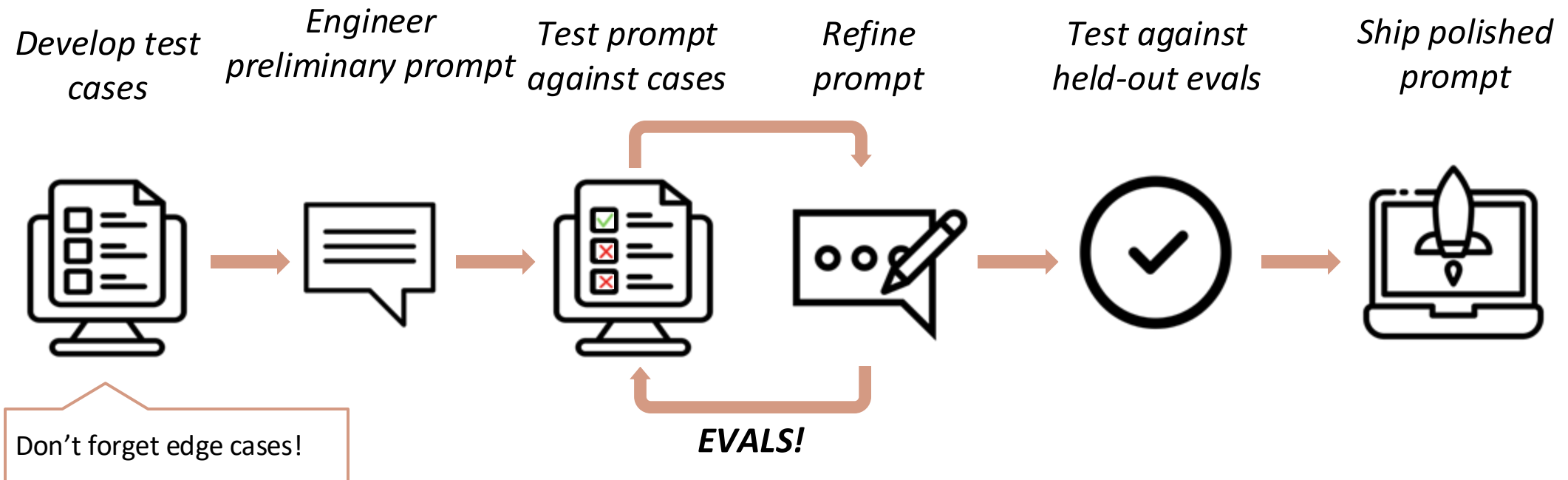
# What is prompt engineering?



**Prompt engineering** is the process of **controlling model behavior** by **optimizing your prompt to elicit high performing LLM responses** (as assessed by rigorous evaluations tailored to your use case).

# How to engineer a good prompt

**Empirical science:** always test your prompts & iterate often!







Amazon Bedrock

▼ Getting started

Overview

Examples

Providers

▼ Foundation models

Base models

Custom models

Imported models [Preview](#)

▼ Playgrounds

Chat

**Text**

Image

▼ Safeguards

Guardrails

Watermark detection

▼ Builder tools

Knowledge bases

Agents

Prompt management [Preview](#)

Prompt flows [Preview](#)

▼ Assessment & deployment

Amazon Bedrock > Text playground

## Text playground [Info](#)

Load examples

**A** Claude 3 Sonnet v1 | [ODT](#)  
[Change](#)

在日常聊天中出現的字詞。  
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。  
4.標點符號必須使用全形。例如，不可以使用小寫的“，”，必須使用“，”。  
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。

<example>  
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI applications.  
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。  
</example>

<text>  
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the thousands of you wonderful people who get value out of all of my courses, prints, presets and ebooks over at patk.com."  
</text>

<result>「我們稍後會涵蓋所有這些內容，但在開始之前，這段影片並沒有贊助商，不過它獲得來自成千上萬位從我在 patk.com 上的各種課程、美工圖案、預設和電子書中獲益良多的出色觀眾支持。」</result>

▶ Run



### Configurations

Reset

#### ▼ Randomness and diversity [Info](#)

Temperature

1

Top P

0.999

Top K

250

#### ▼ Length [Info](#)

Maximum length

2000

Stop sequences

Add

Human: ✕

#### ▼ Guardrail

# Knowledge Bases for Amazon Bedrock

## NATIVE SUPPORT FOR RAG

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multiturn conversations

Automatic citations with retrievals to improve transparency

USER  
QUERY

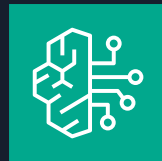
AUGMENTED  
PROMPT

MODEL

ANSWER



1



AMAZON  
BEDROCK

2

KNOWLEDGE  
BASES FOR  
AMAZON BEDROCK

3

4



5

AI21 Labs—Jurassic2

Amazon Titan Text

Anthropic—Claude

Meta—Llama

6



# Bedrock Knowledge Base in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

TestDelete

Knowledge base overview

Edit

Knowledge base name

knowledge-base-cwa

Knowledge base description

—

Service Role

AmazonBedrockExecutionRoleForKnowledgeBase\_cwa

Knowledge base ID

CJOMCDOKNC

Status

Ready

Created date

February 29, 2024, 20:13 (UTC+08:00)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
-----	-------

111年度預算執行情形。

111年度本署歲入預算共編列2,897萬元,決算數為2,929萬4千元,決算數占預算數101.12%。111年度本署歲出預算加計動支第一、二預備金共21億5,114萬5千元,決算數為21億4,675萬元,決算數占預算數99.8%。<sup>[1]</sup>

Show source details >

113年度施政計畫。請提供摘要。

根據第3號搜索結果,113年度中央氣象署的4個主要施政計畫為:1)強化氣象觀測 2)精準預報技術發展 強化預報



中華民國 112 年 10 月  
立法院第 10 屆第 8 會期

交通部中央氣象署 113 年度  
施政計畫及預算編列口頭報告

## 一、111 年度預算執行情形

(一)111 年度本署歲入預算共編列 2,897 萬元，決算數為 2,929 萬 4 千元，決算數占預算數 101.12%。

(二)111 年度本署歲出預算加計動支第一、二預備金共 21 億 5,114 萬 5 千元，決算數為 21 億 4,675 萬元，決算數占預算數 99.8%。

## 二、112 年度截至 9 月止預算執行情形

(一)112 年度本署歲入預算共編列 2,997 萬元，截至 9 月止預算分配數 1,782 萬 4 千元，實收數 2,548 萬 2 千元，占預算分配數 142.96%。

(二)112 年度本署歲出預算(含預備金動支數)共編列 19 億 692 萬 8 千元，截至 9 月止預算分配數 12 億 2,413 萬 6 千元，執行數 11 億 7,312 萬 7 千元，執行數占預算分配數 95.83%。



麥麥聊天



中央氣象署的署長是誰。

9:32 PM



根据提供的內容可以知道,交通部中央氣象署署長是程家平。

9:32 PM

113年度新聞媒體從業人員氣象資訊推廣說明會,總共有幾場。

9:32 PM



從上下文可以知道,113年度新聞媒體從業人員氣象資訊推廣說明會總共有4場,分別是北區、南區、東區和中區。

9:32 PM

北區場次的時間,與地點。

9:32 PM



根據提供的內容,北區場次的時間是2月29日(星期四),地點是中央氣象署617會議室(臺北市中正區公園路64號6樓)。

9:32 PM

111年度預算執行情形。

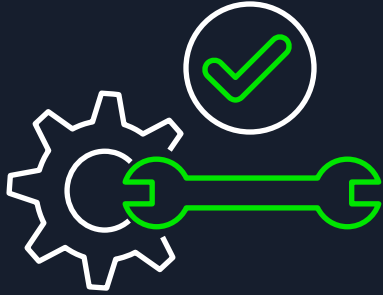
9:33 PM



據文件內容顯示,111年度本署歲入預算共編列2,897萬元,決算數為2,929萬4千元,決算數占預算數101.12%。111年度本署歲出預算加計動支第一、二預備金共21億5,114萬5千元,決算數為21億4,675萬元,決算數占預算數99.8%。

9:33 PM

# Customizing model responses for your business



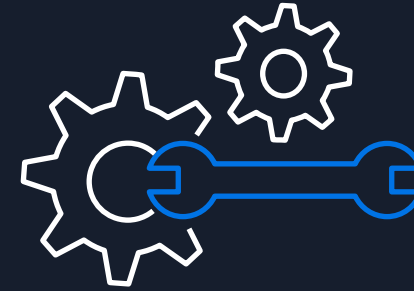
## Fine-tuning

### PURPOSE

Maximizing accuracy  
for **specific tasks**

### DATA NEED

**Small number** of  
labeled examples



## Continued pretraining

### PURPOSE

Maintaining model  
accuracy for  
**your domain**

### DATA NEED

**Large number** of unlabeled  
datasets



# Fine-Tuning in Action: Text Summarization

```
prompt = """"
Summarize the simplest and most interesting part of the following conversation.

#Person1#: Hello. My name is John Sandals, and I've got a reservation.
#Person2#: May I see some identification, sir, please?
#Person1#: Sure. Here you are.
#Person2#: Thank you so much. Have you got a credit card, Mr. Sandals?
#Person1#: I sure do. How about American Express?
#Person2#: Unfortunately, at the present time we take only MasterCard or VISA.
#Person1#: No American Express? Okay, here's my VISA.
#Person2#: Thank you, sir. You'll be in room 507, nonsmoking, with a queen-size bed. Do you approve, sir?
#Person1#: Yeah, that'll be fine.
#Person2#: That's great. This is your key, sir. If you need anything at all, anytime, just dial zero.

Summary:
""""

body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}
```

# Fine-Tuning in Action: Baseline Completion

```
response = bedrock_runtime.invoke_model(  
    modelId="meta.llama2-13b-chat-v1", # compare to chat model  
    body=json.dumps(body)  
)  
  
response_body = response["body"].read().decode('utf8')  
print(json.loads(response_body)["generation"])
```

A man named John Sandals checks into a hotel and provides identification and a credit card. The hotel only takes MasterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.

# Fine-Tuning in Action: Improved Completion

```
response = bedrock_runtime.invoke_model(  
    modelId=provisioned_model_arn, # custom fine-tuned model  
    body=json.dumps(body)  
)  
  
response_body = response["body"].read().decode('utf8')  
print(json.loads(response_body)["generation"])
```

John Sandals checks in the hotel with VISA and is assigned room 507, nonsmoking, with a queen-size bed.

[+ New chat](#)

Query documents ⓘ

Index

canvas-blogs-EN ▼



Titan Express



Based on the information provided for the client and their goals and risk profile, recommend funds to invest on, and explain why.

Client Background: A 38-year-old married couple with moderate monthly expenses and a mortgage. They aim to retire comfortably in 20 years and want to support clean energy initiatives. They have a medium risk tolerance.

Goals: To retire comfortably and invest in clean energy.

Risk Profile: Medium

Recommendation:

Based on the client's goals and risk profile, I recommend a mix of equity and bond funds. Here are the details: Equity Funds:

1. Vanguard Total Stock Market ETF (VTI): This ETF tracks the performance of the entire U.S. stock market. It has a low expense ratio of 0.04% and a diversified portfolio of over 3,000 stocks. VTI provides exposure to a wide range of sectors, including technology, healthcare, and finance, which aligns with the client's clean energy interests.
  2. iShares ESG Aware MSCI USA ETF (ESGU): This ETF invests in companies that meet ESG criteria, such as sustainability and social responsibility. It has a low expense ratio of 0.05% and a diversified portfolio of over 1,000 stocks. ESGU provides exposure to companies that are contributing to the transition to a greener economy, which aligns with the client's goals.
- Bond Funds:
3. Vanguard Total Bond Market ETF (BND): This ETF tracks the performance of the entire U.S. bond market. It has a low expense ratio of 0.05% and a diversified portfolio of over 3,000 bonds. BND provides stability and income, which can help the client achieve their retirement goals.
  4. iShares ESG Aware MSCI USA Bond ETF (ESGB): This ETF invests in companies that meet ESG criteria, such as sustainability and social responsibility. It has a low expense ratio of 0.05% and a diversified portfolio of over 1,000 bonds. ESGB provides exposure to companies that are contributing to the transition to a greener economy, which aligns with the client's goals.

Explanation: Equity Funds: VTI and ESGU are both excellent choices for equity funds. VTI provides broad exposure to the U.S. stock market, which is a



[+ New chat](#)☐ Query documents ⓘIndex  
canvas-blogs-EN FinancialPortfolio Titan Express

Based on the information provided for the client and their goals and risk profile, recommend funds to invest on, and explain why.

Client Background: A 38-year-old married couple with moderate monthly expenses and a mortgage. They aim to retire comfortably in 20 years and want to support clean energy initiatives. They have a medium risk tolerance.

Goals: To retire comfortably and invest in clean energy.

Risk Profile: Medium

Recommendation:

Fund Allocation: 40% Clean Energy Impact Fund, 30% Early Retirement Portfolio, 30% Global Diversification Portfolio

**Customised  
Results**



Based on the information provided, I recommend the following funds to invest in:

Vanguard Total Stock Market Index Fund Admiral Shares: This fund provides broad exposure to the U.S. stock market and has a low expense ratio of 0.04%. It is a good choice for investors with a medium risk tolerance.

iShares ESG Aware MSCI USA ETF: This fund invests in companies that have strong environmental, social, and governance (ESG) practices. It has a low expense ratio of 0.05% and is a good choice for investors who want to support clean energy initiatives.

Vanguard Total International Stock Index Fund Admiral Shares: This fund provides exposure to stocks from around the world and has a low expense ratio of 0.04%. It is a good choice for investors with a medium risk tolerance who want to diversify their portfolio.

Conclusion:

**Generic  
Results**



Ask me anything...

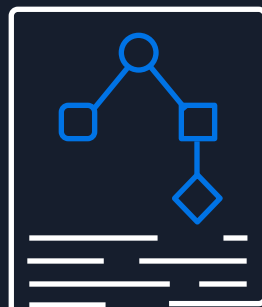


# Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES



Decompose into steps  
using available actions  
and Knowledge Bases  
for Amazon Bedrock



Execute action or  
search knowledge base

Observe results

Think about next step



Until final answer

# Agent in Action

Hi, I am looking for shoes

Could you please provide your name so I can look up your account?  
[Show trace >](#)

Jane Smith

I found two hiking boots in stock - ShoeID 2: Durable leather hiking boots in Brown or Black for \$240 and ShoeID 5: Waterproof hiking boots in Tan or Olive for \$99.99. Which would you like to order?  
[Show trace >](#)

Place an order for Shoe ID 5

I have placed an order for ShoeID 5 for you. The waterproof hiking boots in tan or olive green should be on their way soon! Let me know if you need any other assistance with finding shoes.  
[Show trace >](#)

# Thank you!

**Michael Lin**

[linmicht@amazon.com](mailto:linmicht@amazon.com)