



# Amazon Bedrock

The easiest way to build and scale generative AI  
applications with powerful tools and foundation models

Michael Lin

Sr. Solutions Architect  
Amazon Web Services

# Generative AI Stack

## APPLICATIONS THAT LEVERAGE LLMs AND FMs



Amazon Q



AWS App Studio

## TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails

| Agents

| Studio

| Customization

| Custom Model Import

| Amazon Models

## INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron





# Amazon Bedrock

The easiest way to build and scale generative AI applications with powerful tools foundation models

Choice of leading FMs through a single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and data governance

# 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 1.5 Large	Amazon Titan Text Premier	Claude 3.5 Sonnet	Command	Llama 3.2	Mistral Large 2 (24.07)	Stable Image Ultra
Jamba 1.5 Mini	Amazon Titan Text Lite	Claude 3 Opus	Command Light	Llama 3.1	Mistral Large (24.02)	Stable Diffusion 3 Large
Jamba-Instruct	Amazon Titan Text Express	Claude 3 Sonnet	Embed English	Llama 3 8B	Mistral Small	Stable Image Core
Jurassic-2 Ultra	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed Multilingual	Llama 3 70B	Mixtral 8x7B	Stable Diffusion XL1.0
Jurassic-2 Mid	Amazon Titan Text Embeddings V2	Claude 2.1	Command R+	Llama 2 13B	Mistral 7B	Stable Diffusion XL 0.8
	Amazon Titan Multimodal Embeddings	Claude 2	Command R	Llama 2 70B		
	Amazon Titan Image Generator	Claude Instant				



# Claude 3.5 Sonnet



## Frontier intelligence

Claude 3.5 Sonnet raises the industry bar for intelligence, outperforming Claude 3 Opus on a wide range of evaluations, with the speed and cost of mid-tier model, Claude 3 Sonnet.

## Frontier visual reasoning

These improvements are most noticeable for tasks that require visual reasoning, such as interpreting charts and graphs. Claude 3.5 Sonnet can also accurately transcribe text from imperfect images.

# Anthropic's Claude Models

FOUNDATION MODELS CRAFTED FOR THE ENTERPRISE

Claude excels at use cases that require **speed & intelligence** and with new capabilities it now can:



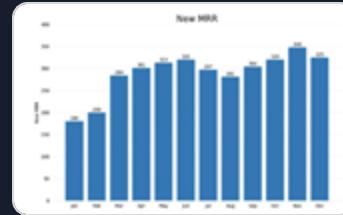
**ANALYZE THE CONDITION OF A PACKAGE**

Evaluate images for insurance claims, adjustments, etc



**DESCRIBE THE CONDITION OF THIS VEHICLE**

Generate product information & insights from images



**RECREATE THIS GRAPH IN PYTHON**

Transcribe handwritten notes, diagrams



**SUMMARIZE THIS REPORT**

Extract data from documents, charts, graphs



[Amazon Bedrock](#) > Chat/text playground

Mode Chat ▾



Claude 3.5 Sonnet



根據這張台灣高鐵車票的資訊，我可以提取以下內容：

姓名：未顯示

出發站：南港 (Nangang)

抵達站：左營 (Zuoying)

出發日期：2018/10/24

出發時間：11:20

抵達時間：13:05

車次：125

車廂：5號車廂（標準廂）

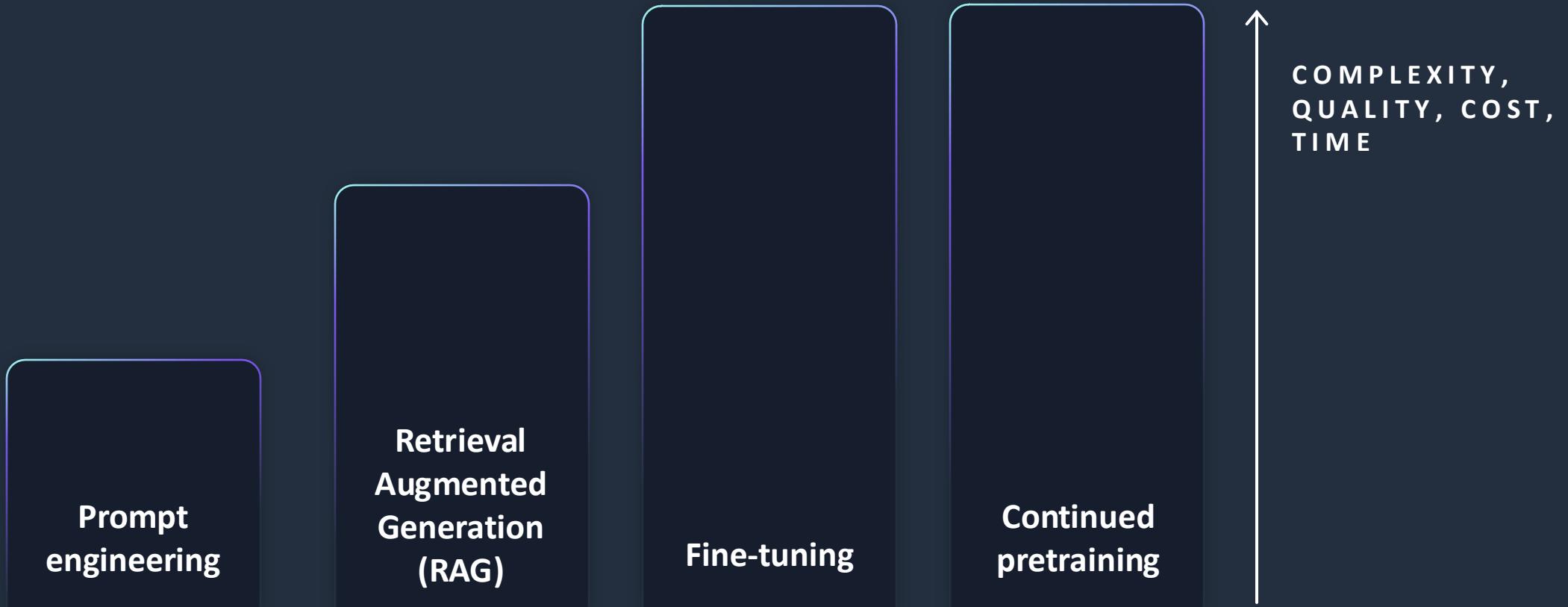
座位：2C

票價：NT\$1530

付款方式：現金

票種：單程票，成人票

# Common approaches for customizing FMs



# Amazon Bedrock Prompt Management

STREAMLINE YOUR GENERATIVE AI  
APPLICATIONS DEVELOPMENT

Built-in Prompt Builder in Amazon Bedrock  
console and SDK APIs

Prompt Library for easy cataloguing and  
management

Seamless integration with Flows and Bedrock  
Studio

The screenshot shows the 'Prompt management' section of the Amazon Bedrock console. It includes sections for 'How it works', 'Create a prompt', 'Test the prompt', and 'Use the prompt'. Below these is a table listing three prompts:

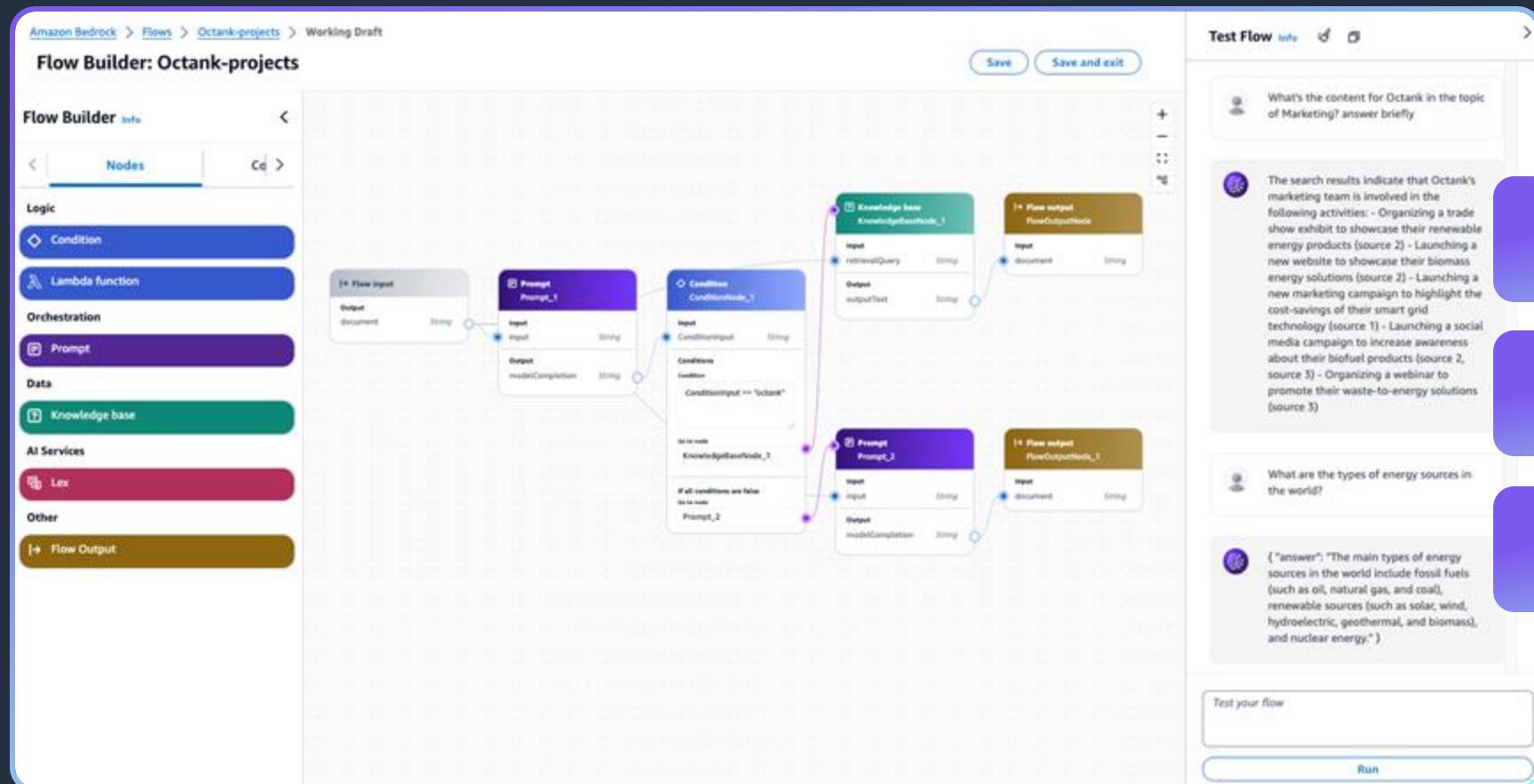
Name	Description	Last updated
MyTestPrompt-20240628-135832	This is my test prompt for the translation use case	June 28, 2024, 13:58 (UTC+02:00)
CustomerService	This is my prompt for Customer Service use case.	June 27, 2024, 11:43 (UTC+02:00)
MyTestPrompt-001	Prompt template for the virtual assistant use case.	June 26, 2024, 14:03 (UTC+02:00)

The screenshot shows the 'Prompt builder' interface for a 'CustomerServiceSummary' prompt. It includes sections for 'Prompt', 'Configurations - Optional', 'Test variables (4)', and a 'Test window'. The 'Prompt' section contains text input fields for user context and conversation summary. The 'Configurations' section allows setting model (Claude 3 Haiku), inference parameters, system prompts, and randomness diversity. The 'Test variables' section lists language (English), format (JSON), context (customer has an Android phone), and conversation (customer phone not charging). The 'Test window' section displays a conversation transcript and a generated response.



# Amazon Bedrock Prompt Flows

VISUALIZE AND ACCELERATE GENERATIVE AI DEVELOPMENT WORKFLOWS



Drag-and-drop interface

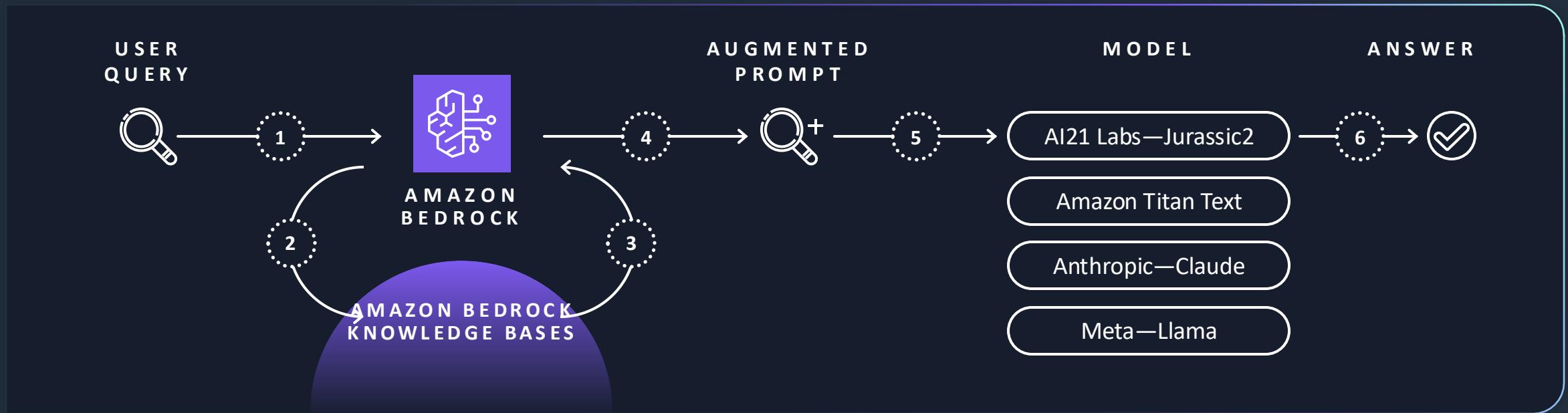
Direct testing and deployment

Version control and aliasing



# Amazon Bedrock Knowledge Bases

## 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 multturn conversations

Automatic citations with retrievals to improve transparency

### 台北大眾捷運股份有限公司票證作業管理要點修訂意見表

1 / 10 | - 100% + | ☰ ⌂

下載 打印 ⌂



1



2



3

## 臺北捷運系統旅客須知

中華民國八十五年三月二十八日公告  
中華民國八十七年十二月二十四日修正  
中華民國八十八年十一月十日公告  
中華民國九十一年八月二十日修正  
中華民國九十三年十二月二十四日公告  
中華民國九十七年三月二十八日修正  
中華民國九十七年六月三日公告  
中華民國九十八年三月十二日修正  
中華民國一〇三年五月二十七日公告  
中華民國一〇四年三月十八日修正  
中華民國一〇四年十二月十一日公告  
中華民國一〇六年五月十二日修正  
中華民國一〇六年十月五日公告  
中華民國一〇七年十月二十六日修正  
中華民國一一一年三月二十九日公告  
中華民國一一二年二月七日公告  
中華民國一一二年十一月八日公告  
中華民國一一三年六月十三日公告

### 壹、一般規定

一、臺北大眾捷運股份有限公司(以下簡稱本公司)為提供捷運系統旅客安全、可靠、便捷、舒適之服務，特依「臺北市大眾捷運系統旅客運送自治條例」規定訂定本須知，並於車站公告，變更或調整時亦同。

### 二、本須知用語定義如下：

- (一) 捷運範圍：為本公司所經營之大眾捷運系統路網範圍內所有路線、場、站與列車等區域。
- (二) 旅客：指搭乘本公司列車，或持有有效車票並進出乘車處所車站大廳之人。

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1?modelId=a... ☆

aws Services Search [Option+S] Oregon WSParticipantRole/Participant @ 1561-5387-8293

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

## 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

### knowledge-base-quick-start-6zlo3

Test Delete Edit

#### Knowledge base overview

Knowledge base name: knowledge-base-quick-start-6zlo3

Knowledge base ID: GRNV74BLMR

Knowledge base description: —

Status: Ready

Service Role: AmazonBedrockExecutionRoleForKnowledgeBase\_6zlo3

Created date: July 27, 2024, 15:51 (UTC+08:00)

Log Deliveries: Configure log deliveries and event logs in the [Edit](#) page.

#### 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
No tags No tags to display	

Manage tags

#### Test knowledge base

Generate responses

Claude 3 Sonnet v1 | QDT

Change

Configure your retrieval and responses  
To customize the search strategy for your knowledge base, select the configurations icon .

Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.

在台北捷運內可不可以溜直排輪？

Run

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1?modelId=a... ☆

AWS Services Search [Option+S] Oregon WSParticipantRole/Participant @ 1561-5387-8293

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

## 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](#)

### Amazon Bedrock > Knowledge bases > knowledge-base-quick-start-6zlo3

## knowledge-base-quick-start-6zlo3

Test Delete Edit

#### Knowledge base overview

Knowledge base name: knowledge-base-quick-start-6zlo3

Knowledge base ID: GRNV74BLMR

Knowledge base description: —

Status: Ready

Service Role: AmazonBedrockExecutionRoleForKnowledgeBase\_6zlo3

Created date: July 27, 2024, 15:51 (UTC+08:00)

#### Log Deliveries

Configure log deliveries and event logs in the [Edit](#) page.

#### 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
No tags	

Enter your message here ▶ Run

### Test knowledge base

Generate responses

Claude 3 Sonnet v7 | ODT Change

Configure your retrieval and responses To customize the search strategy for your knowledge base, select the configurations icon .

在台北捷運內可不可以溜直排輪？

根據台北捷運系統旅客須知的規定,在捷運範圍內是不允許使用直排輪的。具體來說,第8條明確規定"在捷運範圍內,不得有下列行為:(一)使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。"因此,在台北捷運內是禁止溜直排輪的。[\[1\]](#)

Show source details >

## 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 &amp; deployment

## Test Knowledge base

 Generate responses**Claude 3 Sonnet v1** | [ODT](#)[Change](#)**Configure your retrieval and responses**

To customize the search strategy for your knowledge base, select the configurations icon



在台北捷運內可不可以溜直排輪？

 根據台北捷運系統旅客須知的規定,在捷運範圍內是不允許使用直排輪的。具體來說,第8條明確規定"在捷運範圍內,不得有下列行為:(一)使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。"因此,在台北捷運內是禁止溜直排輪的。[\[1\]](#)[< Hide source details](#)

## Source details (1)

Each container below shows a chunk that comes from a data source and that relates to your query.

## ▶ Query configurations (1)



## ▼ Source chunk 1



(七) 其他經本公司公告禁止之事項。八、在捷運範圍內，不得有下列行為：(一) 使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。

(二) 推擠或影響自己或他人安全。(三) 坐、臥於車廂、車站地板、設備或設施上。北捷024184臺北捷運系統旅客須知 3 (四) 吸菸。

(五) 其他行為有造成大眾捷運系統設備損壞、運轉障礙或構成危險之虞。九、在捷運範圍內為下列行為，應向本公司申請許可後，始得為之：(一) 聚眾講演、播放音響、演奏樂器或其他干擾之行為。(二) 張貼、塗抹、刻畫任何文字、圖畫或其他類似東西於各項設施及建築物上。(三) 於車站或車廂內，照相、拍攝或攝影，而妨礙他人或系統安全之虞者。(四) 非營運時間內，於車站或車廂內逗留。(五) 向他人為傳教、市場調查或其他類似行為。(六) 散發報紙、傳單、廣告物或宣傳品。(七) 使用車站、車廂內未開放使用之電源插座。十、在捷運範圍內，電扶梯之搭乘規定及注意事項如下：

(一) 年長及行動不便者宜改搭電梯。(二) 須遵循電扶梯方向搭乘，握好扶手、站穩踏階，勿倚靠側板。(三) 禁止於電扶梯上奔跑、嬉戲、跳躍、跨越兩側護欄或其他危險行為。(四) 禁止攜帶大型物品搭乘電扶梯。貳、車票使用規定十一、車票種類：(一) 單程票：提供旅客單次使用之車票。(二) 團體票：提供旅客 2人以上，全程同行且起訖站相同使用之車票。(三) 定期票：提供旅客於一定期間內使用之車票。(四) 回數票：可供旅客於一定區間或不限區間搭乘一定次數之車票。(五) 儲值卡：各發行機

# Customizing model responses for your business



## Fine-tuning

### PURPOSE

Maximizing accuracy  
for specific tasks

### DATA NEED

Small number of  
labeled examples



## Continued pretraining

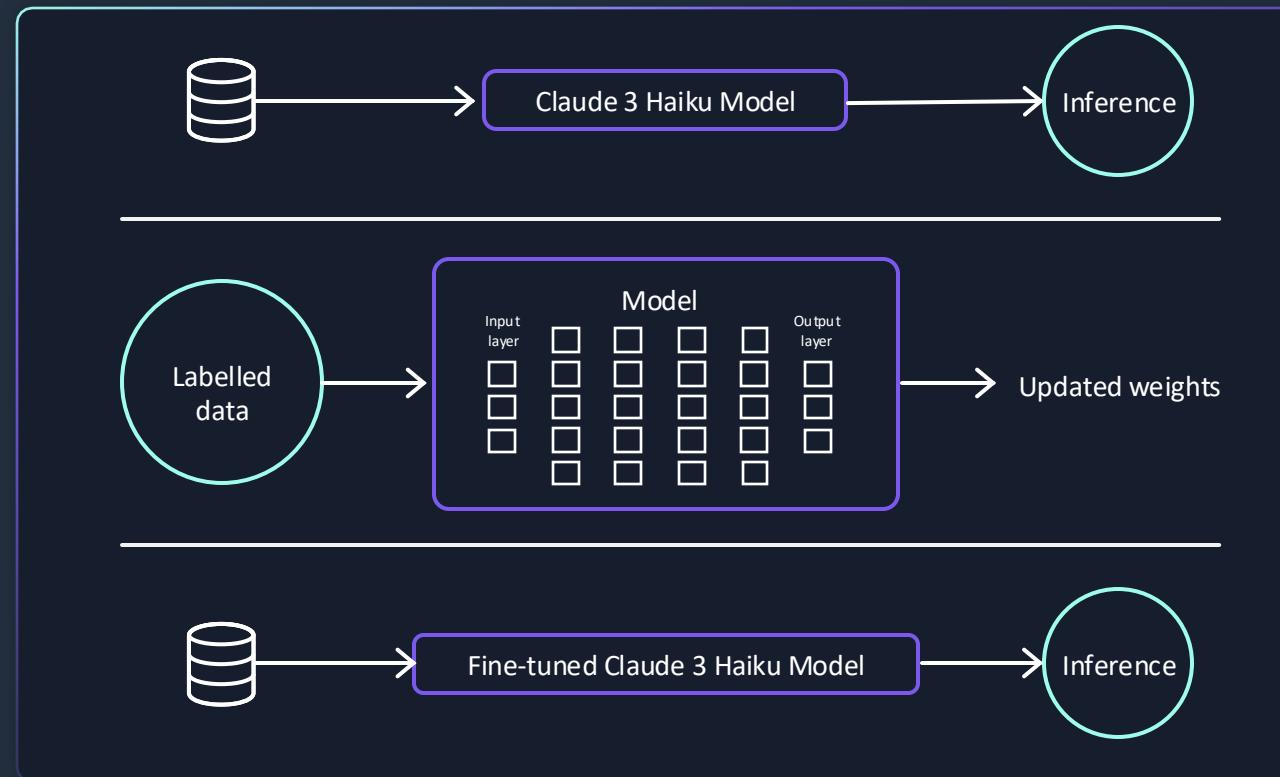
### PURPOSE

Maintaining model  
accuracy for your domain

### DATA NEED

Large number of unlabeled  
datasets

# Fine-tuning for Anthropic's Claude 3 Haiku in Amazon Bedrock



Amazon Bedrock is the only fully-managed service that provides customers with the ability to fine-tune Claude models

# Amazon Bedrock Model Evaluation

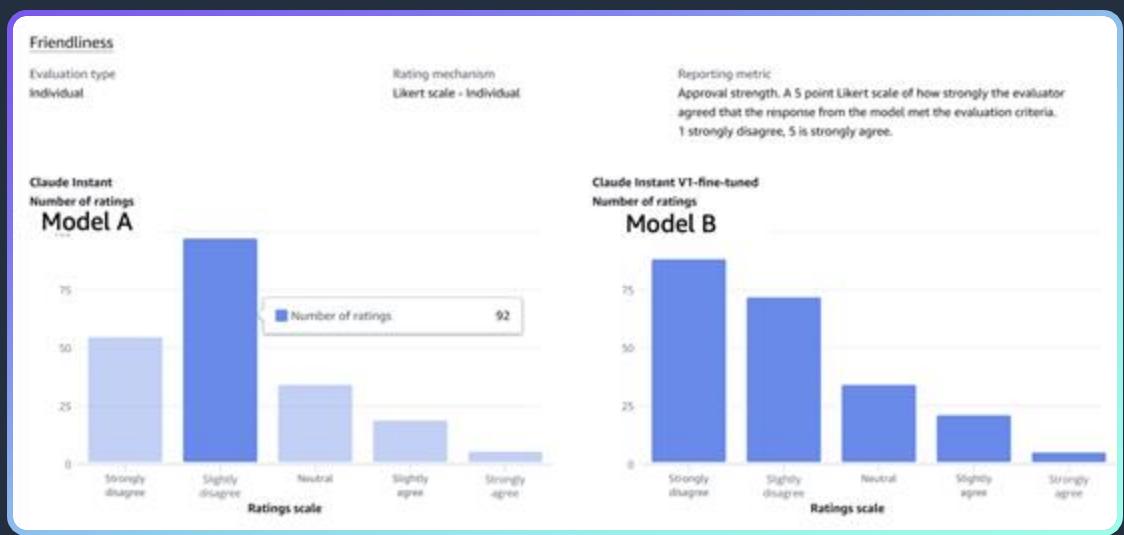
EVALUATE, COMPARE, AND SELECT THE BEST FM FOR YOUR USE CASE

Automatic or human evaluation method

Curated datasets or bring your own

Predefined and custom metrics

## HUMAN EVALUATION REPORT



## AUTOMATIC EVALUATION REPORT

Text summarization evaluation summary (3)	
The results for text summarization consist of accuracy, toxicity, and robustness, which indicate the quality of the summaries generated by the model. <a href="#">Learn more.</a>	
Accuracy	Toxicity
Dataset	Value
CNN/DailyMail	.6
S3 URI 3	.4
Dataset	Value
S3 URI 1	.5
Robustness	
Dataset	Value
CNN/DailyMail	.4
S3 URI 2	.6

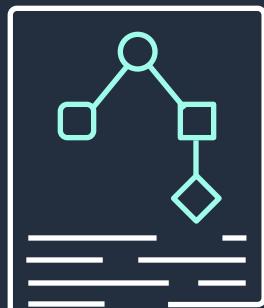


# Amazon Bedrock Agents

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



Decompose into steps  
using available actions  
and Amazon Bedrock Knowledge  
Bases



Execute action or  
search knowledge base



Observe results



Think about next step



Until final answer

# Amazon Bedrock Guardrails

Implement safeguards customized to your application requirements and aligned to your responsible AI policies

Block as much as 85% more harmful content than protection natively provided by some FMs on Amazon Bedrock today, and filters over 75% hallucinated responses for RAG and summarization workloads



Evaluate prompts and model responses for agents, knowledge bases, FMs in Amazon Bedrock, and custom or third-party FMs

Configure thresholds to filter harmful content, jailbreaks and prompt injection attacks

Define and disallow denied topics with short natural language descriptions

Remove personally identifiable information (PII) and sensitive information in gen AI apps

Filter hallucinations by detecting groundedness and relevance of model responses based on context



# Amazon Bedrock Security

Helps keep your data  
secure and private



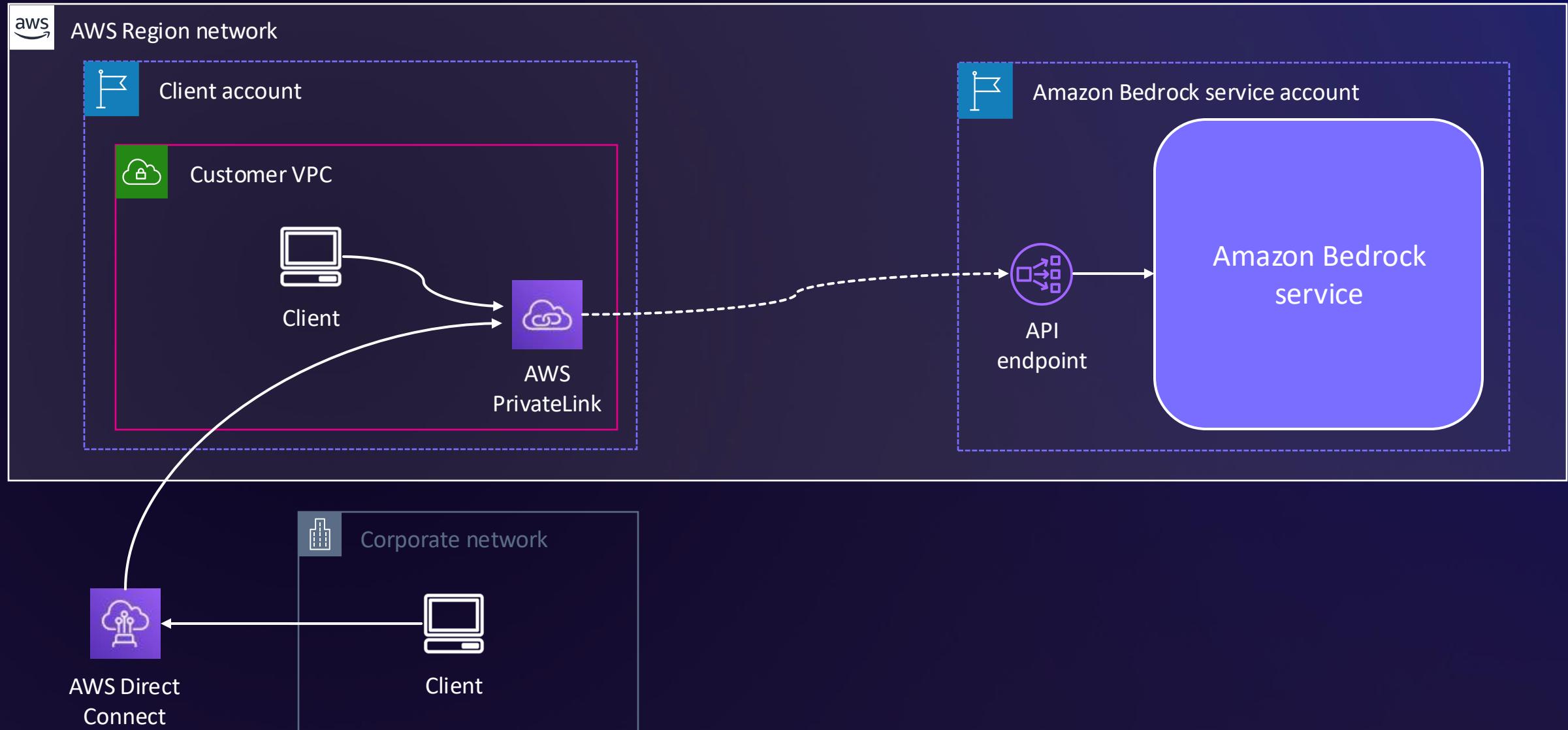
None of the customer's data is used to  
train the underlying model

All data is encrypted in transit and at rest; data  
used for customization is securely transferred  
through customer's VPC

Data remains in the Region where the  
API is processed

Support for GDPR, SOC, ISO, CSA compliance,  
and HIPAA eligibility

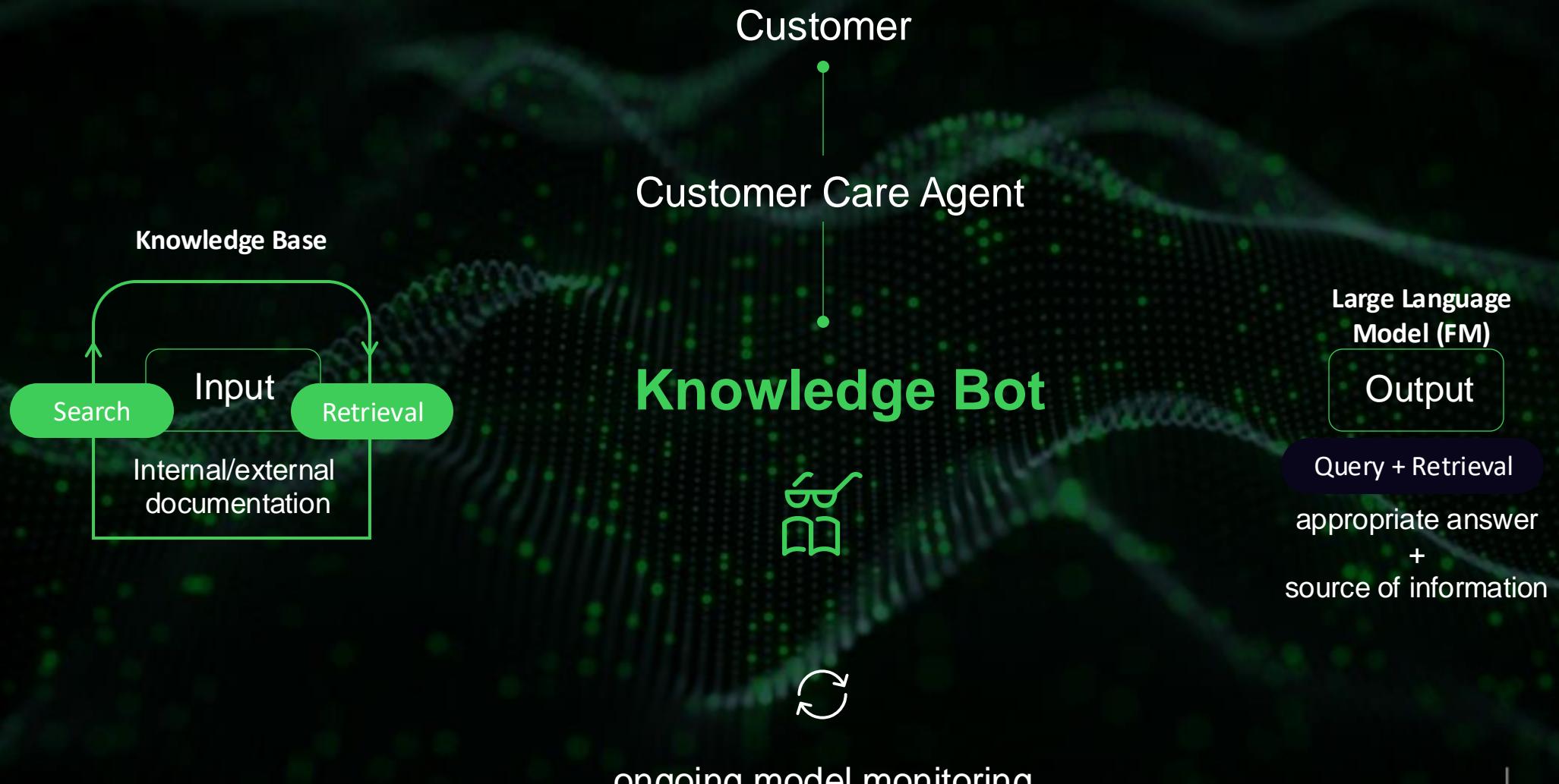
# Client connectivity



# AI @ SCALE

## at Schneider Electric

# Ensuring trustworthy answers to our customers



# Responsible AI Dimensions

## Trustworthiness

Ensuring an AI system operates reliably

## Toxicity & Safety

Preventing from harmful and unfair behavior of a model

## Transparency

Informing stakeholders and users about AI system

## Domain Knowledge Expert

Human supervision and overview at any stage of the AI system

## Explainability

Understanding and evaluation of model outputs

## Security & Privacy

Data privacy and protection for theft and exposure

## Governance

Implementation and enforcement of RAI practices within an organization

## Sustainability

Using the right resource for the right task

# Pillar 1: Trustworthiness

Life Is On

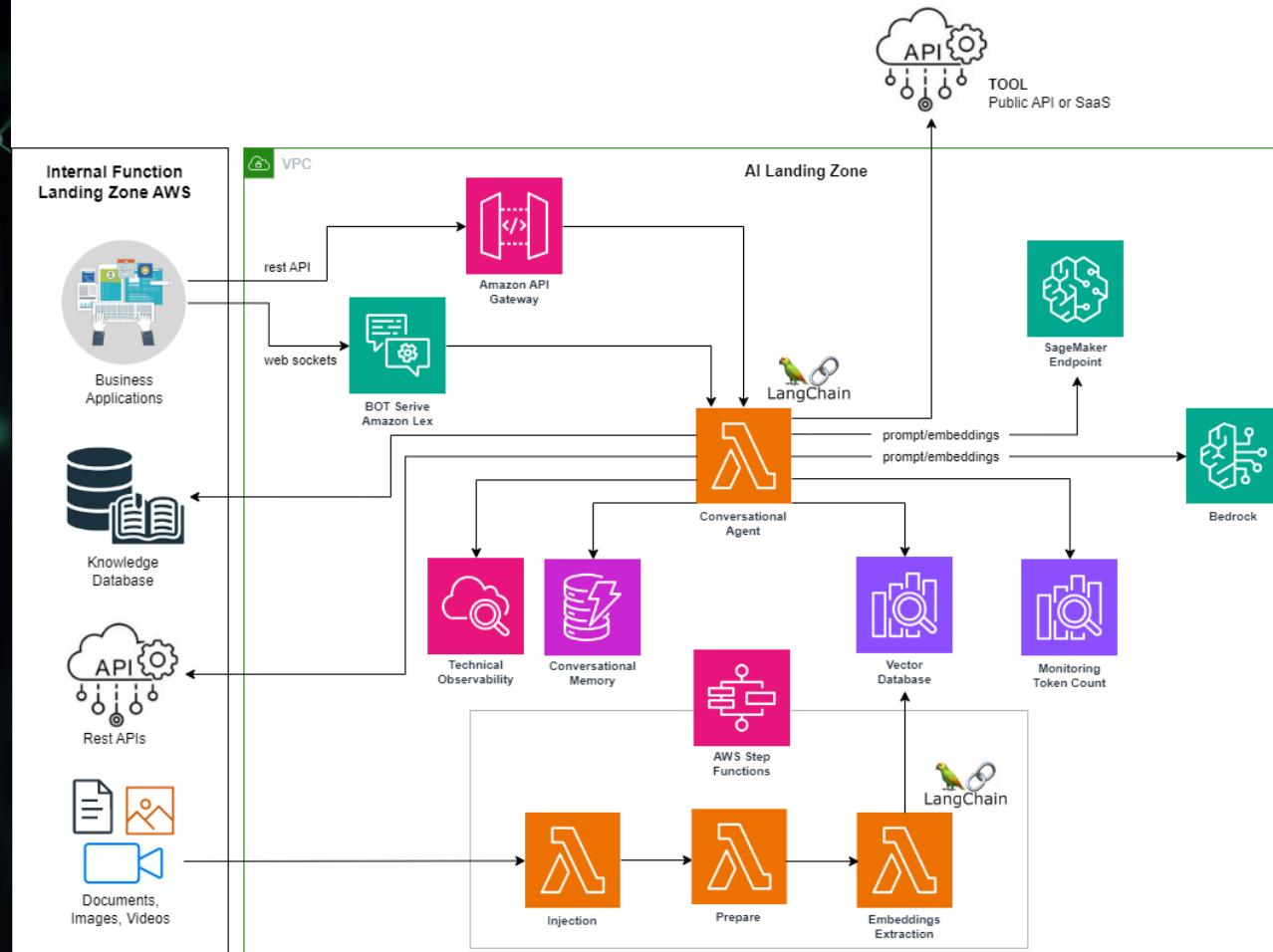
Schneider  
Electric

## Definition

- Ensure an AI system operates reliably
- Earn user's trust from with quality results based on data from validated sources.

## Objectives

- Retrieval Augmented Generation (RAG)
- Model reasoning with Chain-of-thoughts



# Pillar 2: Toxicity & Safety

## Definition

- Prevent from harmful and unfair behavior of a model
- Ensure reliable and correct replies

## Objectives

- Guardrails and input/output filtering
- Constitutional AI

## Services

- Amazon Bedrock

The screenshot shows the Schneider Electric homepage with a dark background featuring a city skyline. At the top, there's a navigation bar with links for Insights, Solutions, Sustainability, Company, Investors, and Careers. A search bar is also at the top right. Below the navigation, a main heading reads "Innovate with Schneider Electric" with a subtext about their purpose: "Our purpose is to empower all to make the most of our energy and resources, bridging progress and sustainability for all. At Schneider, we call this Life Is On." Two buttons are visible: "Our Purpose >" and "Investor Relations >". The main content area features a conversation with the AI assistant SAIA. SAIA says, "Hi! I am SAIA (Schneider AI Assistant), an AI-powered assistant. How can I help you today?" A user asks, "What is the weather forecast for today?", to which SAIA responds, "I'm sorry, the provided context does not contain enough information for me to answer the question about today's weather forecast." Below this, there are two dropdown menus: "Show references" and "Show Reasoning". At the bottom, there's a text input field with placeholder text "Enter your message and press 'Send'" and a "Send" button. There are also "Clear chat history" and "Like" and "Dislike" buttons.

# Pillar 3: Explainability

## Definition

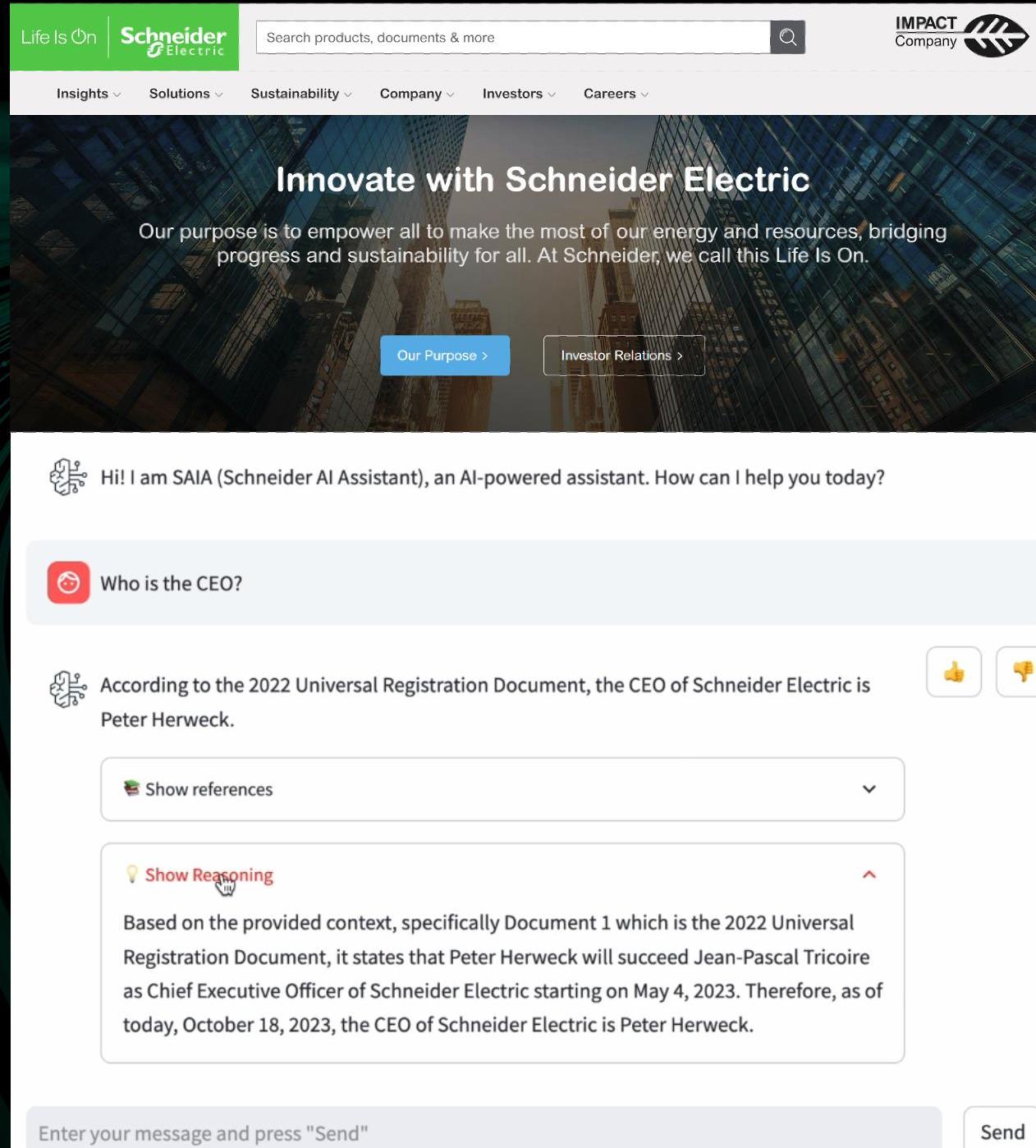
- Understand and evaluate outputs of models
- Identify model's reasoning for giving specific response. Spot important words/tokens.

## Objectives

- Document references
- Ask model to explicitly detail reasoning steps (chain-of-thoughts)

## Services

- Amazon Bedrock



The screenshot shows the Schneider Electric website homepage. At the top, there is a navigation bar with links for Insights, Solutions, Sustainability, Company, Investors, and Careers. A search bar is also present. The main header reads "Innovate with Schneider Electric" with the tagline "Our purpose is to empower all to make the most of our energy and resources, bridging progress and sustainability for all. At Schneider, we call this Life Is On." Below the header, there are two buttons: "Our Purpose >" and "Investor Relations >". The main content area features a large image of a modern skyscraper. A chat interface with the AI assistant SAIA is overlaid on the page. SAIA asks, "Who is the CEO?", and the AI-powered response states, "According to the 2022 Universal Registration Document, the CEO of Schneider Electric is Peter Herweck." There are "Show references" and "Show Reasoning" buttons. The reasoning provided is: "Based on the provided context, specifically Document 1 which is the 2022 Universal Registration Document, it states that Peter Herweck will succeed Jean-Pascal Tricoire as Chief Executive Officer of Schneider Electric starting on May 4, 2023. Therefore, as of today, October 18, 2023, the CEO of Schneider Electric is Peter Herweck." At the bottom, there is a text input field for messages and a "Send" button.

# Pillar 4: Domain Knowledge

## Definition

- Human supervision and overview at any stage of the AI system
- User side: Enable feedback
- Developer side: evaluate model outputs
- Measure model's performance with human annotations
- Evaluate model's output quality (benchmark)
- Amazon SageMaker GroundTruth

## Objectives

## Services

HUMAN QUESTION: Show me the trend of Schneider revenue in the past 10 years?

AI ANSWER:

Based on the provided annual reports and registration documents, Schneider Electric appears to have seen steady organic revenue growth in the range of 3-7% over the past 10+ years.

Reasoning:

Based on the provided context, I can summarize the key information about Schneider Electric's revenue trend over the past 10 years as follows:

- The 2020 Universal Registration Document shows revenue growth for the past 15 years through a revenue distribution diagram, indicating continued growth.
- The 2018 Annual Report states that revenues rose to 25.7 billion EUR in 2018, representing 7% organic growth overall.
- The 2020 Universal Registration Document mentions reaffirming medium-term ambitions for organic growth of +3% to +6% in revenues.
- There are some specific revenue figures provided for 2016 (25.7 billion EUR) and 2018 (25.7 billion EUR) in the annual reports, showing steady growth.

Overall, the context indicates that Schneider Electric has seen consistent organic revenue growth in the 3-7% range over the past 10+ years. More granular yearly figures are not provided, but the trend seems to be one of steady growth.

Select an option

Phrasing and Facts are correct	1
Facts are correct but Phrasing is insufficient	2
Partial answer, some crucial information is missing.	3
Correct but outdated answer	4
Facts are wrong but Phrasing is good	5
Phrasing and Facts are wrong	6

# Pillar 4: Domain Knowledge

## Definition

- Human supervision and overview at any stage of the AI system

## Objectives

- Measure model's performance with human annotations
- Evaluate model's output quality (benchmark)

## Services

- Amazon SageMaker GroundTruth

## Chat

Chat with LLM about your documents

 Hi! I am SAIA (Schneider AI Assistant), an AI-powered assistant. How can I help you today?

 Who is the CEO?

 According to the 2022 Universal Registration Document, the CEO of Schneider Electric is Peter Herweck.

 Show references

 Show Reasoning

 Feedback submitted

Enter your message and press "Send"

Send

Clear chat history

# Options for Solution Adoption

- Develop prototype using AWS components, with full control on customization
- Leverage open-source project and accelerate time-to-market
- Evaluate and buy from vendor's pre-assembled solution

# A minimalist architecture



```
[37]: # adjust modelId for the workshop demo
```

```
modelId = 'anthropic.claude-3-5-sonnet-20240620-v1:0'
```

```
*[38]: # adjust knowledge case id for the workshop demo
```

```
kb_id = "CJOMCDOKNC"
```

```
[ ]:
```

```
[26]: # from langchain.llms.bedrock import Bedrock
```

```
import langchain
from langchain_aws import ChatBedrock
from langchain.retrievers.bedrock import AmazonKnowledgeBasesRetriever
```

```
llm = ChatBedrock(model_id=modelId,
                  client=bedrock_client)
```



```
[56]: from langchain.prompts import PromptTemplate

PROMPT_TEMPLATE = """
Human: You are an intelligent and friendly assistant, and provides answers to questions based on facts wehre possible.
Use the following pieces of information to provide a concise answer to the question enclosed in <question> tags.
Do take yuor time to validate your answer before replying.
If you don't know the answer, just say that you don't know, don't try to make up an answer.
Do not answer any political or religious questions. Just reply "我只針對氣象局內建檔資料，進行問題的回覆。如有疑問，請聯繫管理員 Michael".
Be sure to answer in traditioanl chineese, using phrases most frequently used in Taiwan.

<context>
{context}
</context>

<question>
{question}
</question>

The response should be specific and use statistics or numbers when possible.

Assistant:"""
claude_prompt = PromptTemplate(template=PROMPT_TEMPLATE,
                                input_variables=["context", "question"])
```



```
[39]: query = "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?"
```

```
# query = "北區場次的日期與地點?"
```

```
# query = "111年度預算執行情形?"
```

```
retriever = AmazonKnowledgeBasesRetriever(  
    knowledge_base_id=kb_id,  
    retrieval_config={"vectorSearchConfiguration":  
        {"numberOfResults": 4,  
         'overrideSearchType': "SEMANTIC", # optional  
        }  
    },  
    # endpoint_url=endpoint_url,  
    # region_name=region,  
    # credentials_profile_name=<profile_name>,  
)  
docs = retriever.get_relevant_documents(  
    query=query  
)  
pp.pprint(docs)
```

```
[ Document(metadata={'location': {'s3Location': {'uri': 's3://cwa-rag-demo-bucket-20240220/CWA_2024_MediaBrief.pdf'}}, 'type': 'S3'}, 'score': 0.7734306}, page_content='1 交通部中央氣象署 「113年度新聞媒體從業人員氣象資訊推廣說明會」實施計畫 壹、目的 交通部中央氣象署為增進國內新聞媒體從業人員對本署氣象業務之認識，正確解讀氣象資訊及妥適應用，特以氣象防災及地震防災為主題，分別於北、東、中、南4區各舉辦1場本說明會，以深化氣象資訊之推廣與運用。貳、主(協)辦機關 一、主辦機關：中央氣象署 二、共同主辦：臺東縣政府、國立自然科學博物館 參、舉辦方式 一、參加對象：國內各媒體業者(報紙、廣播、電視等) 二、說明會共4場，日期時間及地點如下表： 場次 日期 地點 北區 2月29日 (星期四) 中央氣象署617會議室 (臺北市中正區公園路64號6樓) 南區 3月7日 (星期四) 中央氣象署臺灣南區氣象中心1樓視聽室 (臺南市中西區公園路21號) 東區 3月20日 (星期三) 中央氣象署臺東氣象站 (臺東市大同路106號) 中區 4月3日 (星期三) 國立自然科學博物館第二科學教室 (臺中市北區館前路1號) 三、研討會議程 時間 講程 講座 09:00~09:20 報到 09:20~09:30 致詞 09:30~11:00 氣候變遷下之氣象防災精進作為 氣象預報中心 11:00~11:10 休息 11:10~12:10 地震速報資訊的通報與應用 地震測報中
```



```
qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    return_source_documents=True,
    chain_type_kwargs={"prompt": claude_prompt}
)
```

```
[42]: answer = qa.invoke(query)
pp.pprint(answer)
```

```
{ 'query': '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？',
  'result': '根據提供的資訊，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。具體場次如下：\n'
            '\n'
            '1. 北區場次：2月29日(星期四)在中央氣象署617會議室舉行\n'
            '2. 南區場次：3月7日(星期四)在中央氣象署臺灣南區氣象中心1樓視聽室舉行\n'
            '3. 東區場次：3月20日(星期三)在中央氣象署臺東氣象站舉行 \n'
            '4. 中區場次：4月3日(星期三)在國立自然科學博物館第二科學教室舉行\n'
            '\n'
            '這4場說明會分別在北、南、東、中四個區域舉辦，目的是為了增進國內新聞媒體從業人員對氣象業務的認識，並正確解讀和應用氣象資訊。',
  'source_documents': [ Document(metadata={'location': {'s3Location': {'uri': 's3://cwa-rag-demo-bucket-20240220/CWA_2024_MediaBrief.pdf'}}, 'type': 'S3'}, score: 0.7734306, page_content='交通部中央氣象署 「113年度新聞媒體從業人員氣象資訊推廣說明會」實施計畫 壹、目的 交通部中央氣象署為增進國內新聞媒體從業人員對本署氣象業務之認識，正確解讀氣象資訊及妥適應用，特以氣象防災及地震防災為主題，分別 於北、東、中、南4區各舉辦1場本說明會，以深化氣象資訊之推廣與運用。 貳、主(協)辦機關 一、主辦機關：中央氣象署 二、共同主辦：臺東縣政府、國立自然科學博物館 參、舉辦方式 一、參加對象：國內各媒體業者(報紙、廣播、電視等) 二、說明會共4場，日期時間及地點如下表： 場次 日期 地點 北區 2月29日 (星期四) 中央氣象署617會議室 (臺北市中正區公園路64號6樓) 南區 3月7日 (星期四) 中央氣象署臺灣南區氣象中心1樓視聽室 (臺南市中西區公園路21號) 東區 3月20日 (星期三) 中央氣象署臺東氣象站 (臺東市大同路106號) 中區 4月3日 (星期三) 國立自然科學博物館第二科學教室 (臺中市北區館前路1號) 三、研討會議程 時間 議程 講座 09:00~09:20 報到 09:20~09:30 致詞 09:30~11:00 氣候變遷下之氣象防災精進作為 氣象預報中心 11:00~11:10 休息 11:10~12:10 地震速報資訊的通報與應用 地震測報中心 12:10~13:30 互動交流及用餐 2 四、欲參加之媒體人員可依其便利性，擇一場次參加。 肆、報名方式及研習證明一、各場次報名截止日為活動舉辦日前2週，請逕至說明會資訊平台(網址：https://reurl.cc/WRX7ly) 下報名頁面填寫報名表。 北區場次：自即日起至113年2月14日(星期三)止。 南區場次：自即日起至113年2月28日(星期三)止。 東區場次：自即日起至113年3月5日(星期二)止。 中區場次：自即
```

A screenshot of a web browser window titled "Knowledge Retrieval for CWA". The interface consists of two main sections: a question input section on the left and an answer output section on the right.

**Question Input Section:**

- Text input field placeholder: 請輸入問題
- Text input field content: 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？直接給答案。
- Buttons: Clear (grey), Submit (orange)

**Answer Output Section:**

- Text output field placeholder: 這裏是答案
- Text output field content: 根據提供的資訊，113年度新聞媒體從業人員氣象資訊推廣說明會總共舉辦4場。
- Buttons: Flag (grey)

這裏是答案

根據提供的資訊，113年度新聞媒體從業人員氣象資訊推廣說明會總共舉辦4場。

Flag

#### 參、舉辦方式

一、參加對象：國內各媒體業者(報紙、廣播、電視等)

二、說明會共4場，日期時間及地點如下表：

場次	日期	地點
北區	2月29日 (星期四)	中央氣象署617會議室 (臺北市中正區公園路64號6樓)
南區	3月7日 (星期四)	中央氣象署臺灣南區氣象中心1樓視聽室 (臺南市中西區公園路21號)
東區	3月20日 (星期三)	中央氣象署臺東氣象站 (臺東市大同路106號)
中區	4月3日 (星期三)	國立自然科學博物館第二科學教室 (臺中市北區館前路1號)

# Illustrative Open Source Project

github.com/aws-samples/bedrock-claude-chat

README Code of conduct MIT-0 license Security

## Bedrock Claude Chat

CDK tests passing

**⚠ Warning**

A major update to v2 is planned soon. v2 will not be backward compatible with v1, and **existing RAG bots will no longer be usable**. For more details, please refer to the [migration guide](#).

**⚠ Warning**

If you are using old version (e.g. `v0.4.x`) and wish to use the latest version, refer [migration guide](#). Without any care, **ALL DATA IN Aurora cluster WILL BE DESTROYED, and NO LONGER USERS CANNOT USE EXISTING BOTS WITH KNOWLEDGE AND CREATE NEW BOTS.**

This repository is a sample chatbot using the Anthropic company's LLM [Claude](#), one of the foundational models provided by [Amazon Bedrock](#) for generative AI.

Packages  
No packages published

Contributors 56

+ 42 contributors

Languages

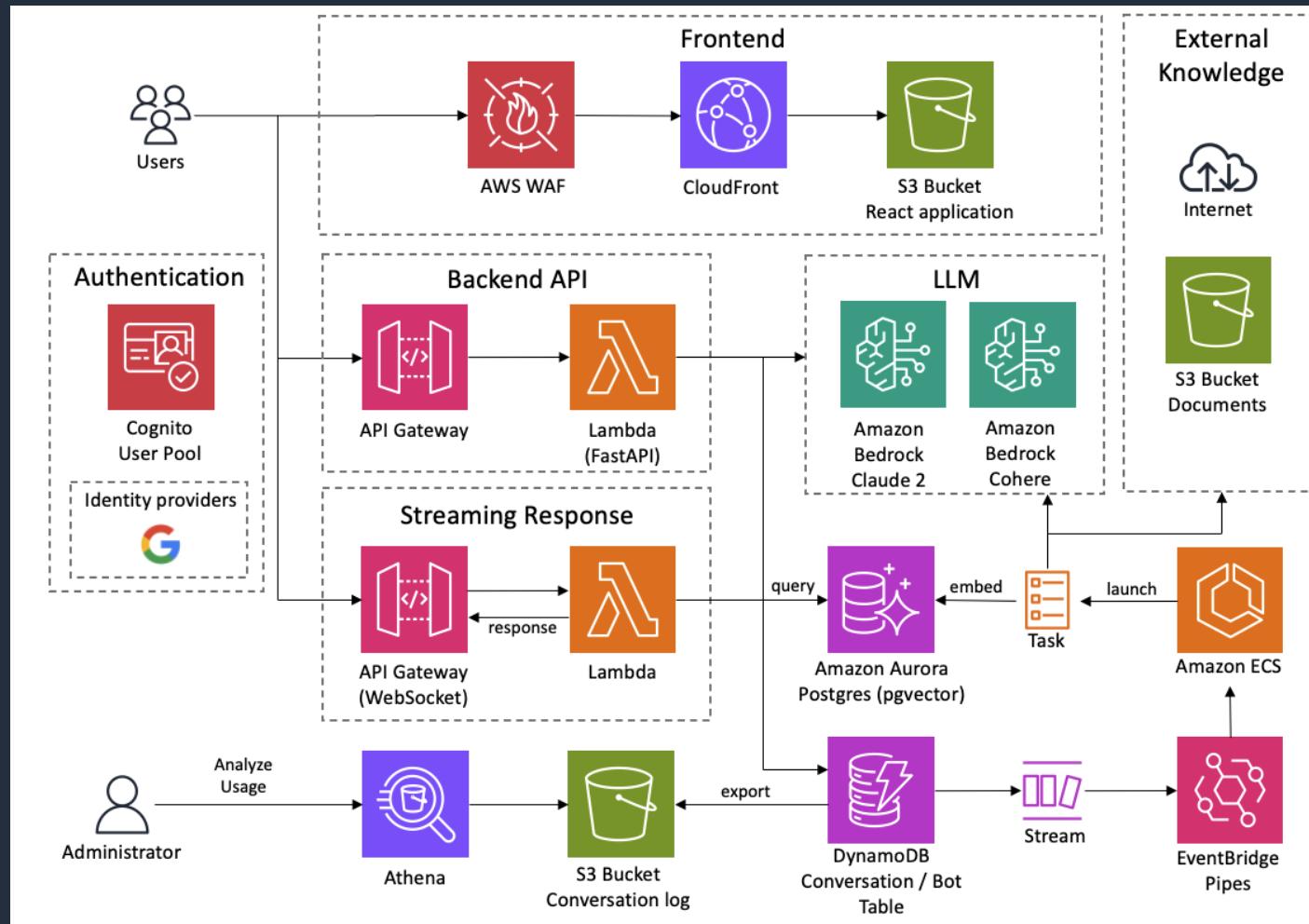
Language	Percentage
TypeScript	61.9%
Python	36.2%
JavaScript	1.2%
Shell	0.4%
Dockerfile	0.2%
HTML	0.1%

<https://github.com/aws-samples/bedrock-claude-chat>



© 2024, Amazon Web Services, Inc. or its affiliates.

# With Enterprise Considerations in Mind



<https://github.com/aws-samples/bedrock-claude-chat/?tab=readme-ov-file>



- 新的聊天
- Bot Console
- 我的最愛 Bots
- 最近用過的 Bots
- demo-bot-mirdc-20241014
- 交談紀錄
- 台北捷運內禁止使用

demo-bot-mirdc-20241014  
沒有描述

同步完成 ⭐ ...  
model: claude-v3-haiku

在台北捷運內可不可以溜直排輪

knowledge\_base\_tool

<result> 根據搜索結果,在台北捷運範圍內不得使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。這是捷運系統的規定,旅客不得在捷運範圍內使用這些器材。主要原因包括:

1. 為了維護捷運系統的安全運作,避免對其他旅客造成危害或干擾。
2. 直排輪等器材可能會對自己或他人的安全造成影響。
3. 在車站或月台上使用這些器材可能會阻礙其他人的通行。
4. 老、幼、重病等需要協助而無人協助的旅客,不得使用這些器材。
5. 攜帶物品可能會造成他人不便。
6. 參與選舉活動、政黨黨務活動、抗議集會或其他政治性議題活動。
7. 其他經本公司公告禁止之事項。

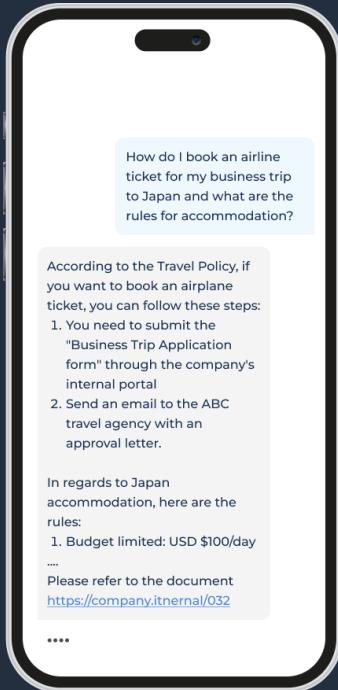
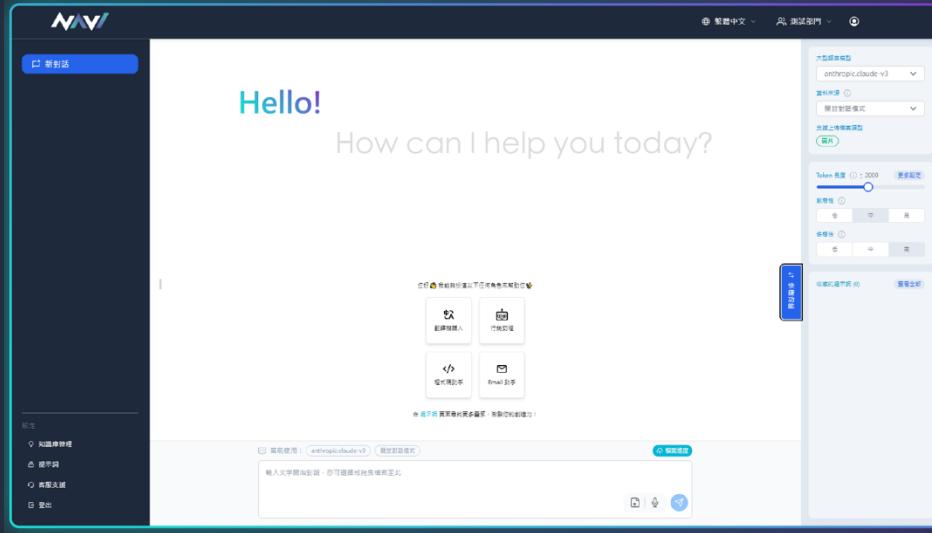
因此,在台北捷運範圍內不得使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。旅客如有違反,除應補繳票價外,並應支付違約金50倍。

</result>

重新生成

請輸入訊息

# Illustrative Vendor Solution



<https://www.netron.net/tw/brand/navi>

<https://www.going.cloud/>

<https://www.no8.io/>





# Thank you!

Michael Lin

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

# Amazon Textract

## Features



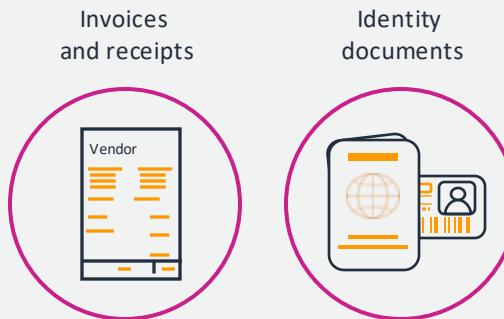
# Amazon Textract - Features



Text



Forms



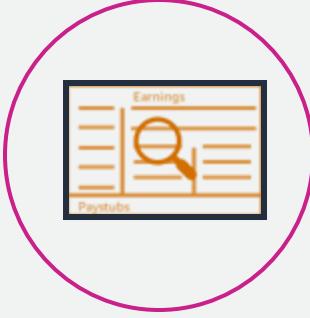
Specialized documents



Handwriting



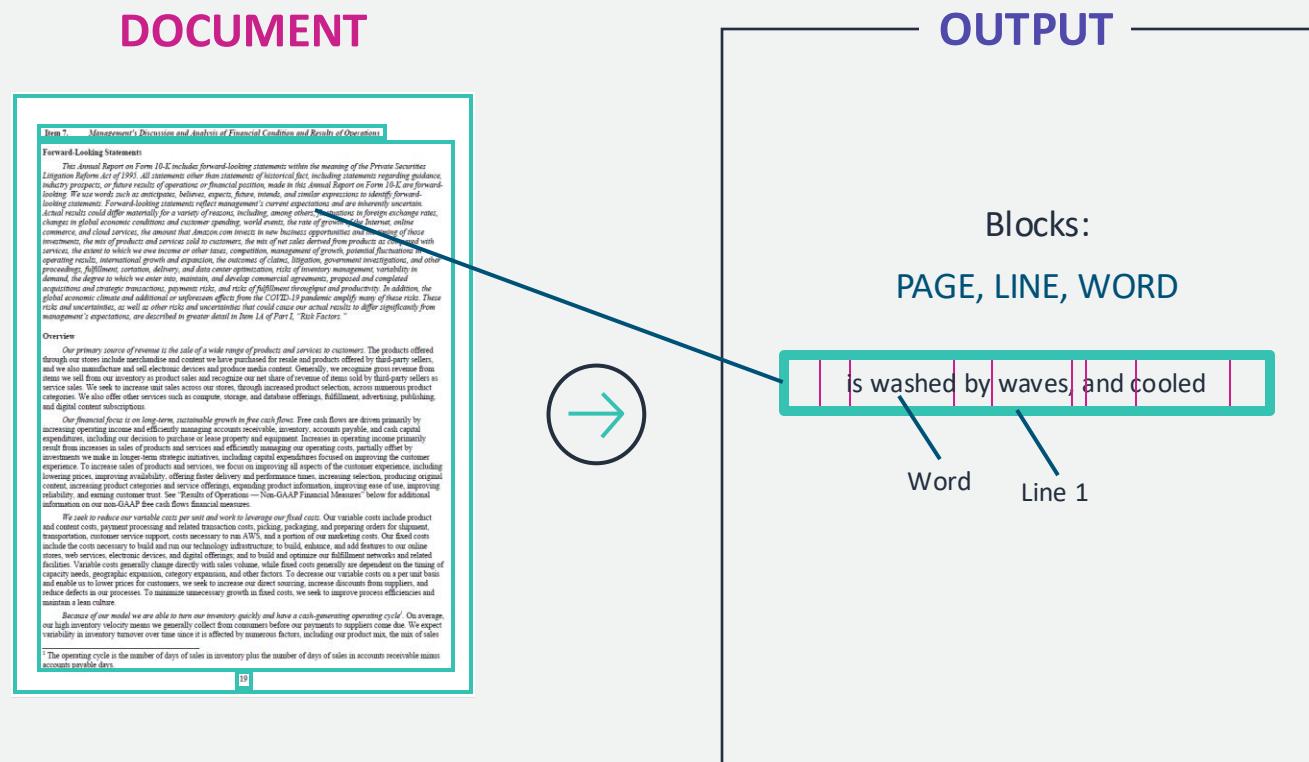
Tables



Queries

# Text extraction

Optimal for dense text extraction with industry leading OCR accuracy



Outputs detected text in 3 hierarchy blocks: Page, Lines and Word

Bounding box for each Line and Word provide visual cues for post-processing

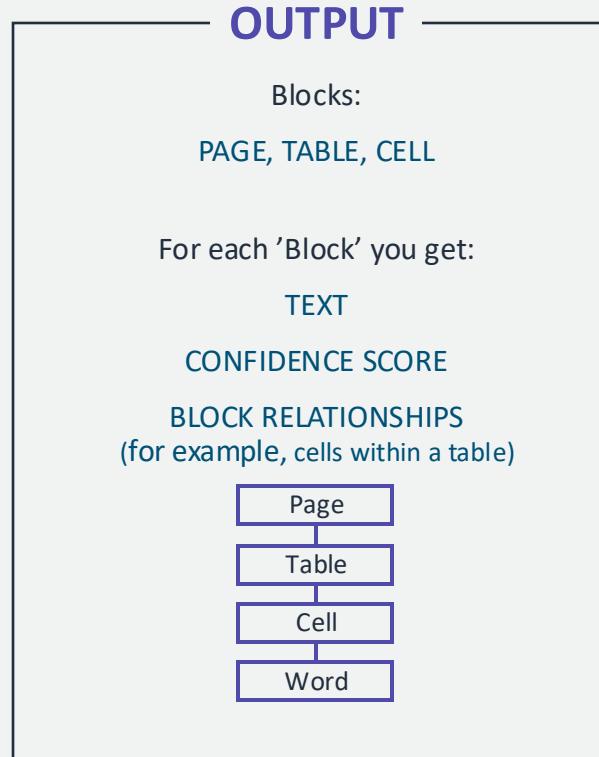
Output includes confidence scores enabling informed decision making for your workflows

# Table extraction

Extract tables from documents while preserving data structure and relationships

**DOCUMENT**

Previous employment history				
Start date	End date	Employer name	Position held	Reason for leaving
1/15/2009	6/30/2013	Any company	Head Baker	Family relocated
8/15/2013	Present	Example corp.	Baker	N/A, current employer



Outputs recognized tables with relationships data intact



Intelligently groups cells within tables and words within each cell



Output includes confidence scores, geometry info, and row/column indexes

# Form extraction

Extract form data from documents as key-value pairs to preserve document structure

**DOCUMENT**

Full Name		
John	X	Doe
First	Middle	Last

Date of Birth		
01	01	1971
MM	DD	YYYY

Gender	
Male	<input checked="" type="radio"/>
Female	<input type="radio"/>



**OUTPUT**

Blocks:  
PAGE, KEY\_VALUE\_SET

Example Output:

First: John
Middle: X
Last: Doe
MM: 01
DD: 01
YYYY: 1971
Male: True
Female: False



Outputs form field name (Key) and field value name (Value) with relationship data intact



Captures logical groupings, relationships, and glyphs



Output includes confidence scores, and geometry info

# Invoices and receipts

Specialized support to process invoices and receipts at scale

## DOCUMENT



## OUTPUT

### Summary Fields:

Vendor Name: WHOLE FOODS MARKET

Subtotal (SUBTOTAL): \$3.50

Net Sales (OTHER): \$3.50

Tax/Fee (TAX): \$0.39

Sold Items (OTHER): 1

Paid (OTHER):

Debit (OTHER): \$3.89

Tax/Fee Total (TAX): \$0.39

Total (TOTAL): \$3.89

### Line Items:

ITEM: Pizza Slice

PRICE: \$3.50



Outputs headline amounts, line item details and inferred fields (like Vendor Name)



Supports any style of invoice or receipt

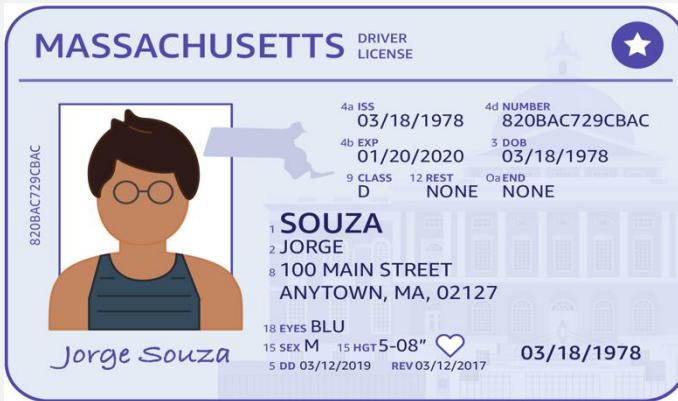


No templates or configuration required

# Identity documents

## Specialized support for identity documents

### DOCUMENT



### OUTPUT

First Name: JORGE  
 Last Name: SOUZA  
 Middle Name:  
 Address Line1: 100 MAIN STREET  
 Address Line 2:  
 City: ANYTOWN  
 State: MA  
 Document Number: 820BAC729CBAC  
 Expiration Date: 01/20/2020  
 Date of Birth: 03/18/1978  
 ID Type: Driver License  
 Date of Issue: 03/18/1978  
 Issued By: MASSACHUSETTS  
 Class: D  
 Restrictions: NONE  
 Endorsements: NONE



95%+ accuracy for US driver licenses and passports



No Templates or Configuration required



Outputs normalized field names and supports implied elements

# Textract Queries

Extract insights from documents using plain language

## Document

CO.	FILE	DEPT.	CLOCK	NUMBER
ABC	126543	123456	12345	0000000 1
<b>Earnings Statement</b>				
ANY COMPANY CORP.			Period ending:	7/18/2008
475 ANY AVENUE			Pay date:	7/25/2008
ANYTOWN, USA 10101				
Social Security Number: 987-65-4321				
Taxable Marital Status: Married				
Exemptions/Allowances:				
Federal: 3, \$25 Additional Tax				
State: 2				
Local: 2				
<b>Earnings</b>				
Regular	rate	hours	this period	year to date
10.00		32.00	320.00	16,640.00
Overtime		1.00	15.00	780.00
Holiday		8.00	80.00	4,160.00
Tuition				1,948.80
				37.43*
				<b>Gross Pay</b>
				<b>\$ 452.43</b>
				23,526.80
<b>Deductions</b>				
Statutory				
Federal Income Tax	- 40.60		2,111.20	
Social Security Tax	- 26.05		1,456.80	
Medicare Tax	- 6.56		341.12	
NY State Income Tax	- 8.43		438.36	
NYC Income Tax	- 5.94		308.88	
NY SUI/SDI Tax	- 0.60		31.20	
				TO \$10.00 PER HOUR.
<b>Other Benefits and Information</b>				
Group Term Life		0.51		27.00
Loan Amt Paid				840.00
Vac Hrs				40.00
Sick Hrs				16.00
Title				Operator
<b>Important Notes</b>				
EFFECTIVE THIS PAY PERIOD YOUR REGULAR HOURLY RATE HAS BEEN CHANGED FROM \$8.00 TO \$10.00 PER HOUR.				

## QUESTIONS

Example questions for a pay stub:

1. What is YTD gross pay?
2. What is gross pay this period?
3. What is employee SSN?
4. How many sick hours did the employee earn?



## OUTPUT

1. What is YTD gross pay?  
\$23,526.80
2. What is gross pay this period?  
\$452.43
3. What is employee SSN? 987-65-4321
4. How many sick hours did the employee earn? 16.00

*No model training or keyword lists required.*



Simple natural language Q&A interface



Extract specific information of interest from documents



Easy integration (Part of existing APIs)



# Textract Queries

Context from implied and nested fields helps to provide answers

Borrower		III. BORROWER INFORMATION				Co-Borrower	
Borrower's Name (include Jr. or Sr. if applicable)						Co-Borrower's Name (include Jr. or Sr. if applicable)	
Social Security Number	Home Phone (incl. area code)	DOB (mm/dd/yyyy)	Yrs. School	Social Security Number	Home Phone (incl. area code)	DOB (mm/dd/yyyy)	Yrs. School
		1/1/1900	18			2/2/1902	18
<input checked="" type="checkbox"/> Married <input type="checkbox"/> Separated	<input type="checkbox"/> Unmarried (include single, divorced, widowed)	Dependents (not listed by Co-Borrower) no.   ages		<input checked="" type="checkbox"/> Married <input type="checkbox"/> Separated	<input type="checkbox"/> Unmarried (include single, divorced, widowed)	Dependents (not listed by Borrower) no.   ages	



## OUTPUT

1. What is the borrower's date of birth? 1/1/1900
2. What is the co-borrower's date of birth? 2/2/1902