AWS Generative AI

Beginner Workshop

Andrew Brown
CEO
ExamPro

Du'An Lightfoot
Sr. Developer Advocate
AWS

Aaron Brighton
Sr. Startup SA
AWS

Agenda

Generative AI overview

Getting started with Amazon Bedrock

Prompt engineering techniques

Comparing large language models

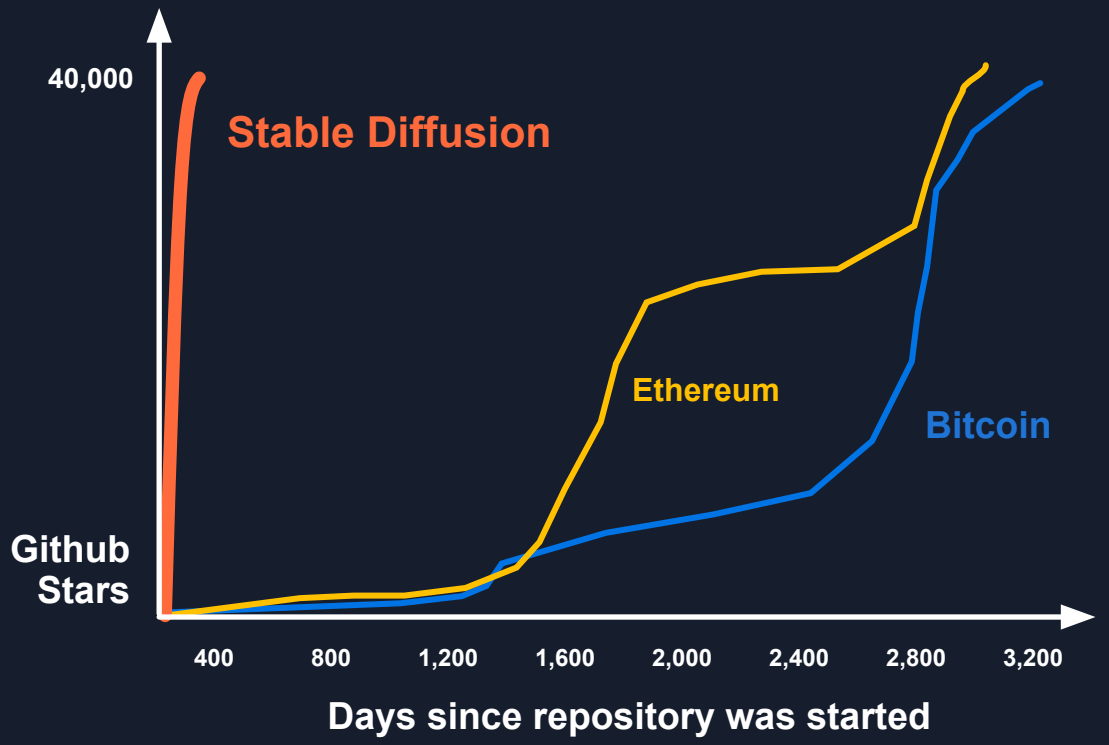
Working with tools

Generative AI overview

Generative AI is the fastest growing trend in AI

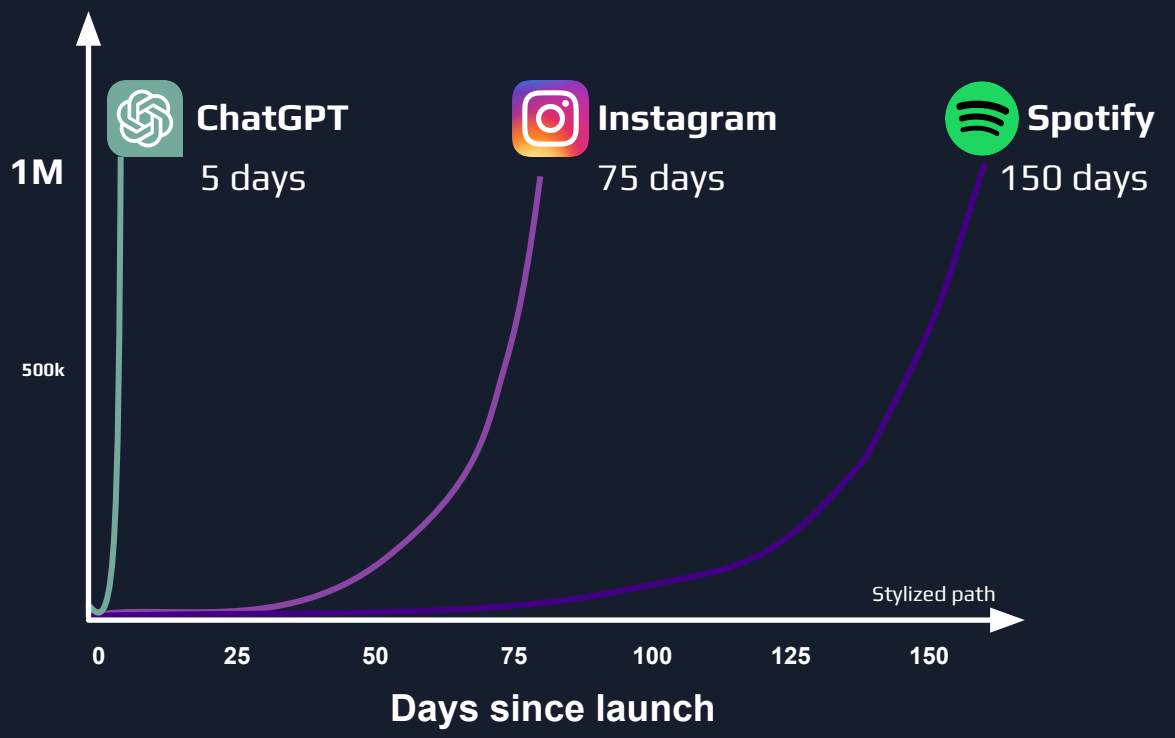
Developer adoption

Stable Diffusion accumulated 40k stars on GitHub in its first 90 days



Consumer adoption

ChatGPT reached the 1 million users mark in just 5 days



What is generative AI?

Working with generative AI

- Text, image, other media, and multi-modal models
- Summarization, analogies, translation and localization, personalization, with long memory and conversational capabilities

A study found office workers using AI assistant Claude were 35.7% more productive, produced higher quality work and made more confident decisi to Claude's data analysis and ins Researchers concluded integrati generative AI like Claude could g boost workplace efficiency and

Un estudio descubrió que los trabajadores de oficina que utilizaban el asistente de IA Claude eran un 35,7% más productivos, producían trabajo de mayor calidad y tomaban decisiones más seguras gracias al análisis de datos y las perspectivas de Claude. Los investigadores concluyeron que integrar la IA generativa como Claude podría aumentar enormemente la eficiencia y la toma de decisiones en el gar de trabajo

Image Examples

IMAGE GENERATION, TRANSFORMATION, UPSCALING

Generated by Stable Diffusion 2.0

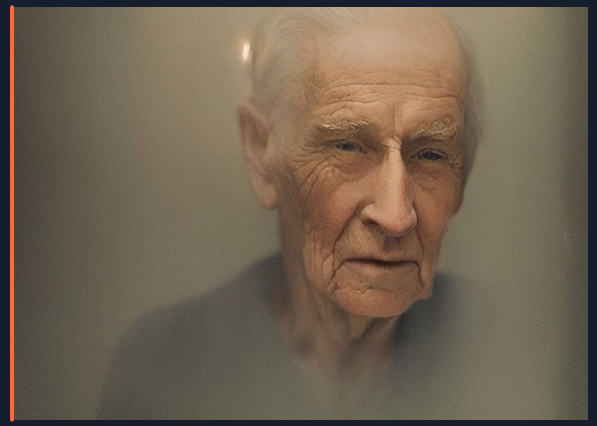


Image transformation

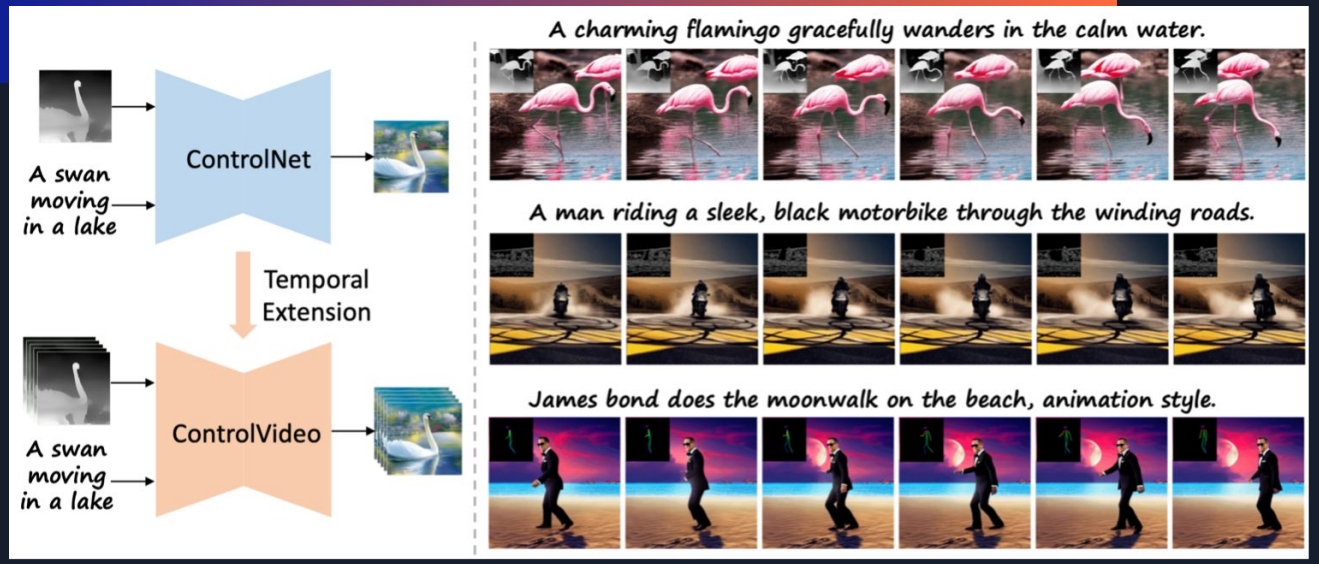
4x

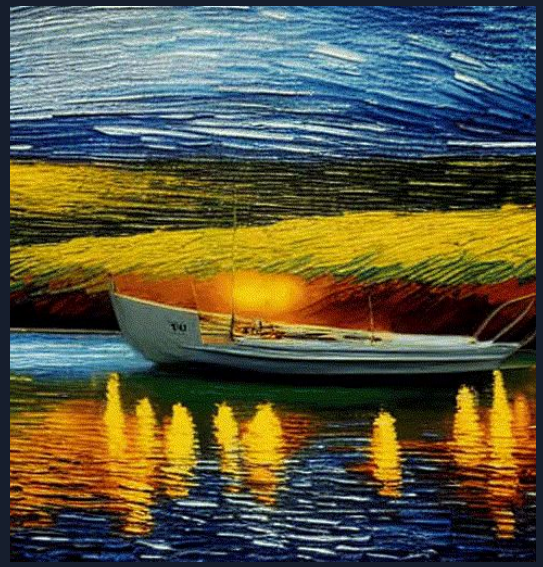
Upscaling

Generating new video content

Open Source project example

ControlVideo (May 2023)

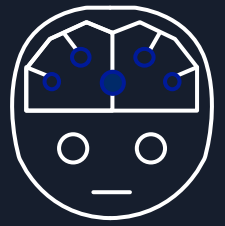




Prompt: "A sleek boat glides effortlessly through the shimmering river, Van Gogh style"

Prompt: "James Bond moonwalk on the beach, animation style"

Where does Generative AI fit?



Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning

Machine learning (ML)

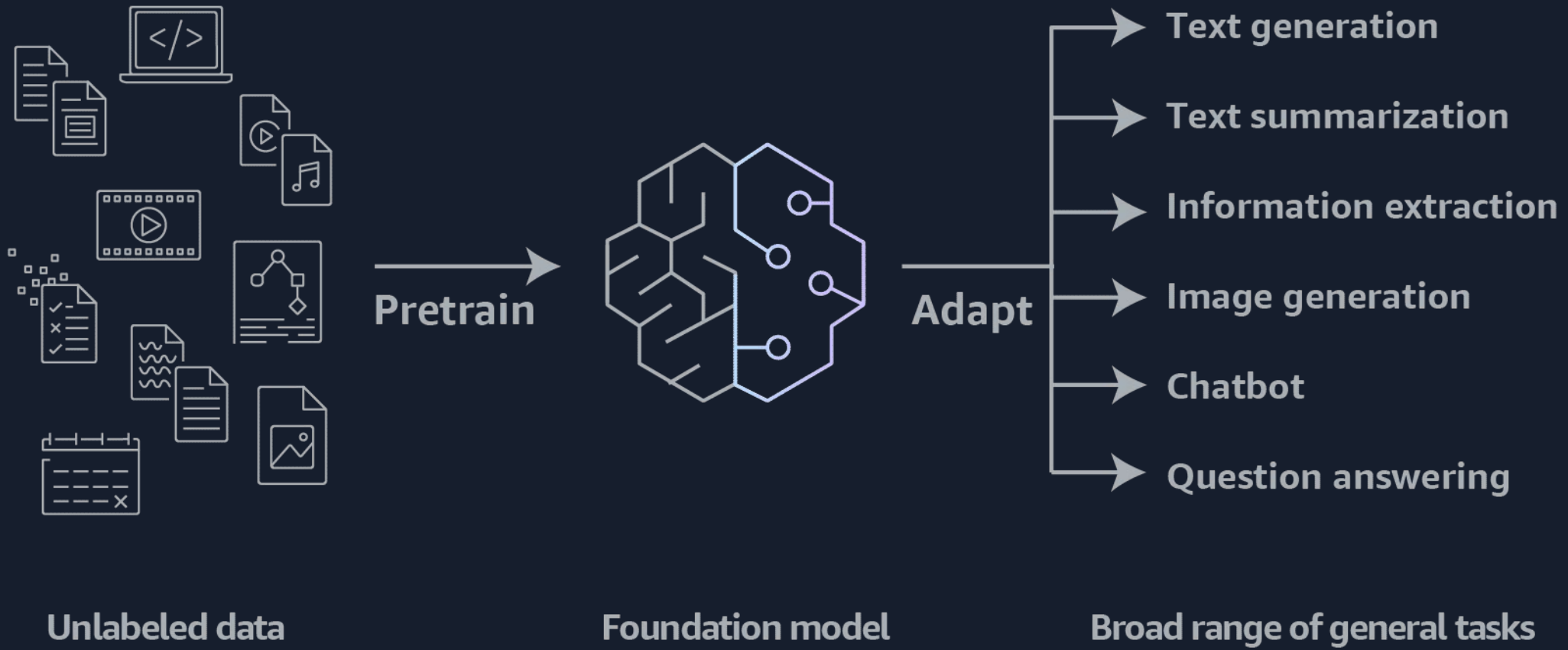
A subset of AI that uses machines to search for patterns in data to build logic models automatically

Deep learning (DL)

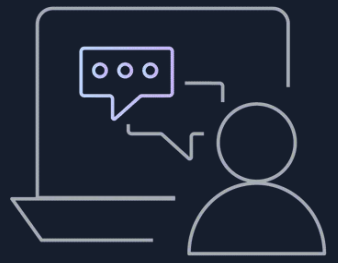
A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition

Generative AI

How does a foundation model function?

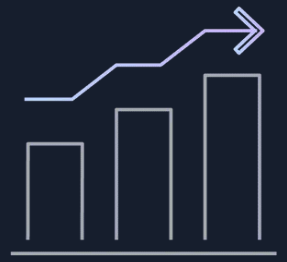


LLM use cases



Improves customer experience

- Chatbots
- Call analytics
- Agent assist



Boosts employee productivity

- Conversational assist
- Code generation
- Automated report generation

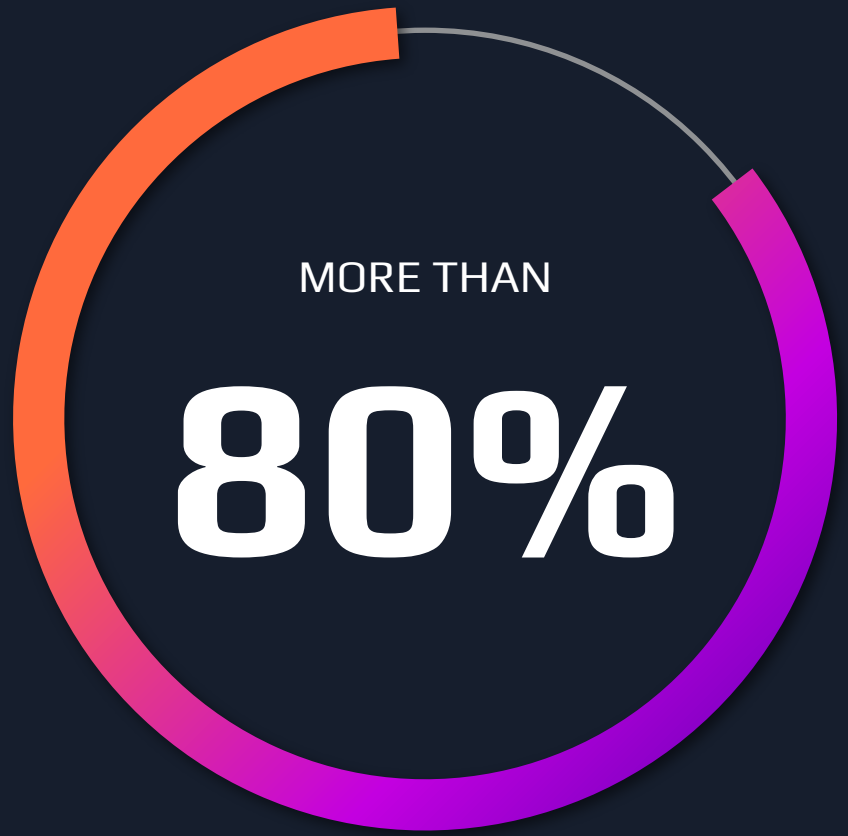
Enhances creativity and content creation

- Marketing
- Sales
- Product development
- Media and entertainment
- News generation

Accelerates process optimization

- Document processing
- Fraud detection
- Supply-chain optimization

Getting started with Amazon Bedrock



ACCORDING TO GARTNER, INC.®

# of enterprises will have used generative AI APIs or deployed generative AI-enabled apps by 2026[1]

[1] Gartner, “More than 80% of Enterprises,” October 11, 2023.

AWS offers a full generative AI stack of tools and services

APPLICATIONS THAT USE LLMs AND OTHER FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

- GPUs
- AWS Trainium
- AWS Inferentia
- Amazon SageMaker
- Amazon EC2 UltraClusters
- Elastic Fabric Adapter (EFA)
- Amazon EC2 Capacity Blocks
- AWS Nitro
- AWS Neuron

AWS offers a full generative AI stack of tools and services

APPLICATIONS THAT USE LLMs AND OTHER FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

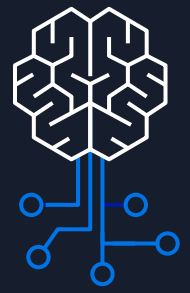
- GPUs
- AWS Trainium
- AWS Inferentia
- Amazon SageMaker
- Amazon EC2 UltraClusters
- Elastic Fabric Adapter (EFA)
- Amazon EC2 Capacity Blocks
- AWS Nitro
- AWS Neuron

Amazon Bedrock

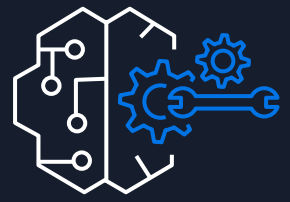
The easiest way to build and scale generative AI applications with foundation models (FMs)

- 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
simplifies

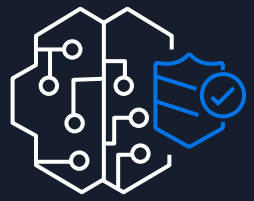


Choice



Customization

Integration



Security and governance

How Amazon Bedrock works

Amazon Bedrock

Build generative AI applications using FMs through a serverless API service

Choose a FM

Use the playground to experiment with FMs and select the one that suits your needs

Use as is or customize

Fine-tune FMs as needed; Bedrock will automatically deploy the FM for inference

Send prompt

Use Bedrock API to send your prompts to the model

Receive response

Receive model response in your application

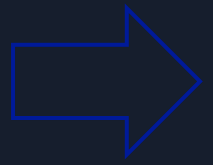
Single API to build with generative AI

Bedrock core API: InvokeModel

- Pass the model ID, type of content, and body of the request
  - Body includes the prompt and execution parameters
  - Returns model response and metadata
- Handles text-to-text, text-to-image, image-to-image, and more
- Supports current and future Amazon Titan models, third-party models, and even fine-tuned models

Bedrock core API: InvokeModel

```
bedrock.invoke_model
(
      modelId =
model_id,
      contentType =
"…",
      accept = "…",
      body = body)
```



Access foundation models

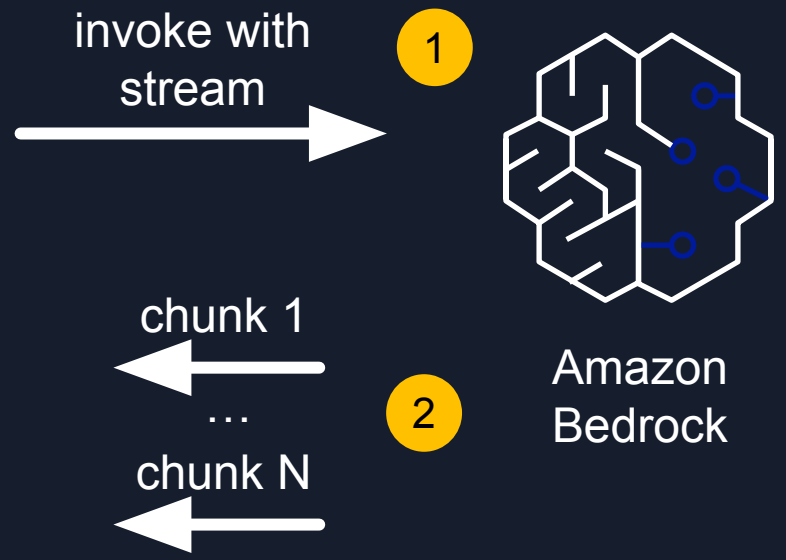
Amazon Titan models

Third-party models

Fine-tuned models NEW

Bedrock core API: Streaming responses

```
response =
bedrock.invoke_model_with_response_strea
m(
        modelId = model_id, body = body)
stream = response.get('body')
if stream:
      for event in stream:
            chunk = event.get('chunk’)
            if chunk:
```




```
print(json.loads(chunk.get('bytes').deco
de()))
```

- Users can start reading the response as soon as the first chunk is available
- Initially supported for Amazon Titan models; Claude and J2 models coming soon

API operations: `invoke_model()`

PARAMETERS

Parameters	Models				
	Titan Text Large	Claude	Command	Jurassic	Stable Diffusion
Temperature	✅	✅	✅	✅	
TopP	✅	✅	✅	✅	
StopSequences	✅	✅	✅	✅	
MaxTokens	✅	✅	✅	✅	
TopK		✅	✅		
CountPenalty				✅	
PresencePenalty				✅	
FrequencePenalty				✅	
Return Likelihoods			✅		
stream			✅		
Number of generations			✅		
Prompt strength (cfg_scale)					✅
Generation step					✅
Seed					✅

invoke_model() – Titan Text

```
prompt_data = """Command: Write me a blog about making strong business decisions as a leader."""

config = {"maxTokenCount":512,"StopSequences":[],"temperature":0.5,"topP":0.9}
body = json.dumps({"inputText": prompt_data},"textGenerationConfig": config)
modelId = "amazon.titan-tg1-large"
accept = "*/*"
contentType = "application/json"
response = bedrock_runtime.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
print(response_body.get("results")[0].get("outputText"))
```

invoke_model() – Anthropic Claude

```
prompt_data = """Command: Write me a blog about making strong business decisions as a leader."""

body = json.dumps({"prompt": prompt_data,
                   "max_tokens_to_sample": 500,
                   "temperature": 0.5,
                   "top_k": 250,
                   "top_p": 1,
                   "stop_sequences": ["\n\nHuman:"]
                   })
modelId = "anthropic.claude-instant-v1"
accept = "*/*"
contentType = "application/json"
response = bedrock_runtime.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
print(response_body.get("completion"))
```

The challenge

As AI models evolve, developers face significant hurdles in keeping pace with changes and leveraging multiple models effectively.

1

Model Versioning

Staying current with various models and their APIs requires constant learning and adaptation.

2

Conversation Management

Implementing features like memory for ongoing dialogues adds another layer of complexity.

3

Multi-Model Integration

Coordinating conversations across different AI models presents technical challenges.

4

API Integration

Incorporating external data from APIs to enhance AI responses requires sophisticated orchestration.

Converse API for Bedrock

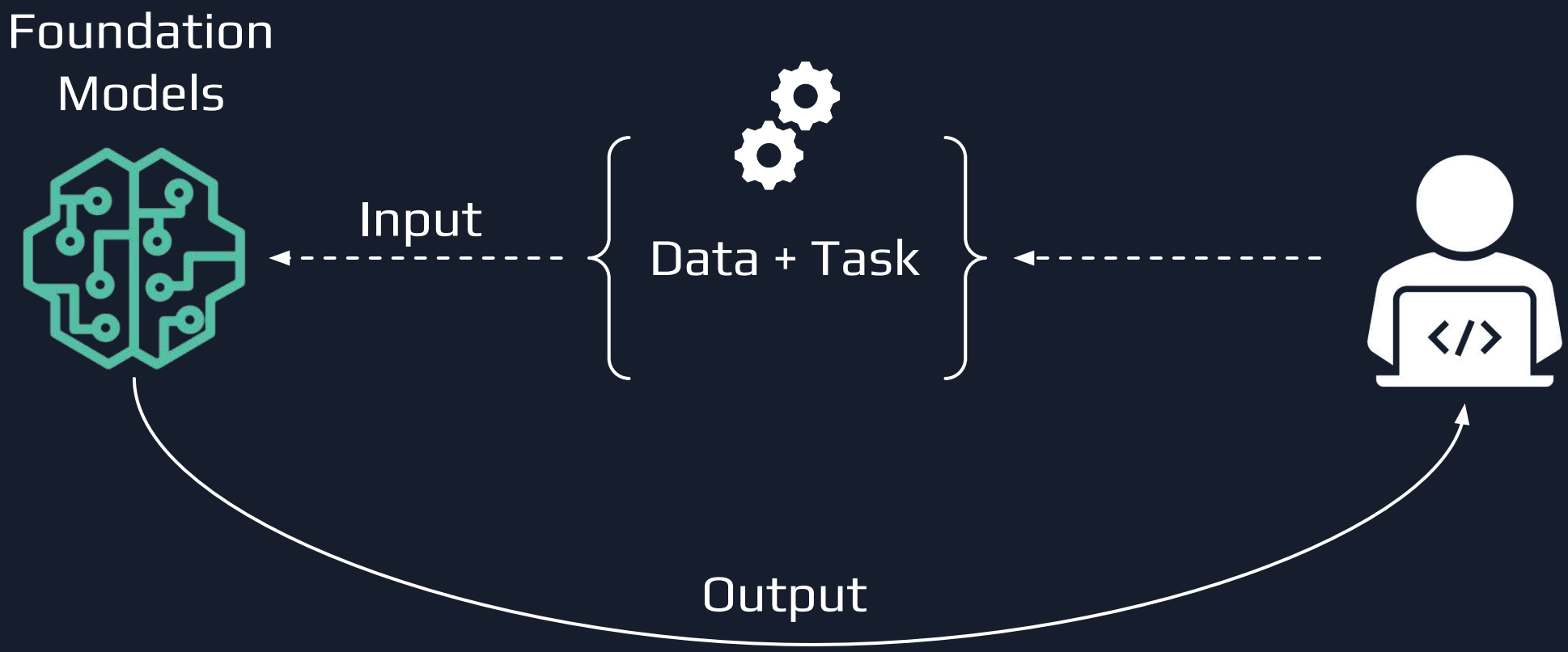
- ✔ New unified message-structured invocations for Bedrock.
- ✔ Same inference parameters and bodies, independently of the model chosen.
- ✔ Bedrock handles basic prompt format translation for system/user/assistant prompts.
- ✔ Consistent output format for all models.
- ✔ Support for native function-calling from Anthropic, Cohere and Mistral via unified Tool configuration.
- ✔ Easy implementations of apps with Bedrock for any supported model.

```
response = bedrock.converse(
  modelId="anthropic.claude-3-sonnet-20240229-v1:0",
  messages=[
    {
      "role": "user",
      "content": [
        {"text": "Is tomorrow raining in Seattle?"}
      ]
    }
  ],
  toolConfig={
    "tools": [
      {
        "function":
          "name": "get_weather",
          "description": "Gets the weather forecast for a
city",
          "inputSchema": {
            "jsonSchema": []
          }
        }
      }
    ]
  }
)
```

Note: The Converse API only supports text generation models. Embeddings models and image generation models still require InvokeModel.

Prompt engineering

Prompt engineering, new way of using ML!



Elements of the prompt example

Instructions and output indicator

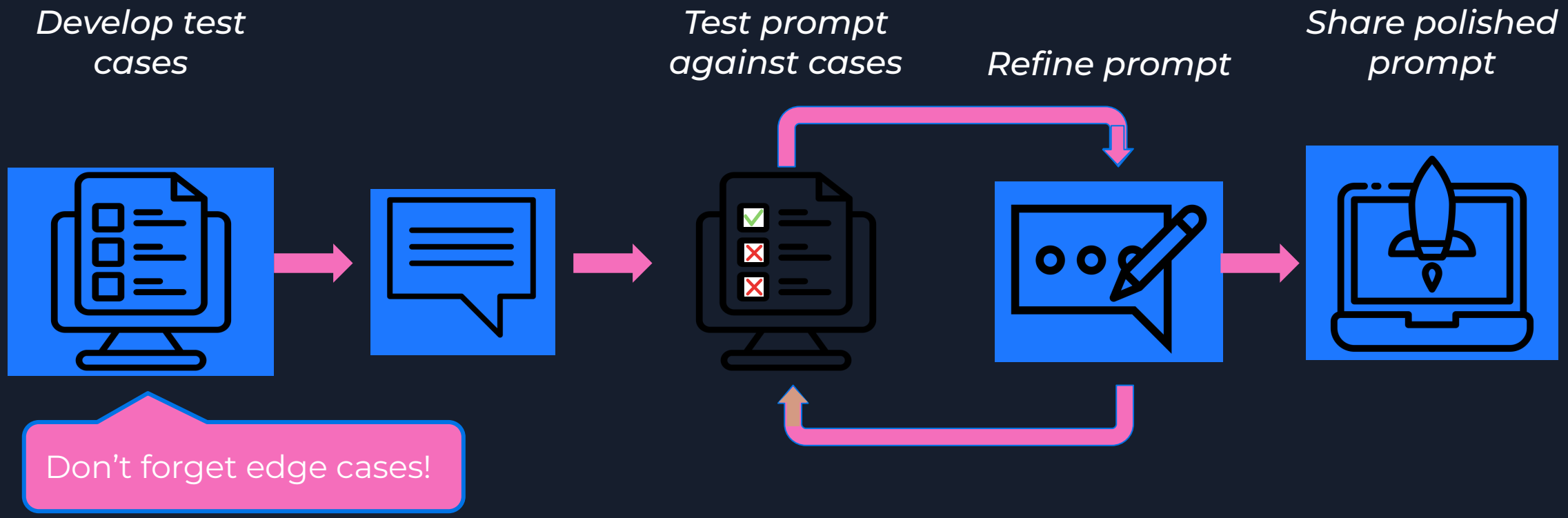
Context

Input data

Prompt	Output
Write a summary of a service review using two sentences. Store: Online Service: Shipping Review: Amazon Prime Student is a great option for students looking to save money. Not paying for shipping is the biggest save in my opinion. As a working mom of three who is also a student, it saves me tons of time with free 2-day shipping, and I get things I need quickly and sometimes as early as the next day, while enjoying all the free streaming services and books that a regular Prime membership has to offer for half the price. Amazon Prime Student is only available for college students, and it offers so many things to help make college life easier. This is why Amazon Prime is the no-brainer that I use to order my school supplies, my clothes, and even to watch movies in between classes. I think Amazon Prime Student is a great investment for all college students. Summary:	Amazon Prime Student is a fantastic option for college students, offering free 2-day shipping, streaming services, books, and other benefits for half the price of a regular Prime membership. It saves time and money, making college life easier.

Prompt engineering philosophy

Empirical science: always test your prompts & iterate often!



Best practices for designing effective prompts

Be clear and concise

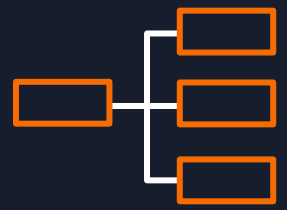
Include context

Use directives

Include output

Start with a question

Provide example responses



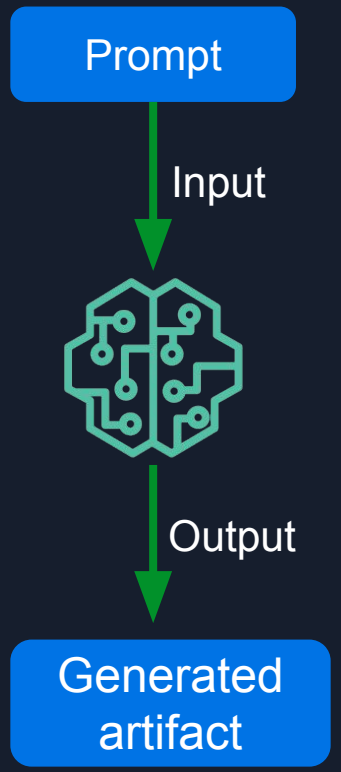
Break up complex tasks

Experiment and be creative

Prompt engineering types

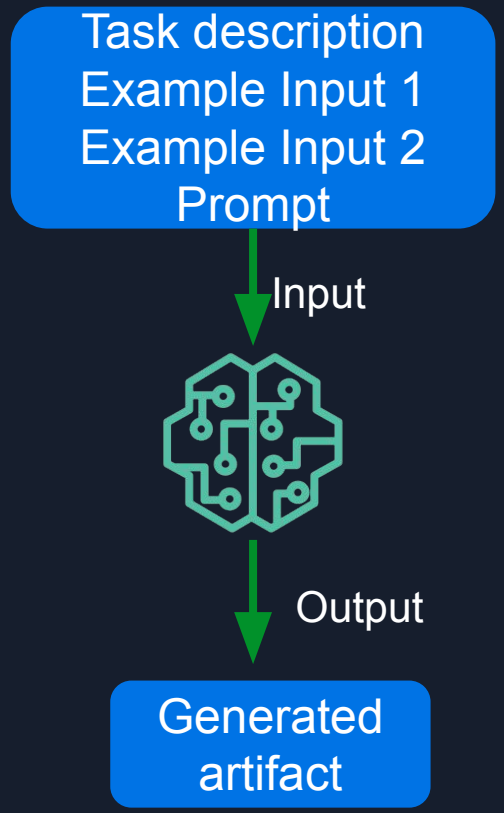
Zero shot prompts

- Direct request with sufficient context



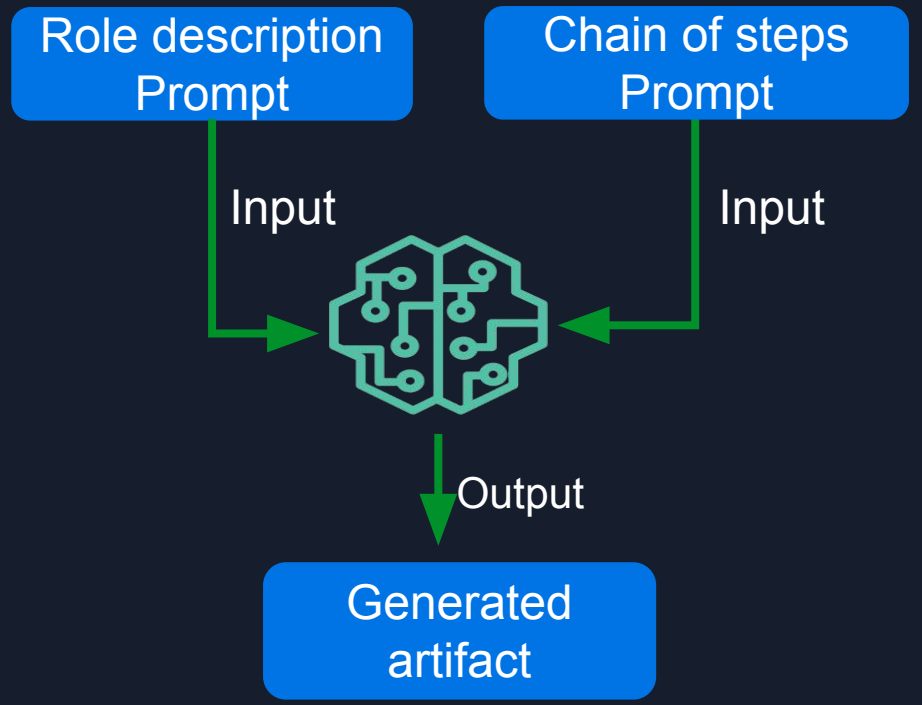
One shot or few shot prompts

- Provide one or more examples with a request

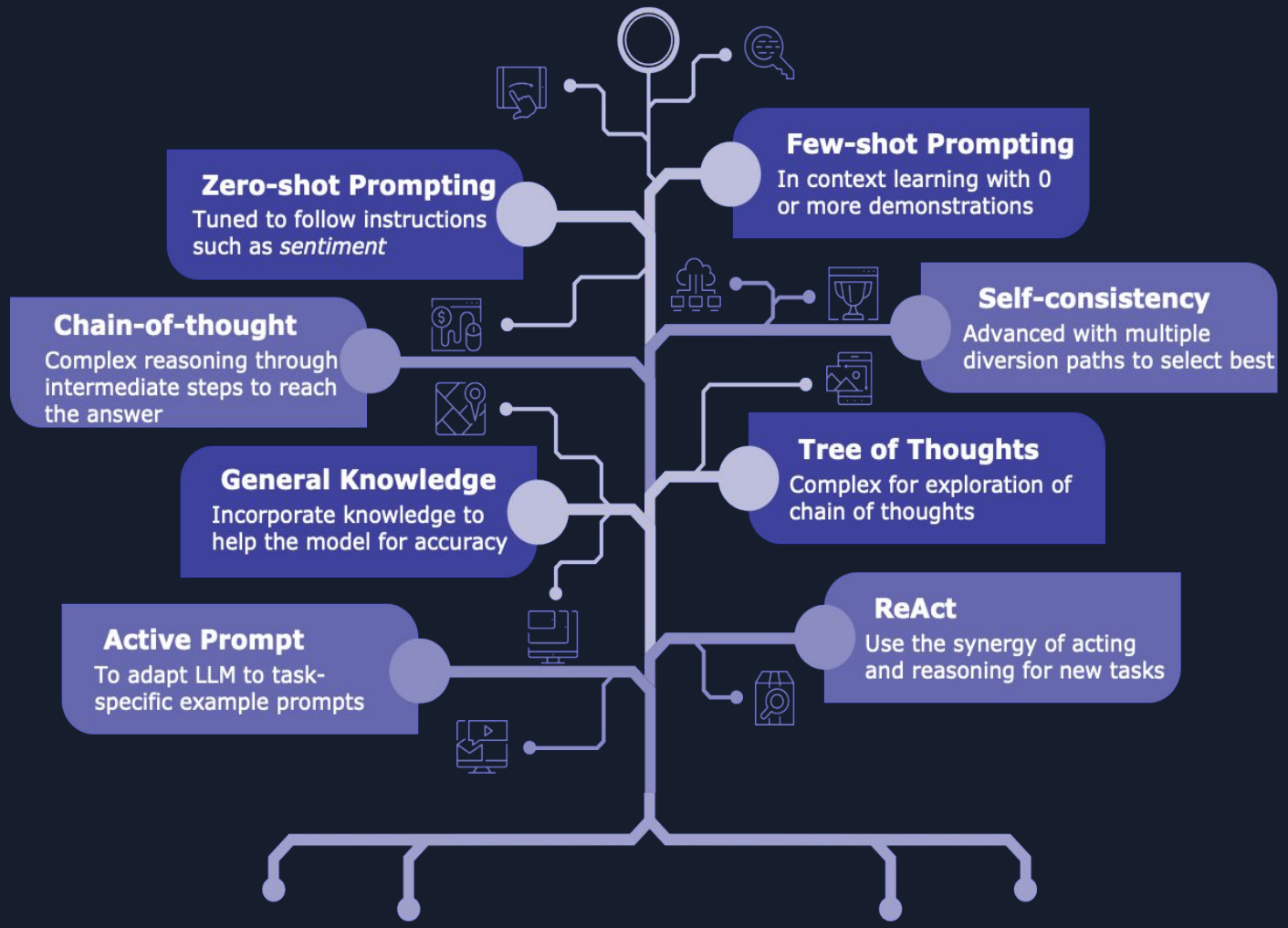


Role or Chain of Thought prompts

- Provide the model with a role or persona for the task
- Provide a chain of steps for the model to follow



Prompting Techniques



Prompt template

Role

Task

Background and objective

Variations

Guidelines

Examples (3-5)

Latest information

Chain of thought

Prompt Engineering

> *you are a personal assistant. you are friendly, polite and casual. you help with...*

> *you are a classifying agent that filters user inputs into categories. your job is to sort these inputs before they are passed along to our function calling agent. the purpose of our function calling agent is to call functions in order to answer user's questions.*

Prompt Engineering

> The user input is between the <question></question> XML tags:
> <question>
> What is the weather in Oslo right now?
> </question>

Prompt engineering techniques

Zero-shot Prompting

Prompt:

```
Classify the text into neutral, negative or positive.
Text: I think the vacation is okay.
Sentiment:
```

Output:

```
Neutral
```

Few-Shot Prompting

Prompt:

```
A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses
the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses
the word farduddle is:
```

Output:

```
When we won the game, we all started to farduddle in celebration.
```

https://www.promptingguide.ai/techniques

Which prompt follows prompting best practices?

INCLUDE CONTEXT IF NEEDED

Prompt 1
Summarize this article: [insert article text]

or

Prompt 2
Provide a summary of this article to be used in a blog post: [insert article text]

Which prompt follows prompting best practices?

INCLUDE CONTEXT IF NEEDED

Prompt 1
Summarize this article: [insert article text]

or

Prompt 2
Provide a summary of this article to be used in a blog post: [insert article text]

Which prompt follows prompting best practices?

PROVIDE AN EXAMPLE RESPONSE

Prompt 1
Determine the sentiment of this social media post: “[insert post]”

or

Prompt 2
Determine the sentiment of the following social media post using these examples: post: “great pen” // Positive post: “I hate when my phone battery dies” // Negative “[insert post]” //

Which prompt follows prompting best practices?

PROVIDE AN EXAMPLE RESPONSE

Prompt 1
Determine the sentiment of this social media post: “[insert post]”

or

Prompt 2
Determine the sentiment of the following social media post using these examples: post: “great pen” // Positive post: “I hate when my phone battery dies” // Negative “[insert post]” //

Comparing LLMs

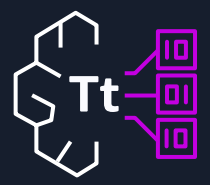
Amazon Bedrock

BROAD CHOICE OF MODELS

AI21 labs	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 3.1	Mistral Large 2 (24.07)	Stable Diffusion XL1.0
Jurassic-2 Ultra	Amazon Titan Text Lite	Claude 3 Opus	Command Light	Llama 3 8B	Mistral Large (24.02)	Stable Diffusion XL 0.8
Jurassic-2 Mid	Amazon Titan Text Express	Claude 3 Sonnet	Embed English	Llama 3 70B	Mistral Small	
	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed Multilingual	Llama 2 13B	Mixtral 8x7B	
	Amazon Titan Text Embeddings V2	Claude 2.1	Command R+	Llama 2 70B	Mistral 7B	
	Amazon Titan Multimodal Embeddings	Claude 2	Command R			
	Amazon Titan Image Generator	Claude Instant				

Amazon Titan

FMS



Amazon Titan Text Embeddings

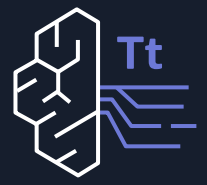
Numerical representations of text

Amazon Titan Text Lite

Summarization, copywriting, fine-tuning

Amazon Titan Text Express

Open-ended text generation, conversational chat, RAG support

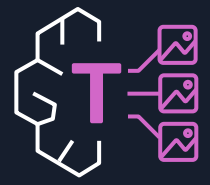


Amazon Titan Text Premier

Enterprise-grade text generation, optimized for RAG and Agents

Amazon Titan Multimodal Embeddings

Search, recommendation, personalization

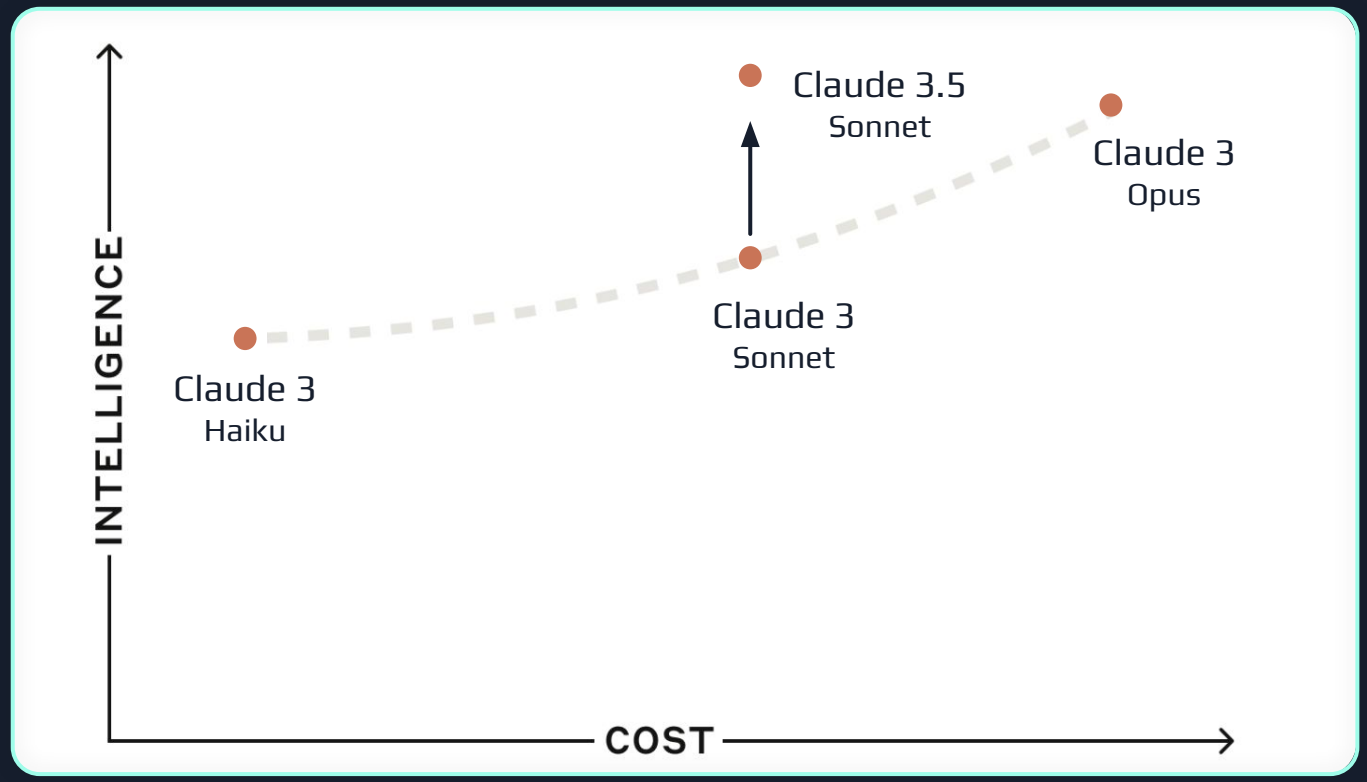


Amazon Titan Image Generator

Realistic, studio-quality images

FOUNDATION MODELS CRAFTED FOR THE ENTERPRISE

Anthropic's Claude Models



The Claude model family

Claude 3.5 Sonnet — Anthropic's most intelligent model to date, priced for high-volume use cases

Claude 3 Opus — Powerful model for complex tasks that excels in sight-unseen scenarios

Claude 3 Sonnet — Balance of speed and performance and intelligence

Claude 3 Haiku — Fastest, most cost-effective model preferred for scaled tasks

Using tools

Tool Use/Function Calling

User prompt

```
What's the weather in San Francisco?
```

List of functions in API (tools)

```
name : get_weather
description: Get the current weather in a
given location
Required: city, state

name : top_song
description: Get the most popular song played
on a radio station
Required: title, artist
```

Chosen tool & input parameters

```
Name: get_weather
Parameters
  city=San Francisco
  state=California
```

Amazon Bedrock Converse API

MODELS THAT SUPPORT TOOL USE

Model	User model/chat	System prompts	Vision	Tool use	Streaming tool use
Amazon Titan	Yes	No	No	No	No
Anthropic Claude 2 and earlier	Yes	Yes	No	No	No
Anthropic Claude 3	Yes	Yes	Yes	**Yes**	**Yes**
Cohere Command R and Command R+	Yes	Yes	No	**Yes**	No
Meta Llama 2 and Llama 3	Yes	Yes	No	No	No
Mistral AI Instruct	Yes	No	No	No	No
Mistral AI Large	Yes	Yes	No	**Yes**	No
Mistral AI Small	Yes	Yes	No	No	No
AI21 Labs Jurassic-2 (Text)	Limited. No chat support.	No	No	No	No
Cohere Command (Text)	Limited. No chat support.	No	No	No	No

Model	Tool choice	Tool use system prompt token count
Claude 3 Opus	`auto`	530 tokens
	`any`, `tool`	281 tokens
Claude 3 Sonnet	`auto`	159 tokens
	`any`, `tool`	235 tokens
Claude 3 Haiku	`auto`	264 tokens
	`any`, `tool`	340 tokens

https://docs.anthropic.com/en/docs/tool-use-pricing-and-tokens

Workshop

Wifi:
Username: **Guest**
Password: **BrokenWires@@2019**

https://github.com/
ExamProCo/gen-ai-training-day-workshops

https://bit.ly/3UGLKhH

Thank you!

Please complete the session survey in the mobile app

Andrew Brown
@exampro

Du'An Lightfoot
@labeveryday

Aaron Brighton
LinkedIn: aaronbrighton