Generative AI with AWS

John Fan
Account Manager, Education
johnfan@amazon.com

Agenda

- Introduction to Generative AI
- Innovation opportunities
- Generative AI with AWS
- Introduction to PartyRock

Innovation can **transform industries**

GENERATIVE AI

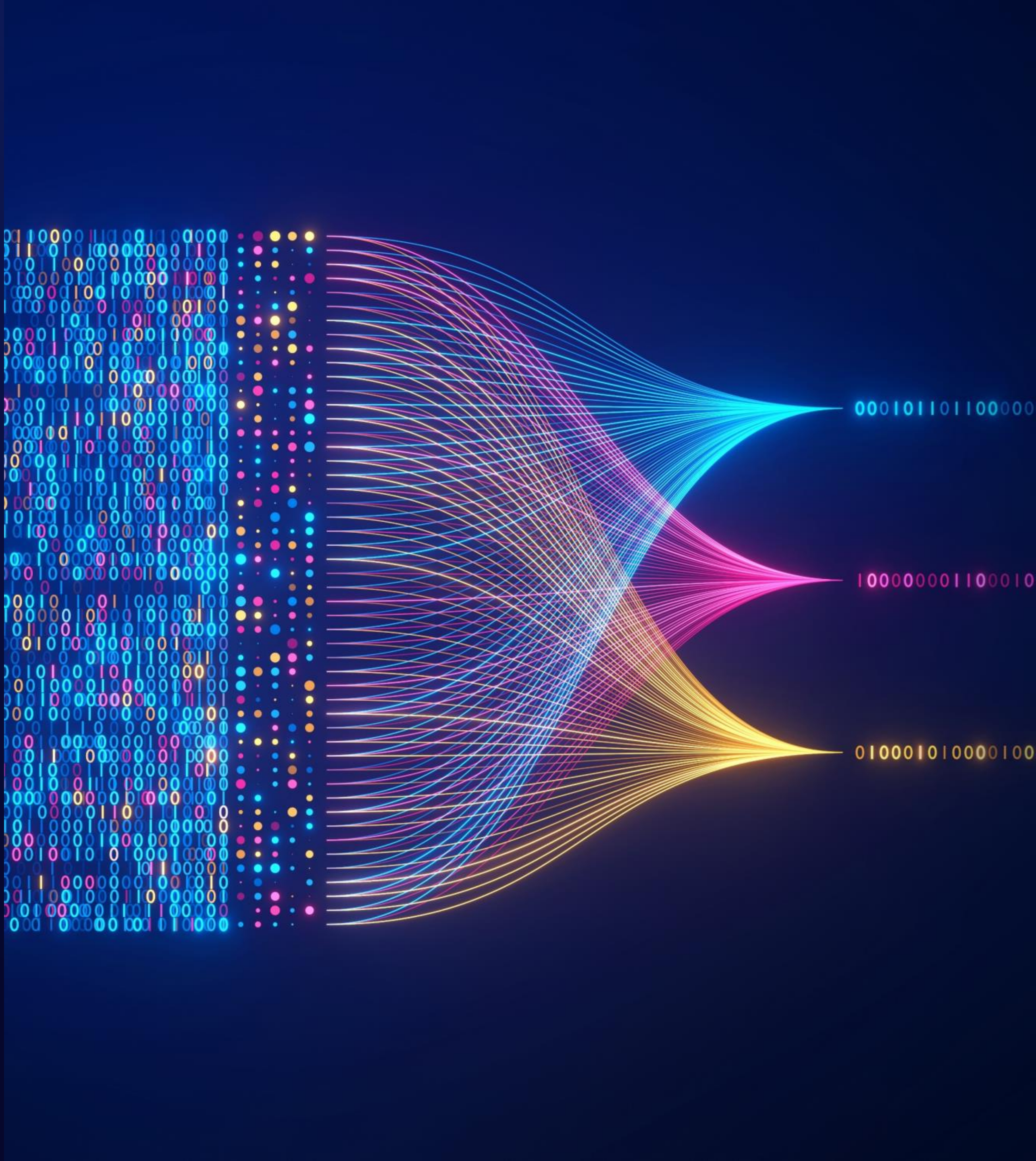
Generative AI is powered by foundation models

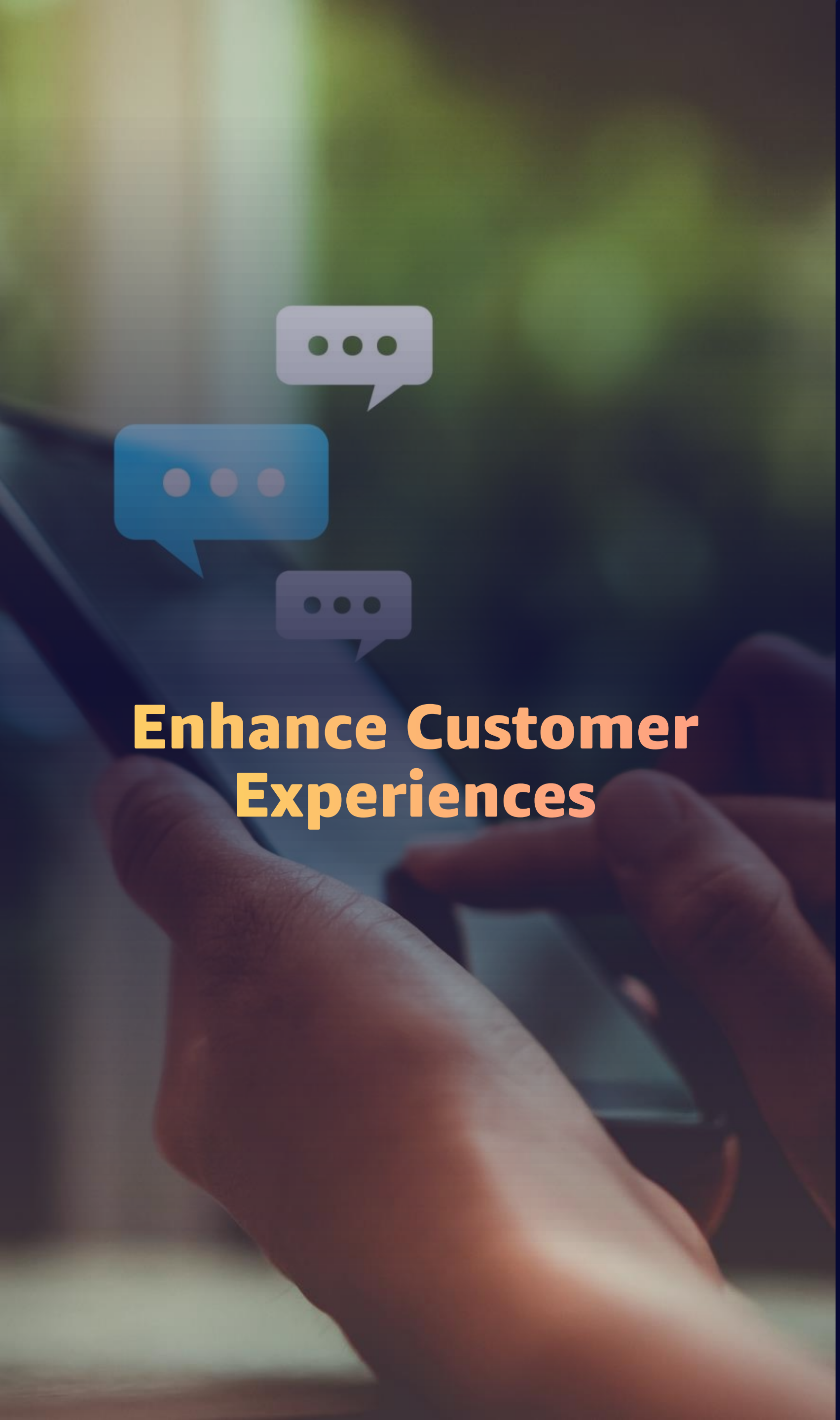
Pretrained on vast amounts of unstructured data

Contain large number of parameters that make them capable of learning complex concepts

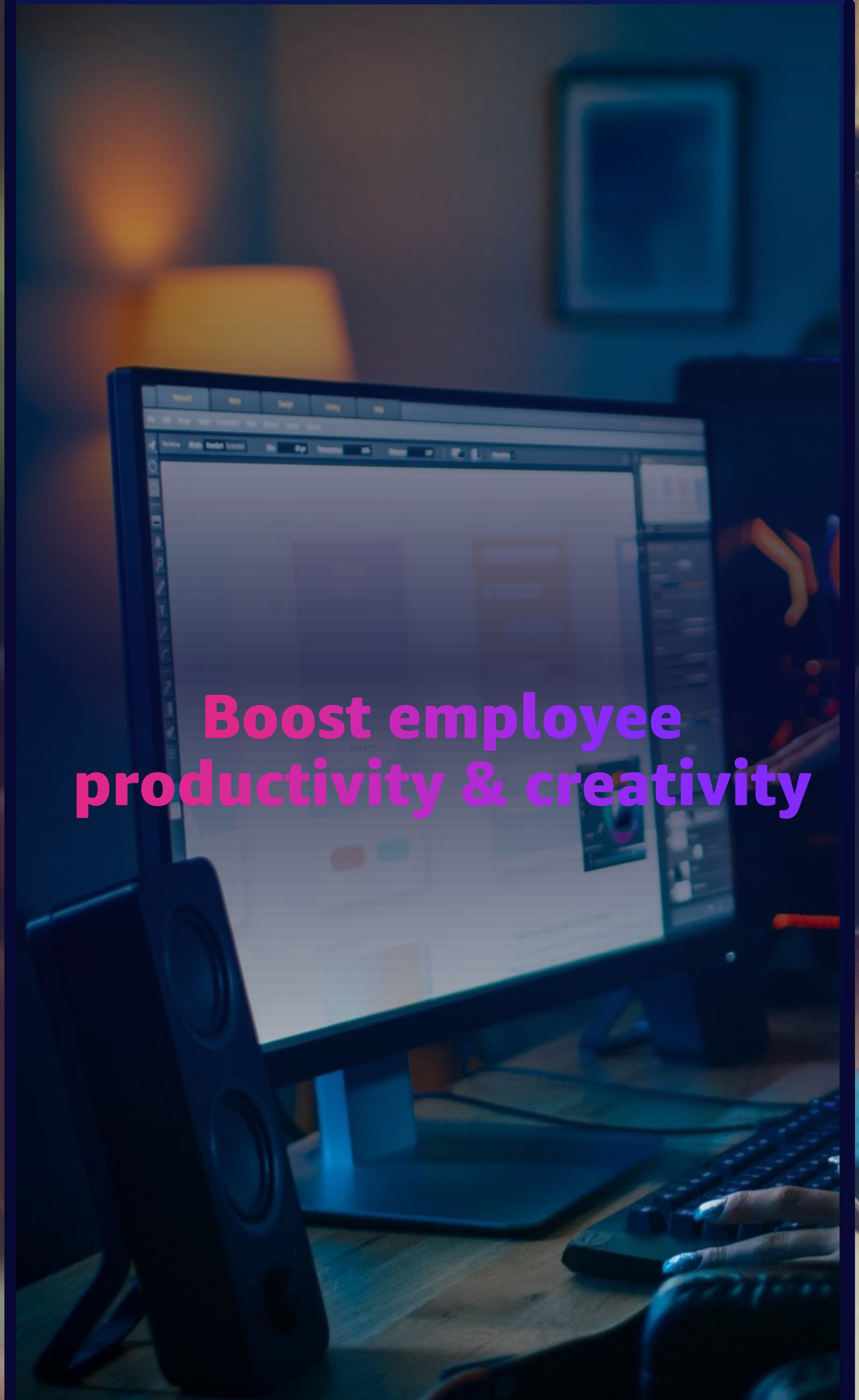
Can be applied in a wide range of contexts

Customize FMs using your data for domain specific tasks



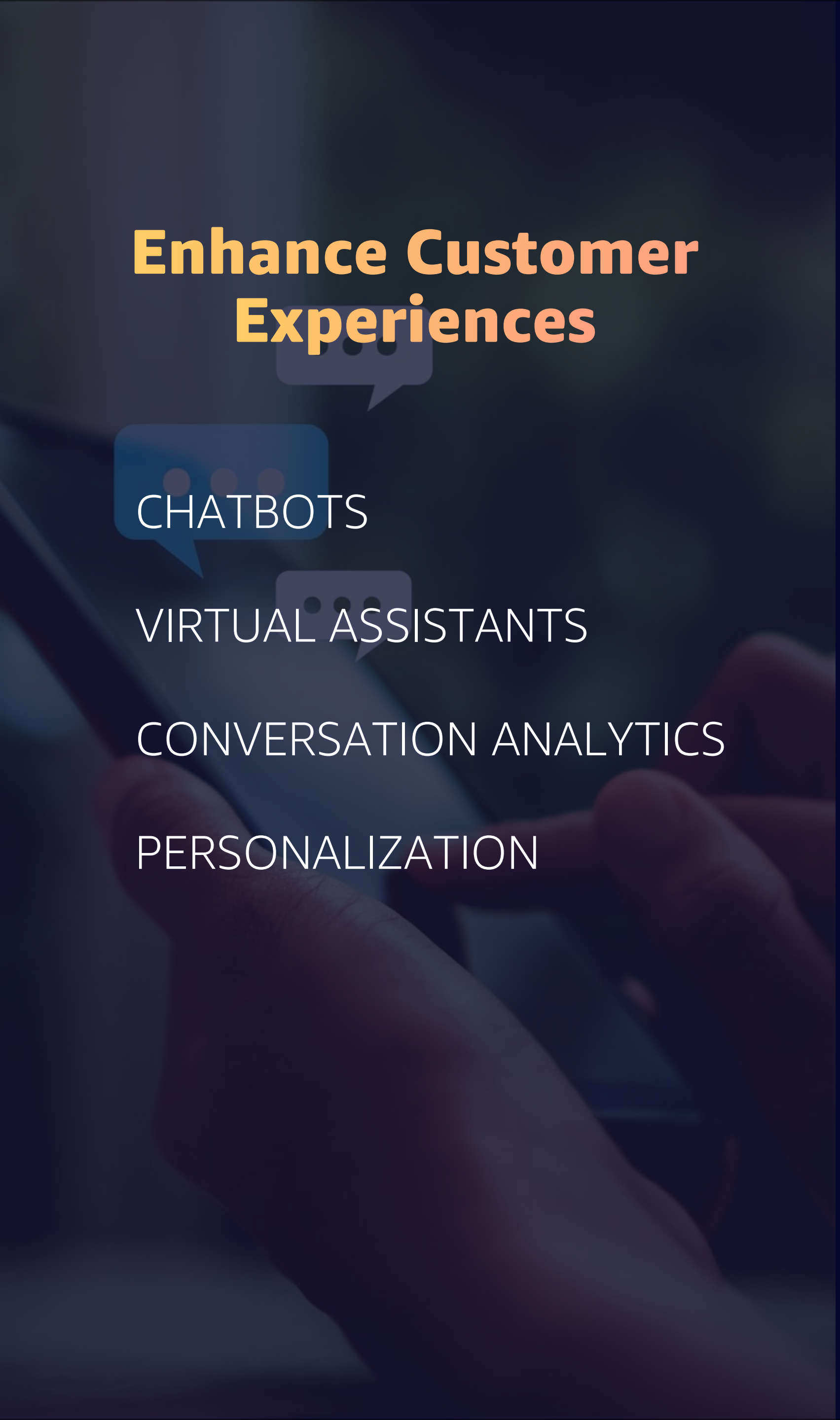


Enhance Customer Experiences



Boost employee productivity & creativity

Optimize business processes



Enhance Customer Experiences

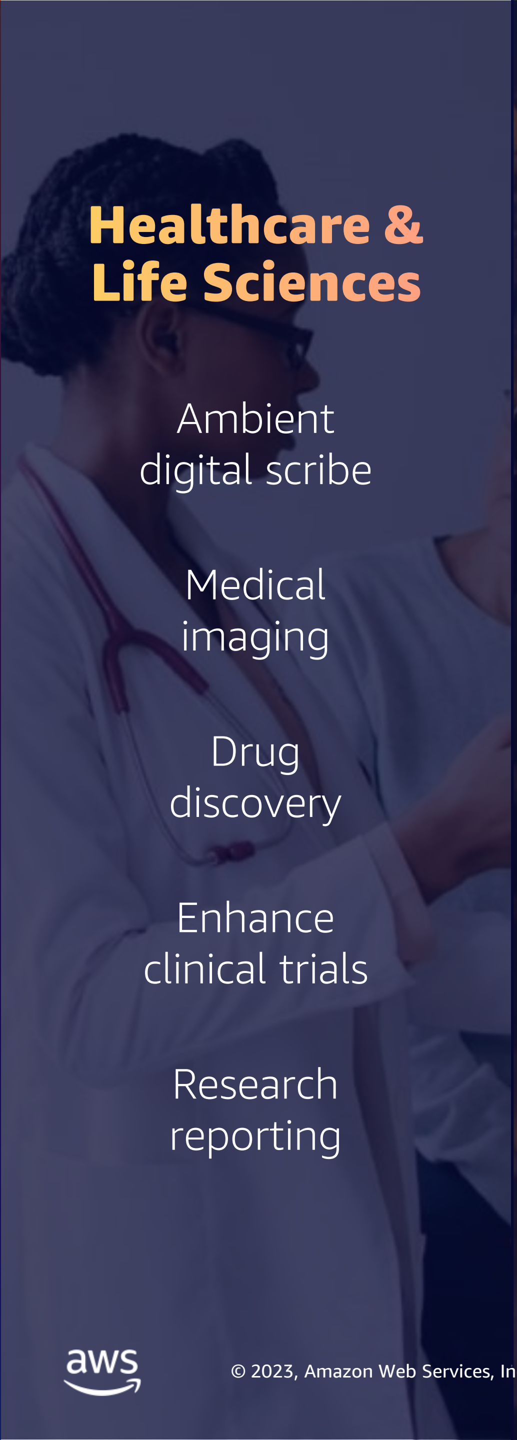
- CHATBOTS
- VIRTUAL ASSISTANTS
- CONVERSATION ANALYTICS
- PERSONALIZATION

Boost employee productivity & creativity

- CONVERSATIONAL SEARCH
- SUMMARIZATION
- CONTENT CREATION
- CODE GENERATION
- DATA TO INSIGHTS

Optimize business processes

- DOCUMENT PROCESSING
- DATA AUGMENTATION
- CYBERSECURITY
- PROCESS OPTIMIZATION

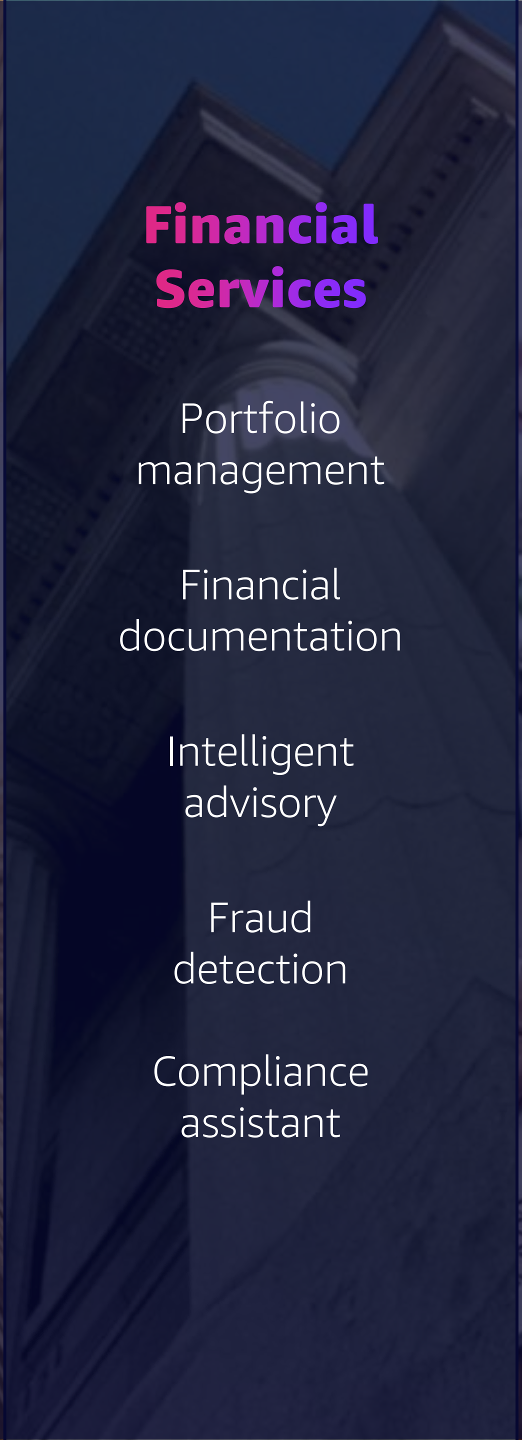


Healthcare & Life Sciences

- Ambient digital scribe
- Medical imaging
- Drug discovery
- Enhance clinical trials
- Research reporting

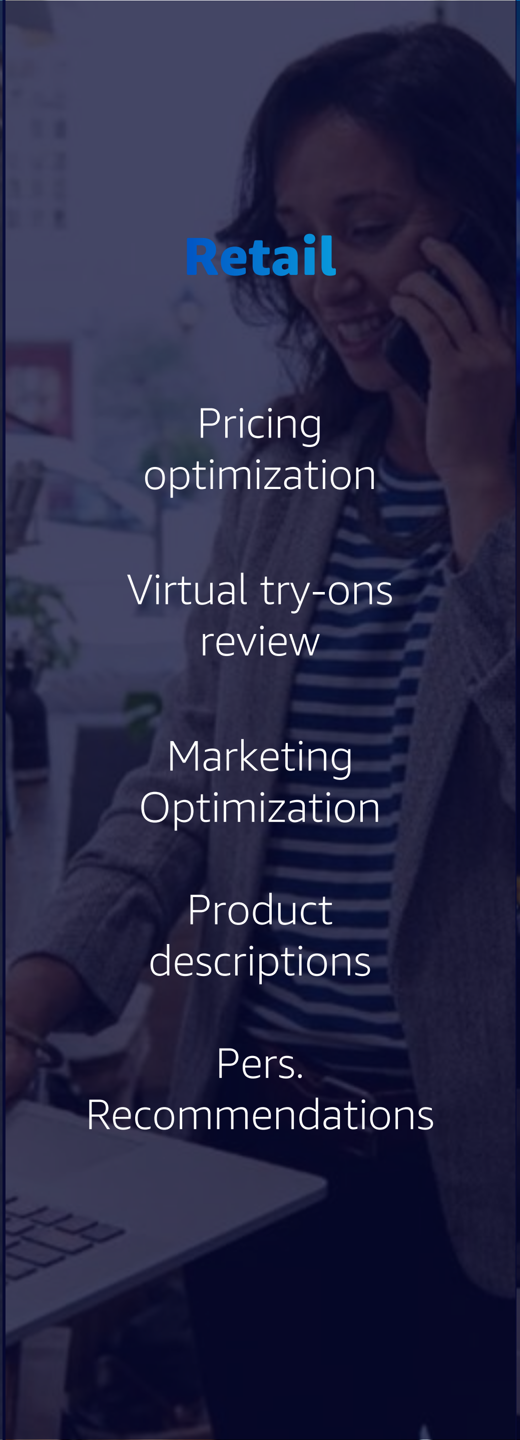
Industrial & Manufacturing

- Product design
- Operational efficiency
- Maintenance Assistants
- Supply chain optimization
- Equipment diagnostics



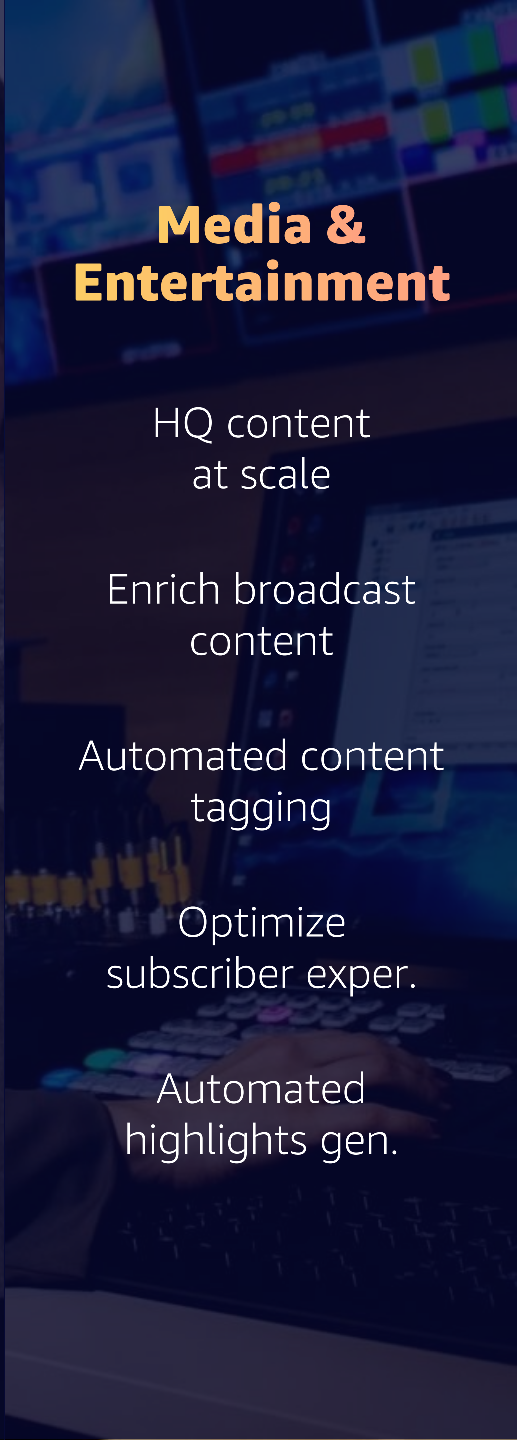
Financial Services

- Portfolio management
- Financial documentation
- Intelligent advisory
- Fraud detection
- Compliance assistant



Retail

- Pricing optimization
- Virtual try-ons review
- Marketing Optimization
- Product descriptions
- Pers. Recommendations



Media & Entertainment

- HQ content at scale
- Enrich broadcast content
- Automated content tagging
- Optimize subscriber exper.
- Automated highlights gen.

For **you.**

LET'S START

CREATING TOGETHER

Reinventing with generative AI

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

Amazon Q Business | Amazon Q Developer | Amazon Q in QuickSight | Amazon Q in Connect | PartyRock

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

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 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

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

Amazon Bedrock

BROADEST SELECTION OF FULLY MANAGED MODELS FROM LEADING AI COMPANIES

AI21labs

Highly efficient processing & grounded generation for long context lengths

JAMBA

amazon

Frontier intelligence & industry leading price performance

NOVA

ANTHROP\C

Excels at complex reasoning, code generation, & instruction following

CLAUDE

cohere

Powering efficient, multilingual AI agents with advanced search & retrieval

COMMAND

EMBED

RERANK

deepseek

Advanced reasoning models that solve complex problems step-by-step

DEEPSEEK-R1

Luma

High-quality video generation with natural, coherent motion & ultra-realistic details

RAY2

Meta

Advanced image & language reasoning

LLAMA

MISTRAL AI_

Specialized expert models for agentic reasoning and multimodal tasks

MISTRAL

MIXTRAL

PIXTRAL

poolside (Coming soon)

Software engineering AI for large enterprises

MALIBU

POINT

stability.ai

Professional-grade images with creative control, deployable at scale

STABLE DIFFUSION

TwelveLabs (Coming soon)

CTRL + F for video data: unlock the full potential of enterprise video assets

MARENGO

PEGASUS

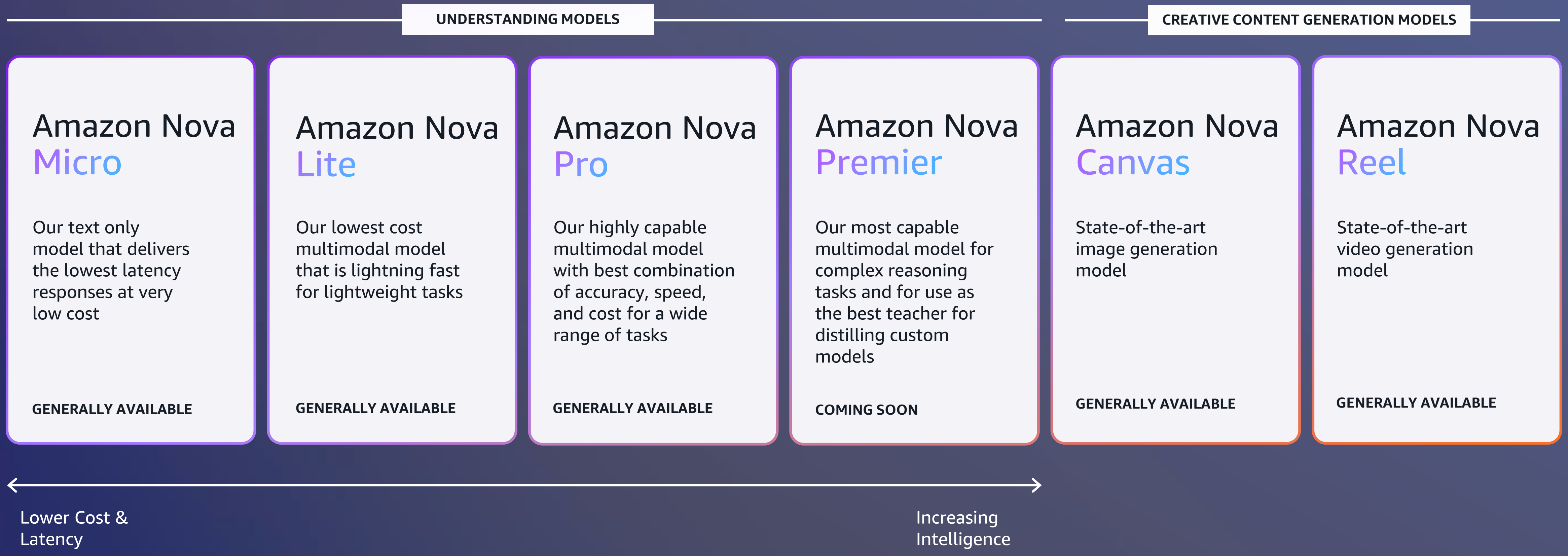
WRITER

Purpose-built models for building & scaling AI agents across the enterprise

PALMYRA

Amazon Bedrock is the only service with industry leading Nova models

Amazon Nova models deliver frontier intelligence and industry leading price performance



UNDERSTANDING MODELS

CREATIVE CONTENT GENERATION MODELS

Amazon Nova Micro

Our text only model that delivers the lowest latency responses at very low cost

GENERALLY AVAILABLE

Amazon Nova Lite

Our lowest cost multimodal model that is lightning fast for lightweight tasks

GENERALLY AVAILABLE

Amazon Nova Pro

Our highly capable multimodal model with best combination of accuracy, speed, and cost for a wide range of tasks

GENERALLY AVAILABLE

Amazon Nova Premier

Our most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON

Amazon Nova Canvas

State-of-the-art image generation model

GENERALLY AVAILABLE

Amazon Nova Reel

State-of-the-art video generation model

GENERALLY AVAILABLE

Lower Cost & Latency

Increasing Intelligence

Introduction to PartyRock

What is PartyRock?

PartyRock, an Amazon Bedrock playground, is a shareable generative AI application-building playground

It’s a hands-on, code-free application builder where you can build, share, and remix applications while playing with generative AI

Anyone can access PartyRock through its intuitive web-based UI

What do I need?

→ Development or ML engineering experience

What do I need?

→ ~~Development or ML engineering experience~~

What do I need?

→ ~~Development or ML engineering experience~~

→ **Coding** experience

What do I need?

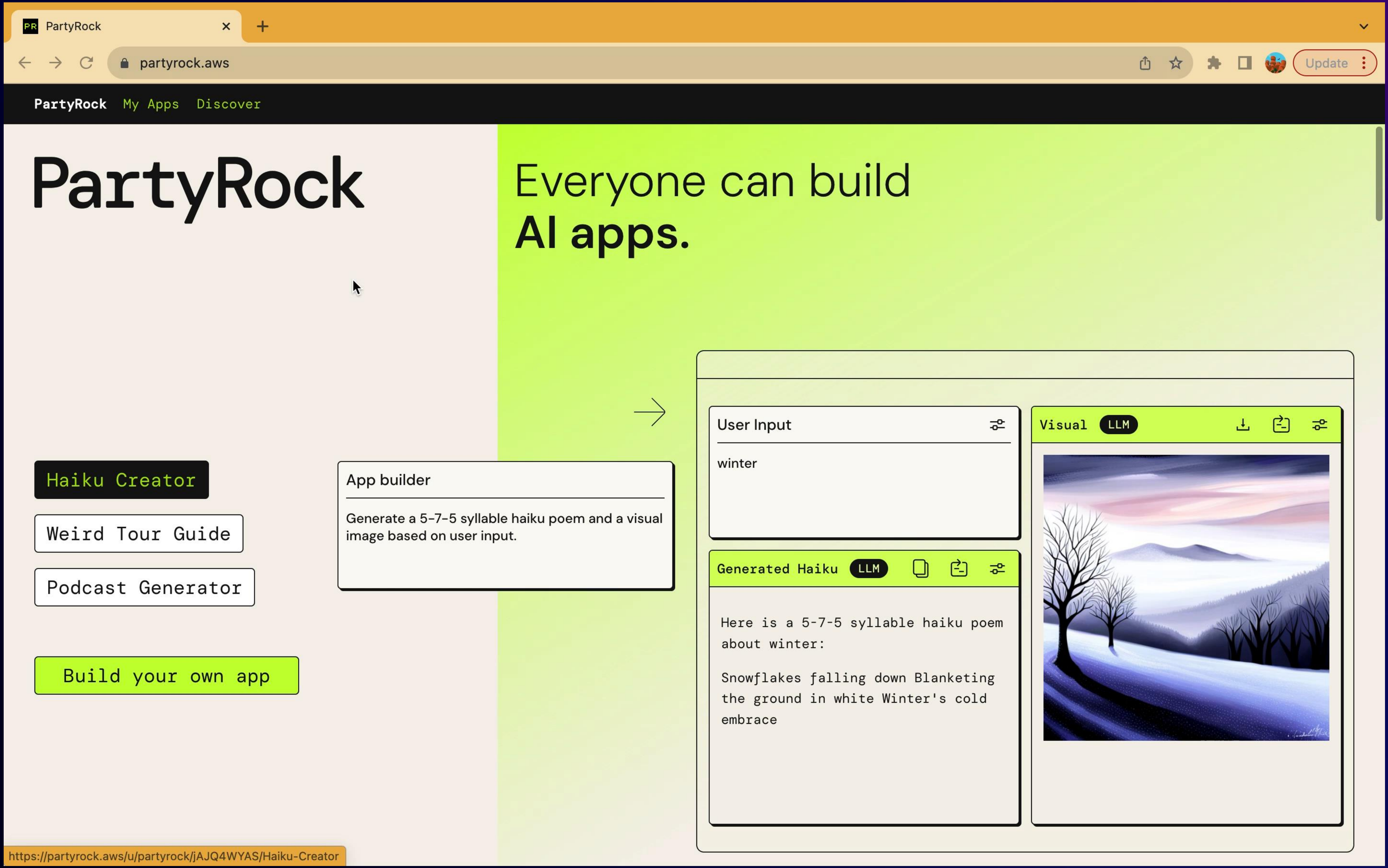
→ ~~Development or ML engineering experience~~

→ ~~**Coding** experience~~

"

Everyone can build **AI apps.**

PartyRock



PartyRock

Everyone can build **AI apps.**

App builder

Generate a 5-7-5 syllable haiku poem and a visual image based on user input.

User Input

winter

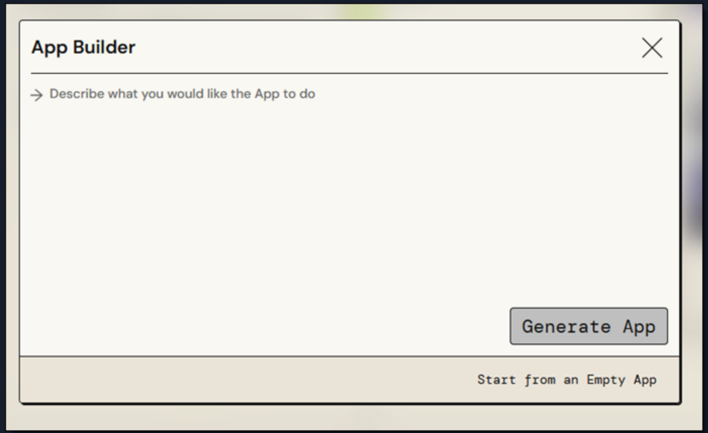
Generated Haiku LLM

Here is a 5-7-5 syllable haiku poem about winter:

Snowflakes falling down Blanketing the ground in white Winter's cold embrace

Build your app

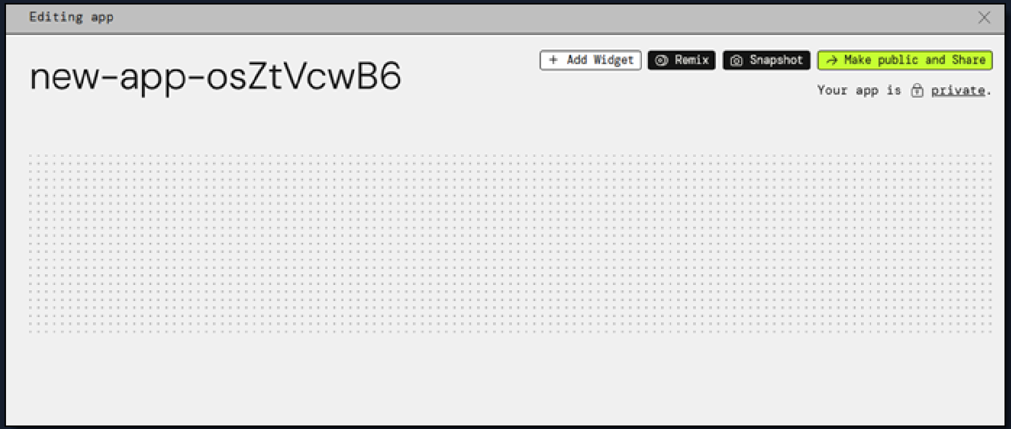
□ Start with a prompt



App Builder

→ Describe what you would like the App to do

□ Start from an empty app

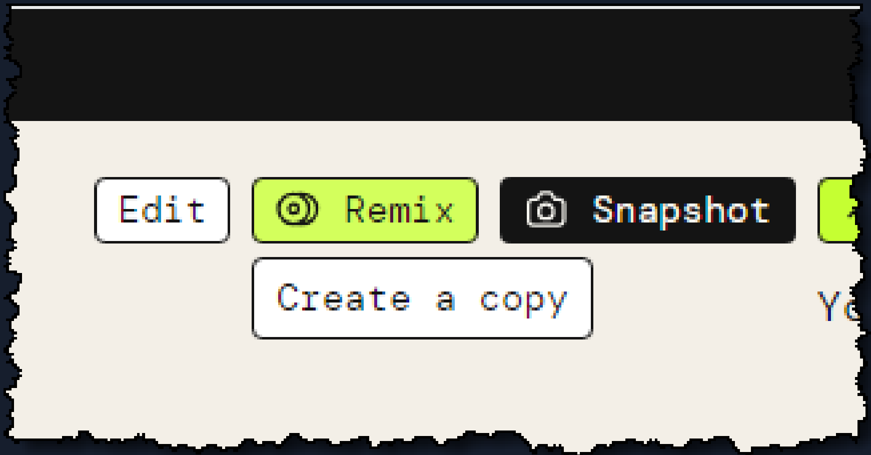


Editing app

new-app-osZtVcwB6

Your app is private.

□ Remix an app



Features

Edit

Snapshot

Remix with other apps

Share your app

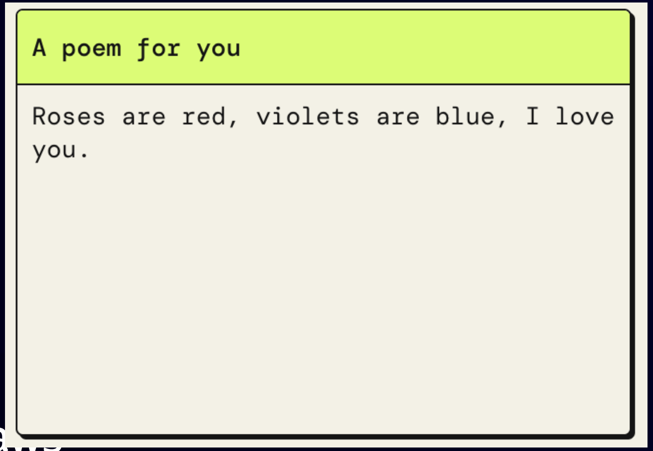
PartyRock

Everyone can build
AI apps

Widgets

A widget is a UI element that you can combine with other widgets to create an application. Widgets display content, take in input, connect to other widgets, and create output. Users interact with the application through widgets that receive input.

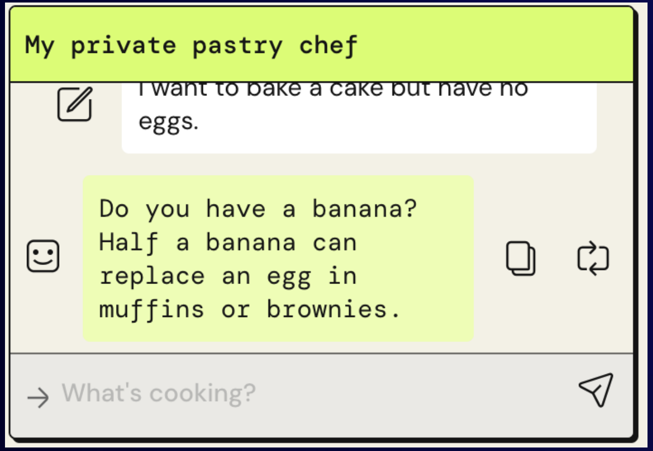
Text generation



A poem for you

Roses are red, violets are blue, I love you.

Chatbot



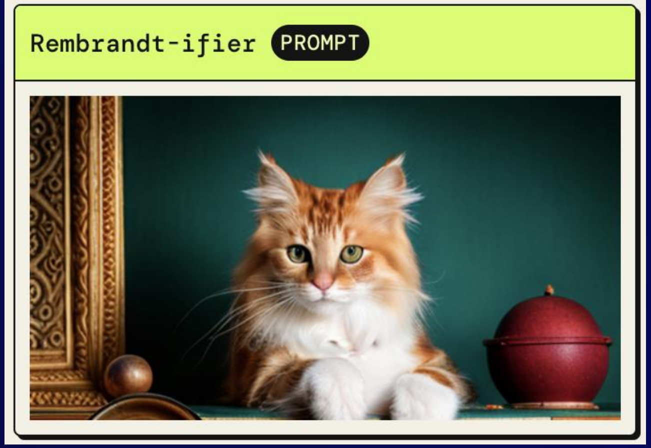
My private pastry chef

I want to bake a cake but have no eggs.

Do you have a banana? Half a banana can replace an egg in muffins or brownies.

What's cooking?

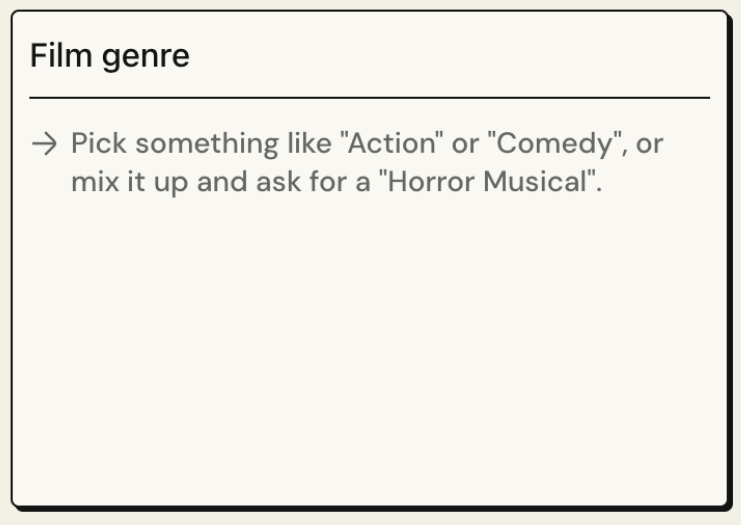
Image generation



Rembrandt-ifier PROMPT

Additional widgets

User input

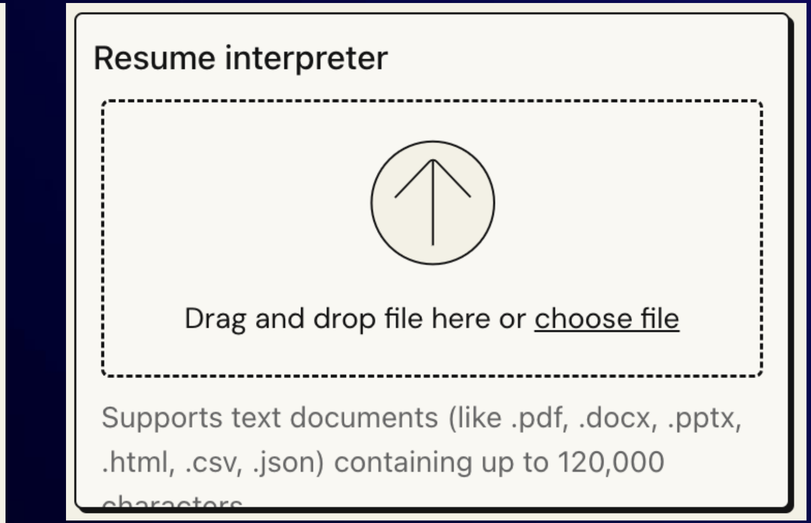


Film genre

→ Pick something like "Action" or "Comedy", or mix it up and ask for a "Horror Musical".

Provides text to change the output of connected AI-powered widgets

Document Select



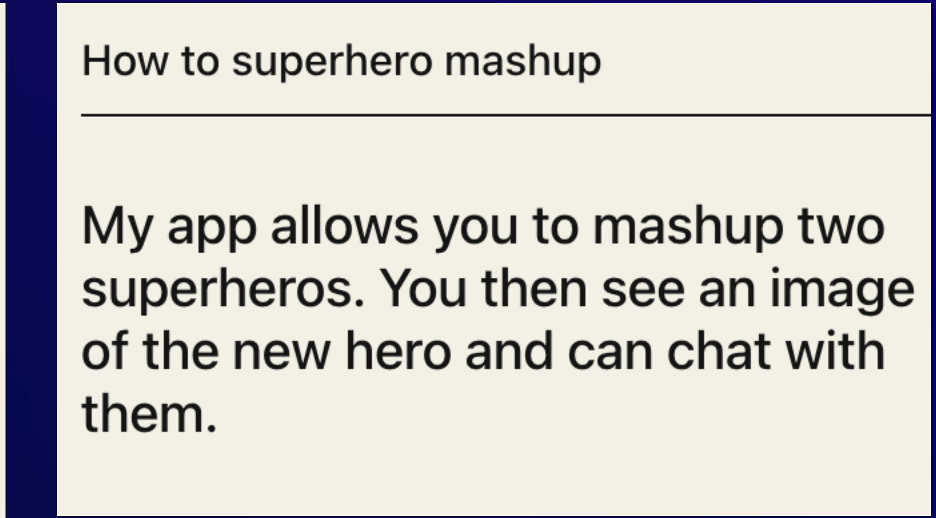
Resume interpreter

Drag and drop file here or choose file

Supports text documents (like .pdf, .docx, .pptx, .html, .csv, .json) containing up to 120,000 characters

An input field for your app user to provide a file or document with text content

Static text



How to superhero mashup

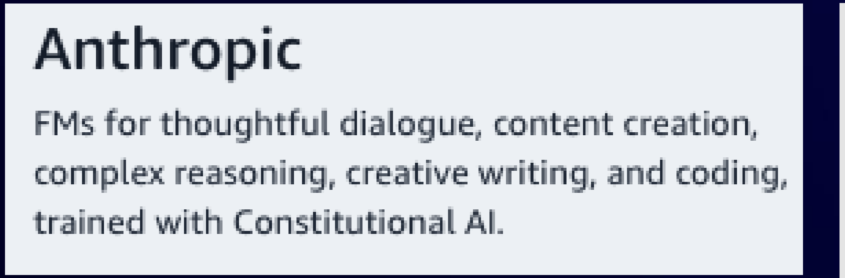
My app allows you to mashup two superheros. You then see an image of the new hero and can chat with them.

Provides a place for text descriptions about your application generation

PartyRock

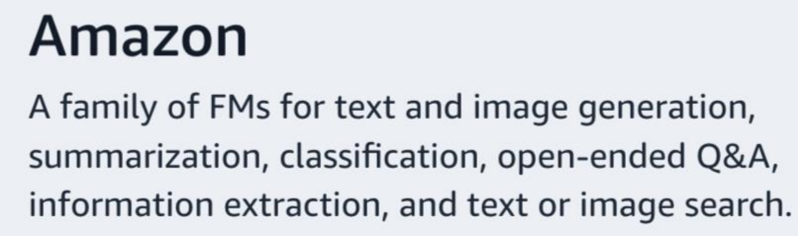
Everyone can build
AI apps

Text generation models



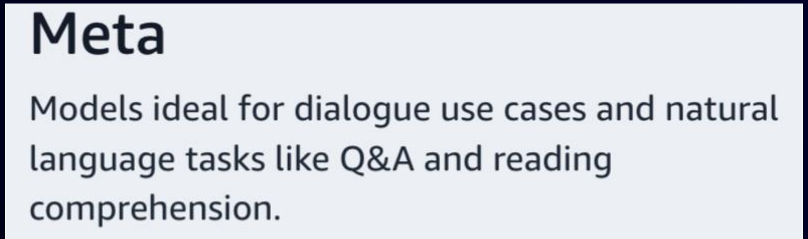
Anthropic

FMs for thoughtful dialogue, content creation, complex reasoning, creative writing, and coding, trained with Constitutional AI.



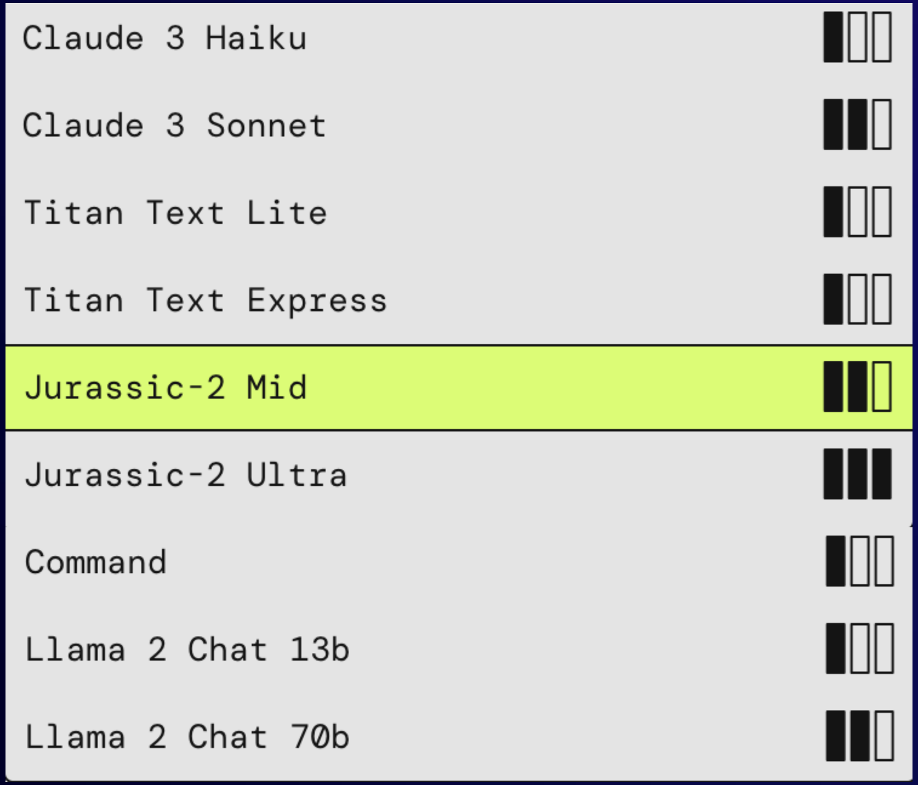
Amazon

A family of FMs for text and image generation, summarization, classification, open-ended Q&A, information extraction, and text or image search.

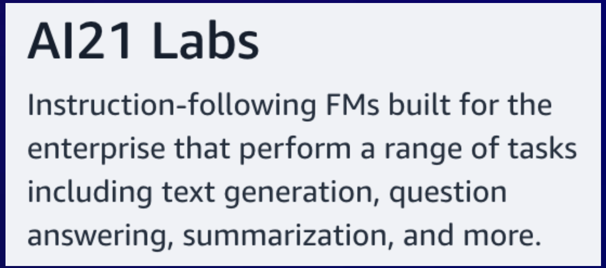


Meta

Models ideal for dialogue use cases and natural language tasks like Q&A and reading comprehension.

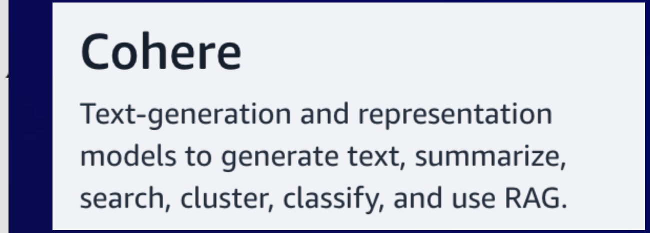


Claude 3 Haiku	■□□
Claude 3 Sonnet	■■□
Titan Text Lite	■□□
Titan Text Express	■□□
Jurassic-2 Mid	■■□
Jurassic-2 Ultra	■■■
Command	■□□
Llama 2 Chat 13b	■□□
Llama 2 Chat 70b	■■□



AI21 Labs

Instruction-following FMs built for the enterprise that perform a range of tasks including text generation, question answering, summarization, and more.



Cohere

Text-generation and representation models to generate text, summarize, search, cluster, classify, and use RAG.

Prompt Like a Pro: Getting the Most from Your PartyRock App

Prompt engineering

Prompts are a specific set of inputs provided by you, the user, that guide LLMs on Amazon Bedrock to generate an appropriate response or output for a given task or instruction

Tips for good prompt writing

- Be more specific (“Start here, turn right when you see this landmark, then go straight”)
- Emphasize the details

Prompt engineering

Prompts are a specific set of inputs provided by you, the user, that guide LLMs on Amazon Bedrock to generate an appropriate response or output for a given task or instruction

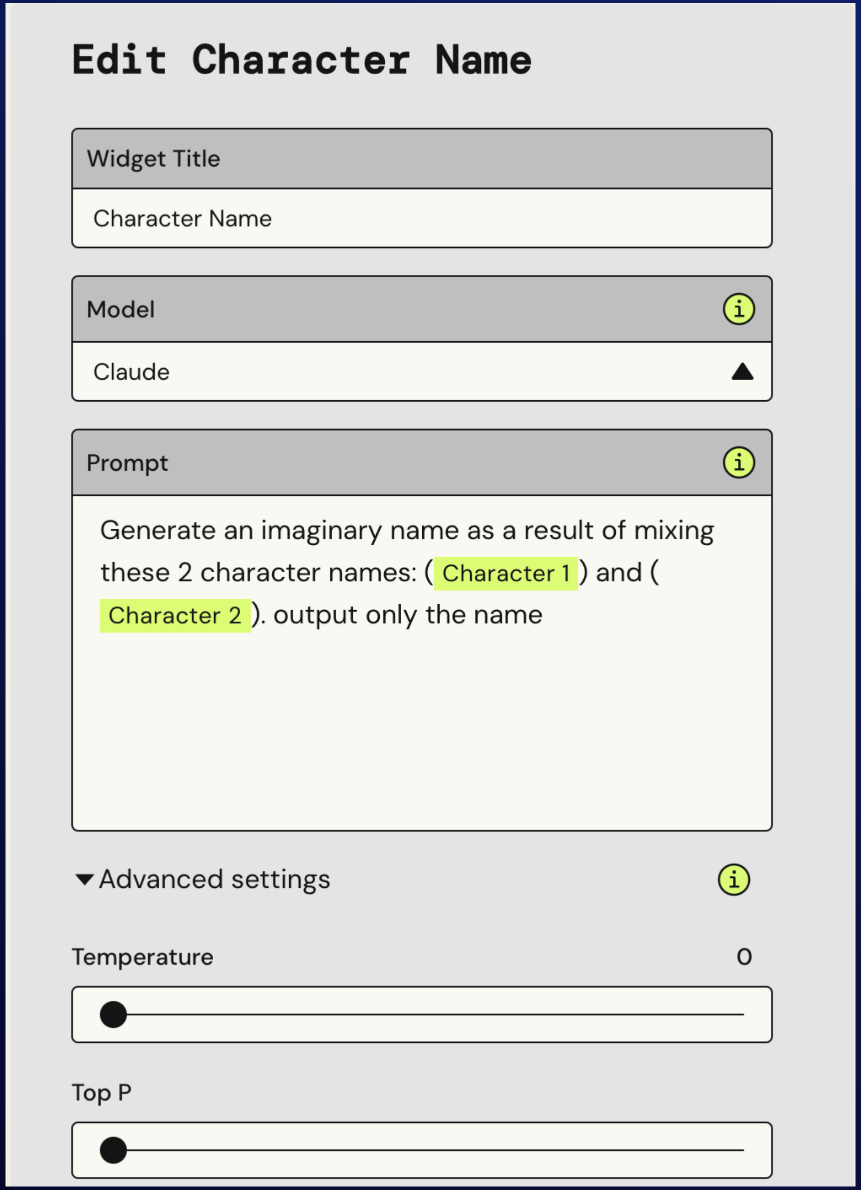
```
User prompt:
Where is re:Invent hosted?

Output:
AWS re:Invent will take place November 27 through December 1, 2023, in Las Vegas,
Nevada across multiple venues
```

Advanced prompting techniques

Prompt chaining

Temperature adjustment



Edit Character Name

Widget Title: Character Name

Model: Claude ▲

Prompt: Generate an imaginary name as a result of mixing these 2 character names: (Character 1) and (Character 2). output only the name

▾ Advanced settings

Temperature: 0

Top P

Everything you need to accelerate **your generative AI journey**

Easiest and most secure way to build generative AI applications

Data as your differentiator and strategic asset for generative AI

Most advanced cloud infrastructure for generative AI

Generative AI applications to enhance productivity

Questions?

Let’s build something today…

https://partyrock.aws/

Thank You

John Fan
Account Manager, Education
johnfan@amazon.com

linkedin.com/in/johnfanyw