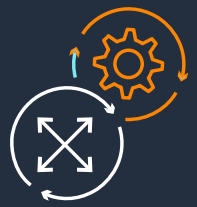
Generative AI on AWS

Arpad Csoke

Solutions Architect
Amazon Web Services

What generative AI customers are asking for

Which model should I use?



How can I move quickly?

How can I keep my data secure and private?

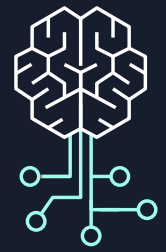
Amazon Bedrock

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

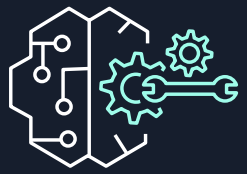
- Choice of leading FMs through a single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

aws

Amazon Bedrock simplifies



Choice



Customization

Integration

Security and governance

aws

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

aws

Amazon Bedrock Model Evaluation

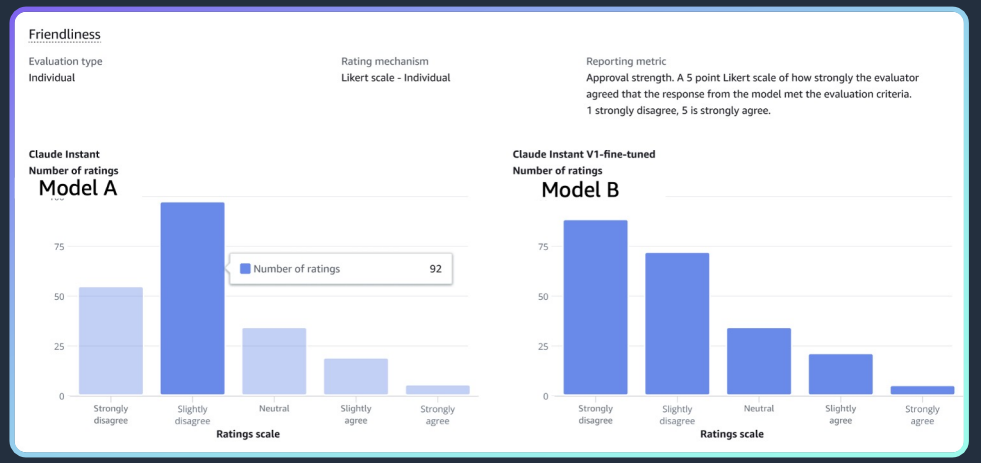
EVALUATE, COMPARE, AND SELECT THE BEST FM FOR YOUR USE CASE

Automatic or human evaluation method

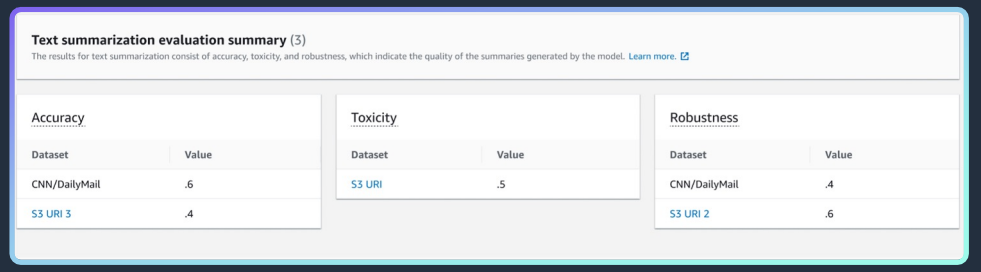
Curated datasets or bring your own

Predefined and custom metrics

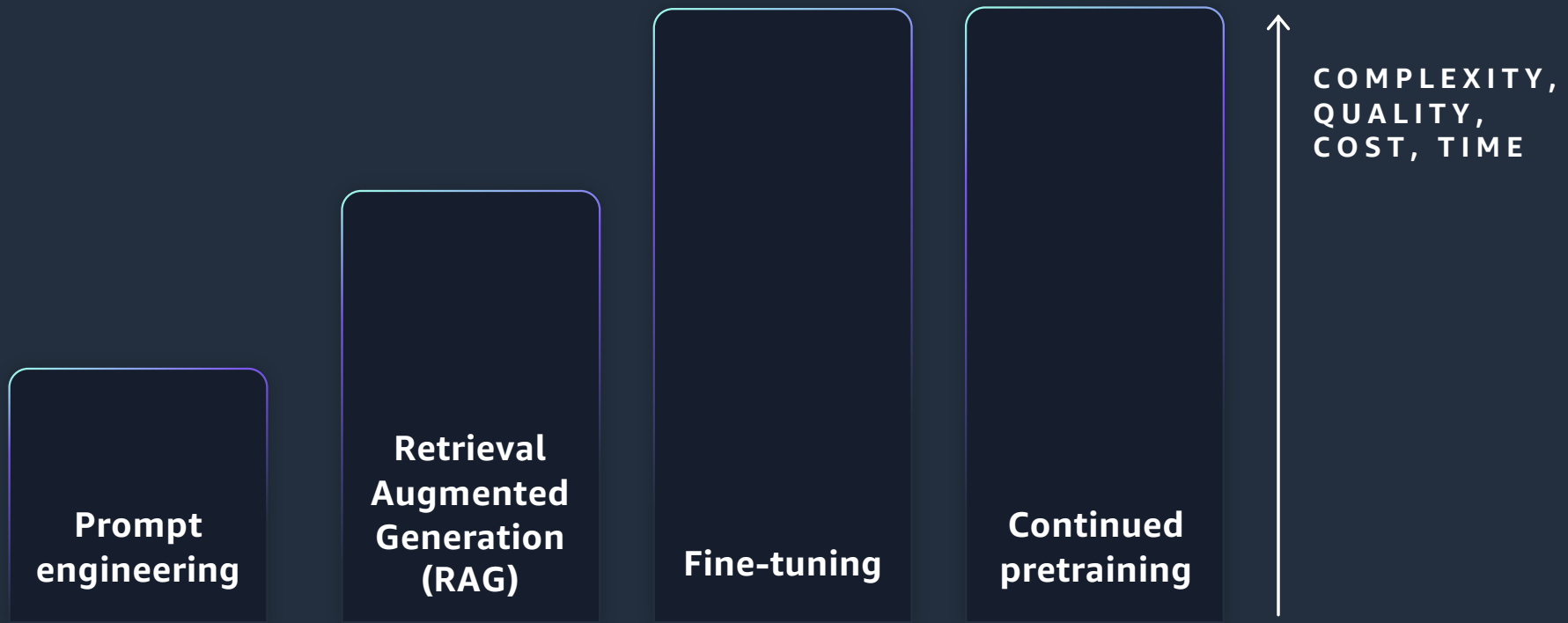
HUMAN EVALUATION REPORT



AUTOMATIC EVALUATION REPORT



Common approaches for customizing FMs



An
Pr

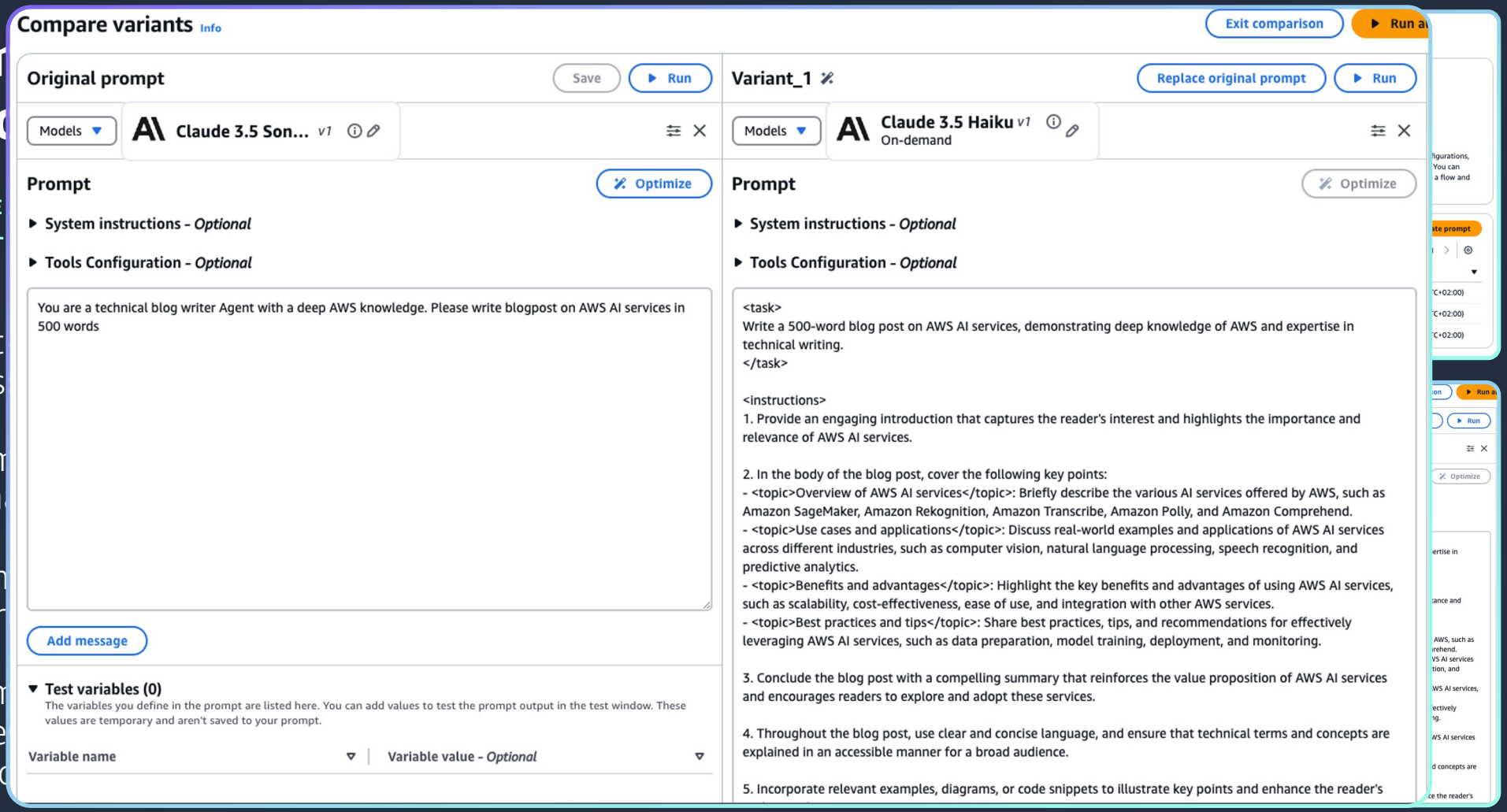
STRE
APPL

Built
cons

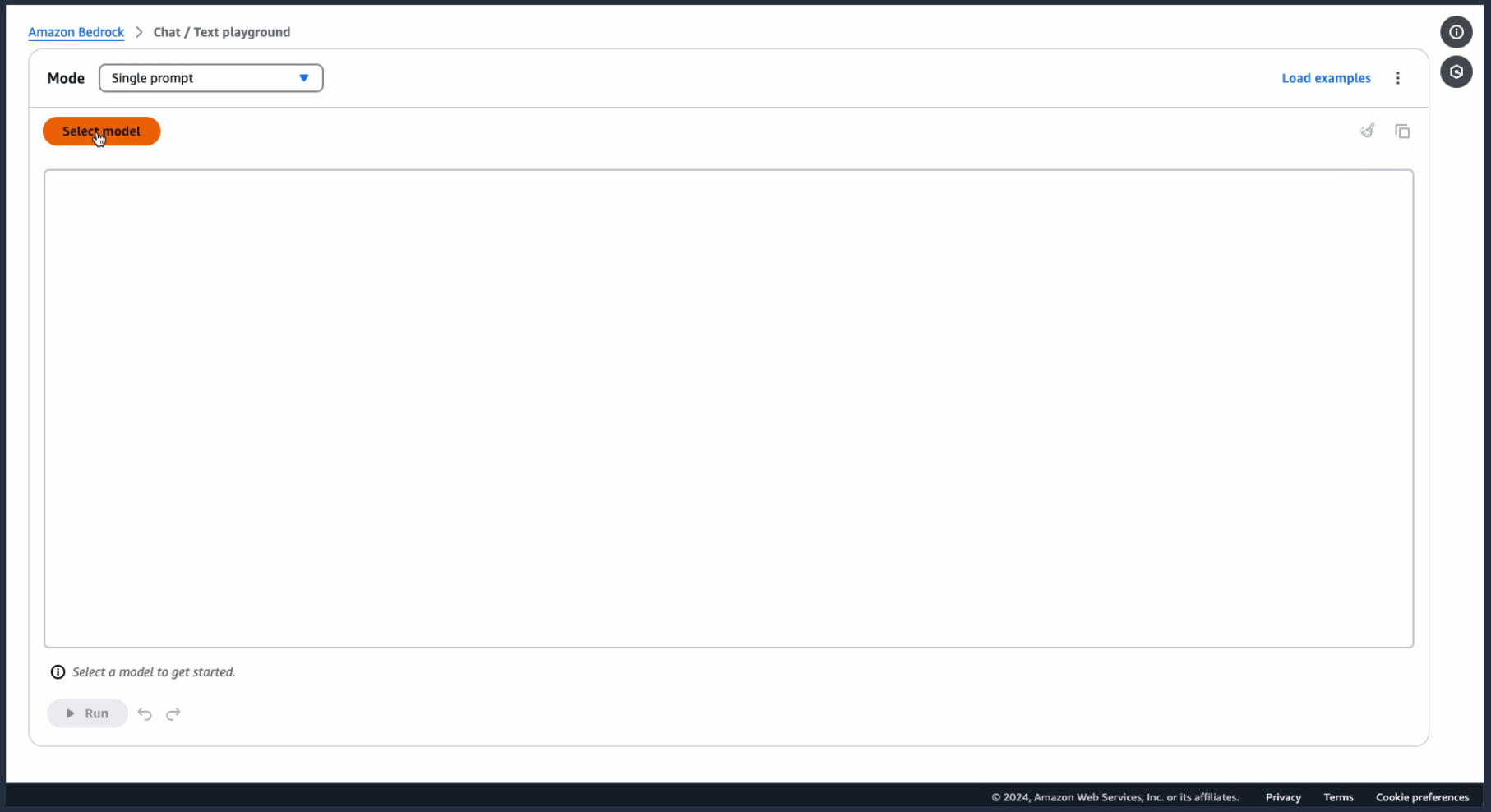
Prom
man

Sean
Bedr

Prom
to re
perfo



Amazon Bedrock Prompt Optimization



Amazon Bedrock Prompt Optimization


```
"model": "claude-3-5-sonnet-20241022"
"amazon-bedrock-invocationMetrics": {
        "inputTokenCount": 33,
        "outputTokenCount": 760,
        "invocationLatency": 20949,
        "firstByteLatency": 337
}
```

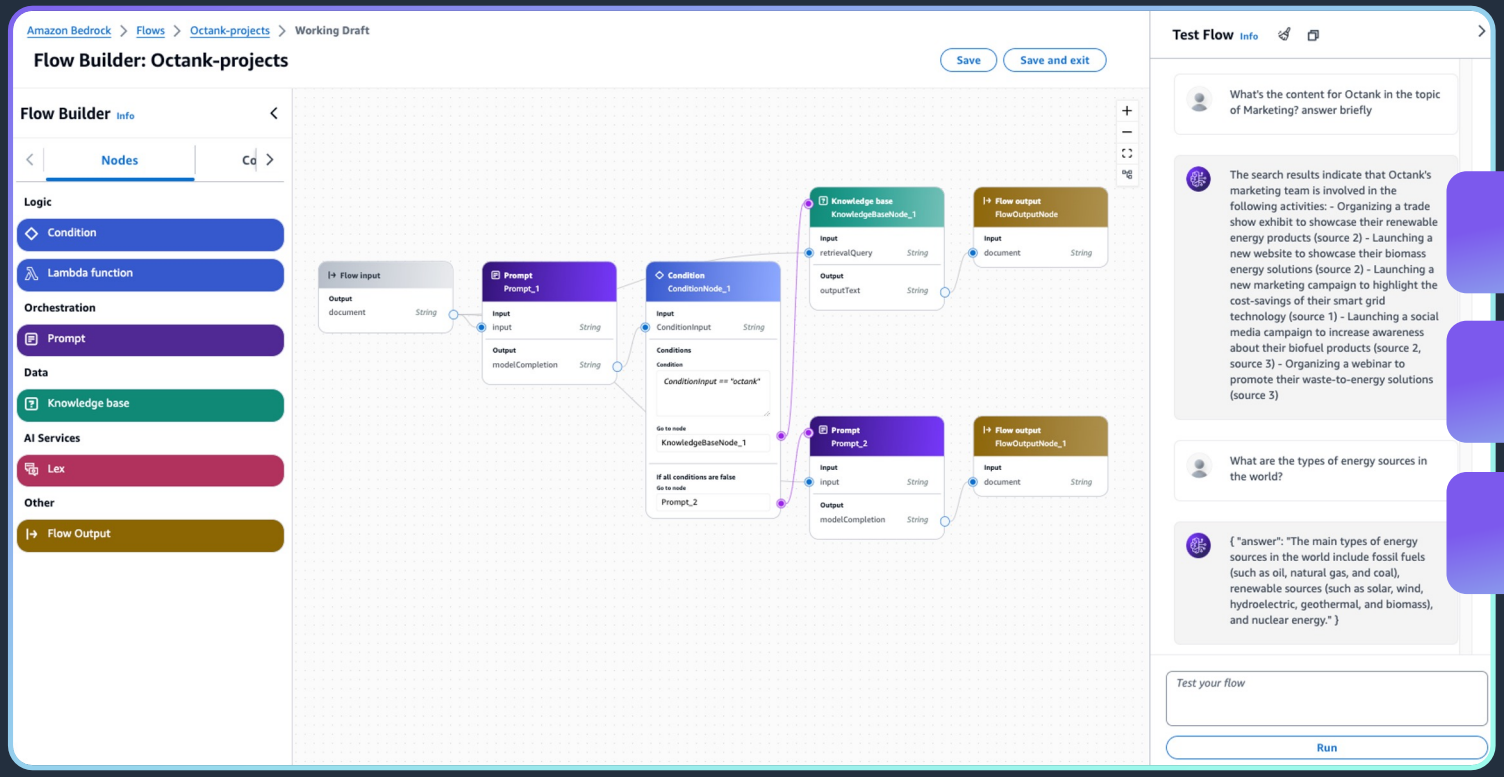
```
"model": "claude-3-5-haiku-20241022"
"amazon-bedrock-invocationMetrics": {
        "inputTokenCount": 464,
        "outputTokenCount": 682,
        "invocationLatency": 14288,
        "firstByteLatency": 345
```


3x !!!

Model	Input price /1000 token	Output price /1000 token	Prompt cost
Claude 3.5 Sonnet v2	0,003 USD	0,015 USD	0,011499 USD
Claude 3.5 Haiku	0,001 USD	0,005 USD	0,003874 USD

Amazon Bedrock Flows

VISUALIZE AND ACCELERATE GENERATIVE AI DEVELOPMENT WORKFLOWS



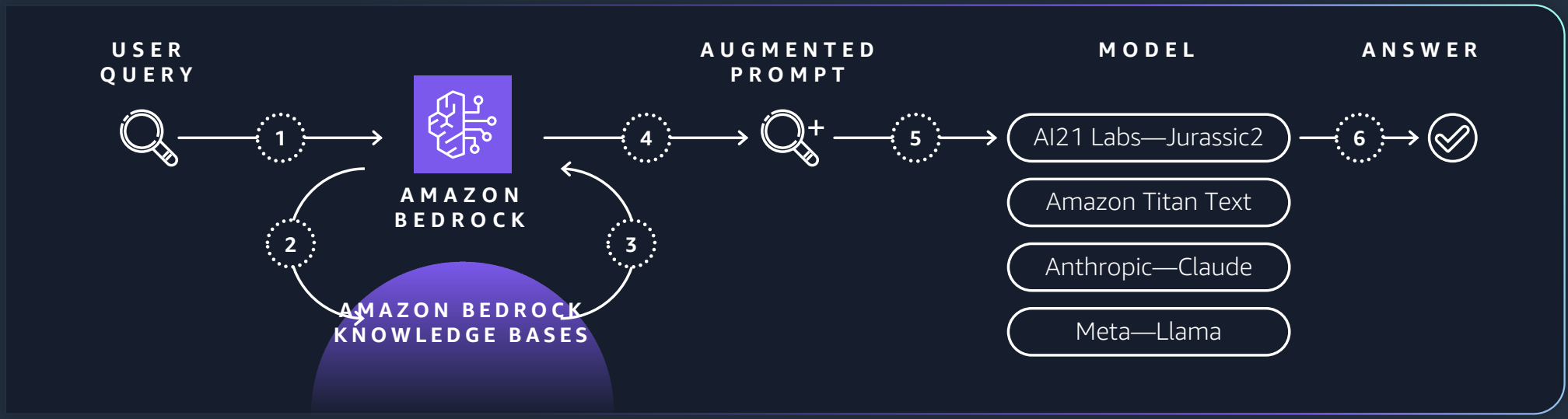
Drag-and-drop interface

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 multiturn conversations

Automatic citations with retrievals to improve transparency

aws

New data sources for Amazon Bedrock Knowledge Bases

Extract structured data, metadata, and other information from documents

Inclusion/exclusion content filters

Incremental content syncs for added, updated, deleted content

Source attribution

Web Crawler
Single or multiple URLs

Atlassian Confluence
Confluence Cloud

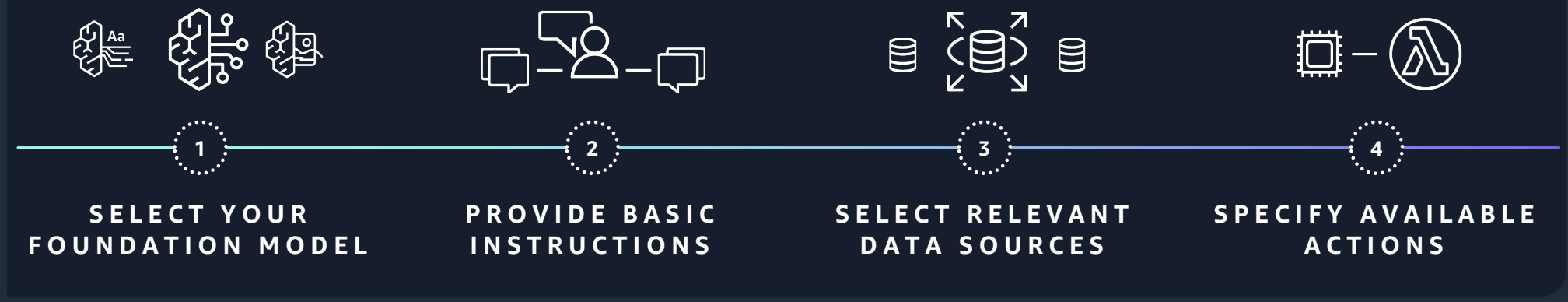
Microsoft SharePoint
SharePoint Online

Salesforce
Salesforce Standard and Custom objects

aws

Amazon Bedrock Agents

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



Breaks down and orchestrates tasks

Securely accesses and retrieves company data for RAG

Takes action by invoking API calls on your behalf

Chain-of-thought trace and ability to modify agent prompts

aws

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

aws

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

aws

Cross-region Inference

Dynamically route requests across AWS Regions to manage traffic bursts

- Use on-demand mode to get higher throughput limits (up to 2x in-Region quota) for optimal availability, performance, and resiliency during periods of peak demand
- Prioritize the Region of the connected Amazon Bedrock API endpoint whenever possible, and minimize latency to a nearby opted-in Region
- Select either a US model or an EU model from two– three geographic Regions for Claude 3 and 3.5 Sonnet models.
- Pay the same price per token for models as in your primary Region, with no additional routing or data transfer costs

aws

Supported AWS Regions in Amazon Bedrock

Intelligent Document Processing (IDP)

and

Generative AI

Faster document processing shortens decision cycles and drives material ROI by enabling enterprises to streamline business operations, boost employee productivity and enhance customer experiences.

IDP ROI Proofpoints

Reduced **cost**

HealthFirst automated medical chart extraction resulting in 10-20x revenue savings

Improved **accuracy**

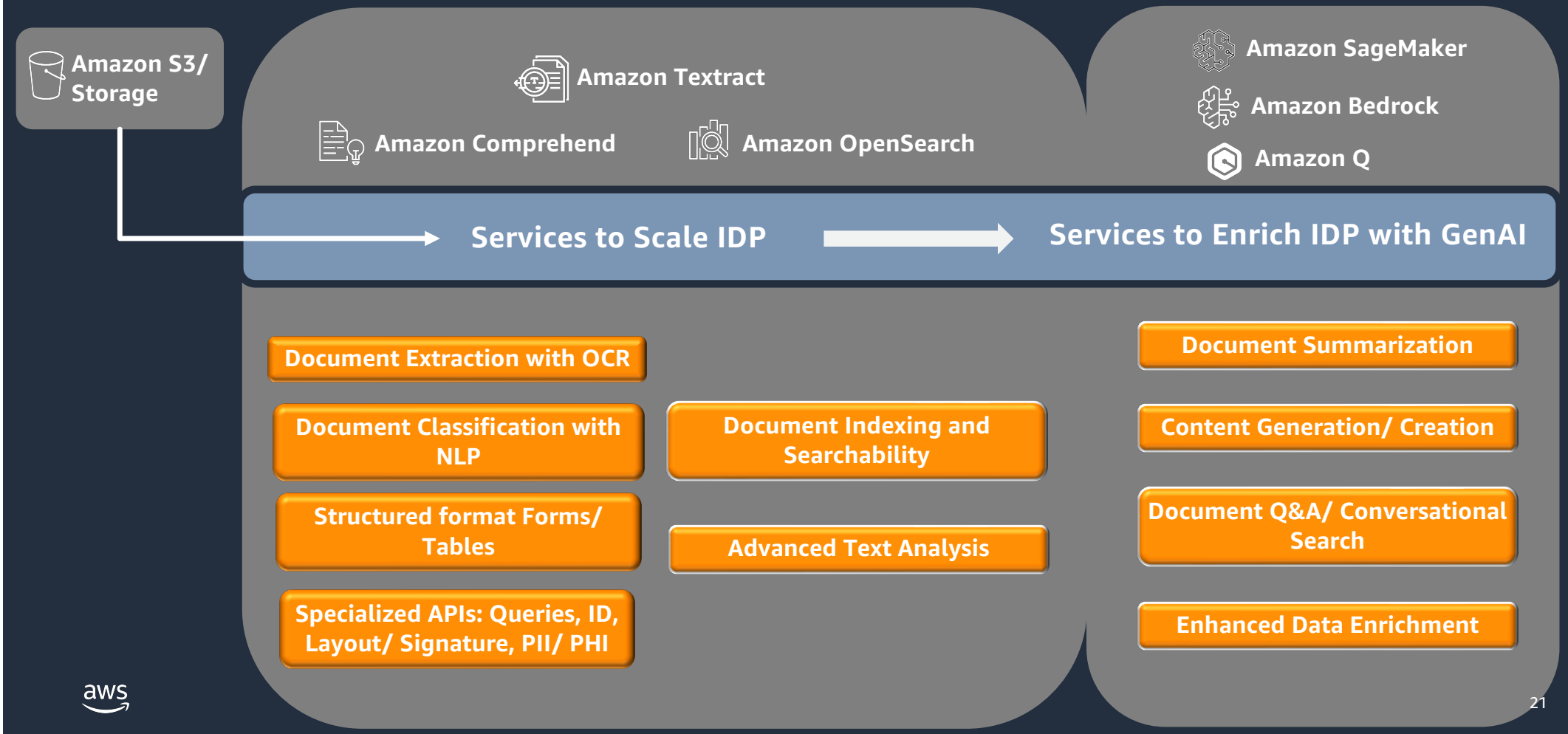
Paytm extracts user data from documents with **97%** accuracy with Amazon Textract

Improved **productivity**

Elevance Health automated classification of attachments for claims process by **90%**

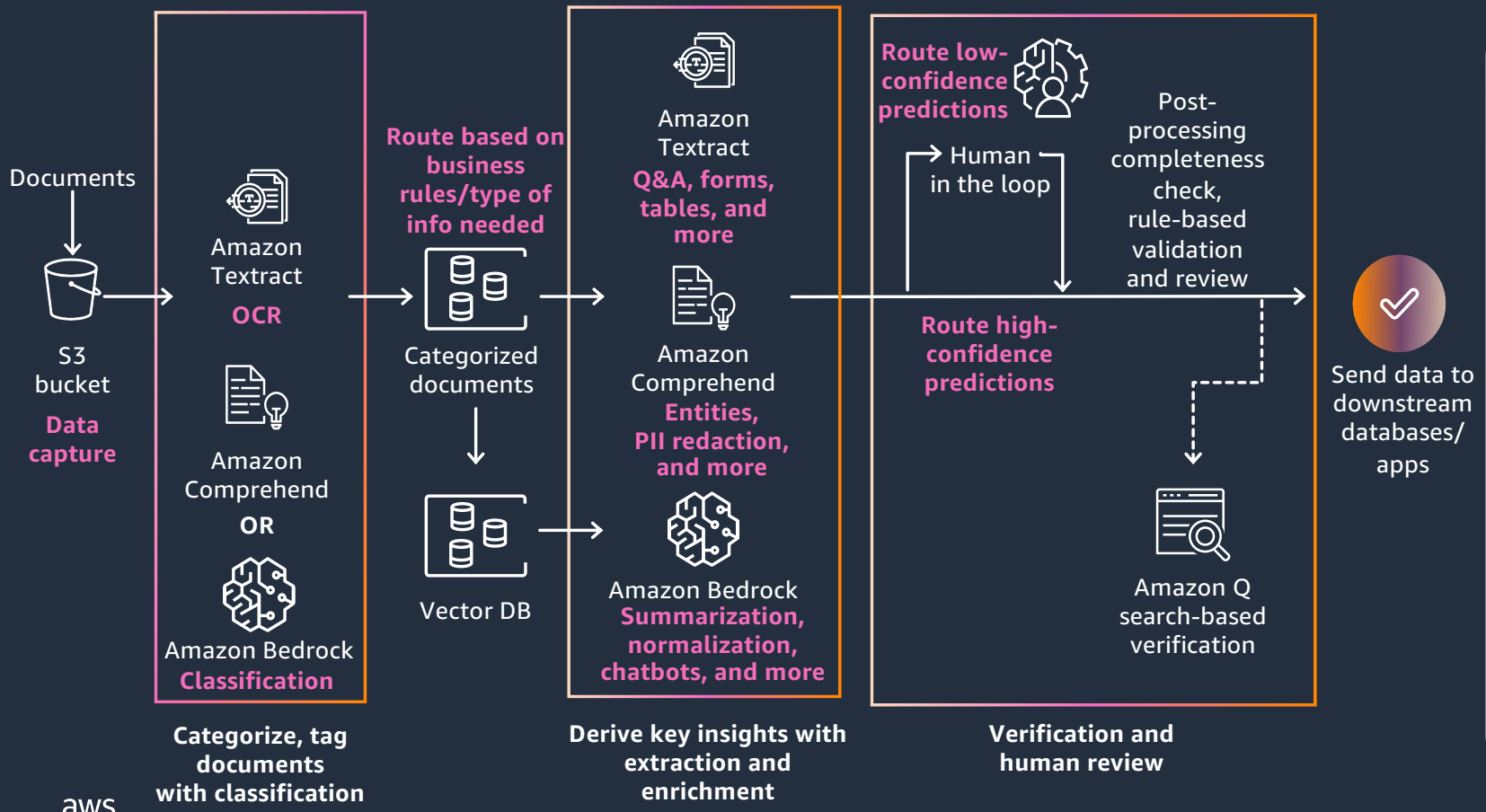
aws

Amazon offers a wide breadth of IDP Services



The Typical IDP Pipeline involves multiple Phases

Example Document pipeline with AWS IDP and generative AI



How to get started

Developers can easily embed AI-powered functionality from Textract and Comprehend into your business workflows and apps

Engage your data science team for FM selection, evaluation and tuning based on your generative AI use case

Link your gen AI/FM modules with AWS IDP (e.g., through chaining) to create an end-to-end document processing pipeline

Storage Services – Amazon S3

Scanned documents

Captured by cell phone cameras or digital scanners

Digital Files

Storing documents in a central location, such as Amazon S3, for each business use-case, so they are ready for processing

Scans from physical mailrooms

Docs from digital mailrooms

Uploaded by end user using a computer or phone

Email attachments

aws

Amazon Textract

Use the different features within Amazon Textract to accurately extract key data from the documents.

OCR

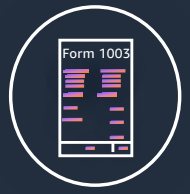
Forms

Tables

Entities

Layout

Queries



Invoices/ID Cards

Signatures

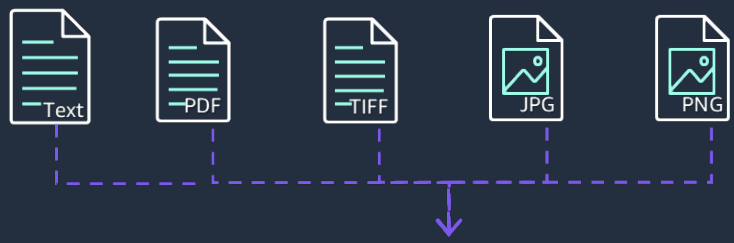
Amazon Comprehend

Use Amazon Comprehend to classify the documents in your Amazon S3 bucket

Identify document types

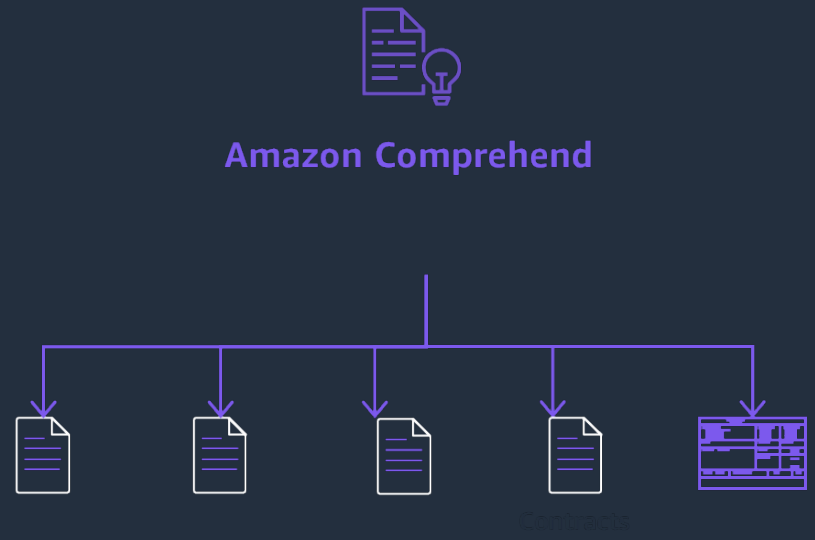
Amazon Comprehend **Custom Classifier,** and send them to the correct document pipeline.

To train a **custom classifier** you need to provide 10-50 samples of each type of document to teach the IDP AI how to classify your documents.



Real-time or asynchronous processing

Perform real time document classification when needed or reduce costs using asynchronous processes.



IDP+GenAI: Amazon Bedrock

Enhance and extend the value of extracted data through advanced GenAI models.

- Choice of industry-leading FMs available via a single API
- Customize your models using your organization's data
- Enterprise-grade security and privacy

IDP + Amazon Bedrock can unlock further use cases:

- ✓ Content Summarization
- ✓ Content Generation
- ✓ Automate personalized Response Generation
- ✓ Error Correction & Validation

AI21labs
Jurassic-2
Contextual answers, summarization, paraphrasing

ANTHROP\C
Claude x.0
Summarization, complex reasoning, writing, coding

cohere
Command & Embed
Text generation, search, classification

Meta
Llama 3
Dialogue use cases and language tasks

Mistral AI
Mistral XB
Text summarization, Q&A, Text classification, Text completion, code generation

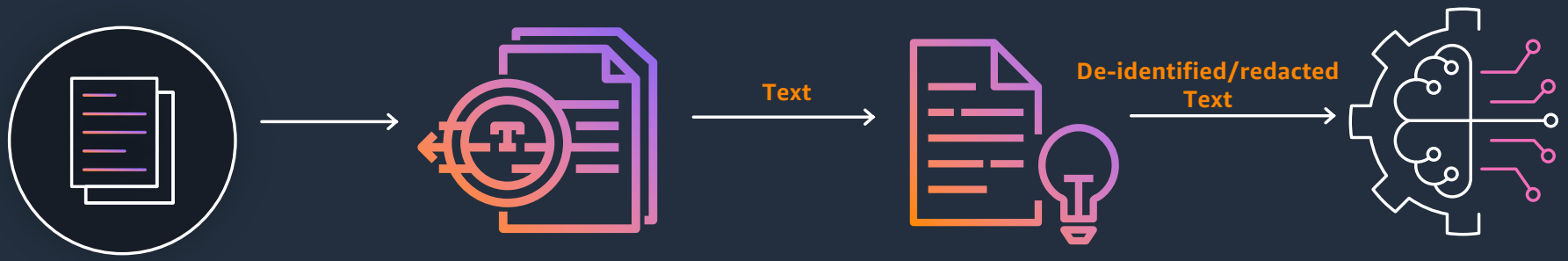
stability.ai
Stable Diffusion XL 1.0
High-quality images and art

amazon
Amazon Titan
Summarization, image and text generation and search, Q&A

aws

Sensitive data handling for Generative AI use-cases

Redacting inputs can improve sensitive data handling with Generative AI models.



Documents stored in Amazon S3

Amazon Textract

Forms, IDs, expenses, medical/healthcare reports.
Structured, Semi-structured

Amazon Comprehend

PII (Personally Identifiable Information)- toxicity, intent, abuse filters

Foundation model

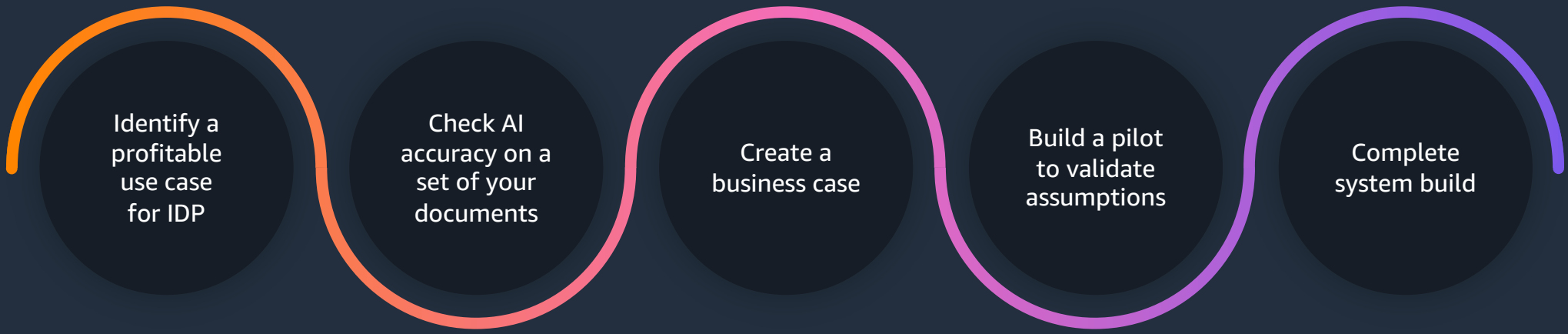
Train/Tune an LLM using your documents

IDP Use-cases with Generative AI

A few very common use cases

- Document Q&A with Chatbots
- Document summarization
- Enhanced data extraction
- Document classification
- Automated content creation
- Medical record analysis
- Translation and localization
- Learning and development
- And more…

IDP Journey



Time to stage: 1 month

Resources: No code, AWS Management Console

1 week

Build and partner paths available

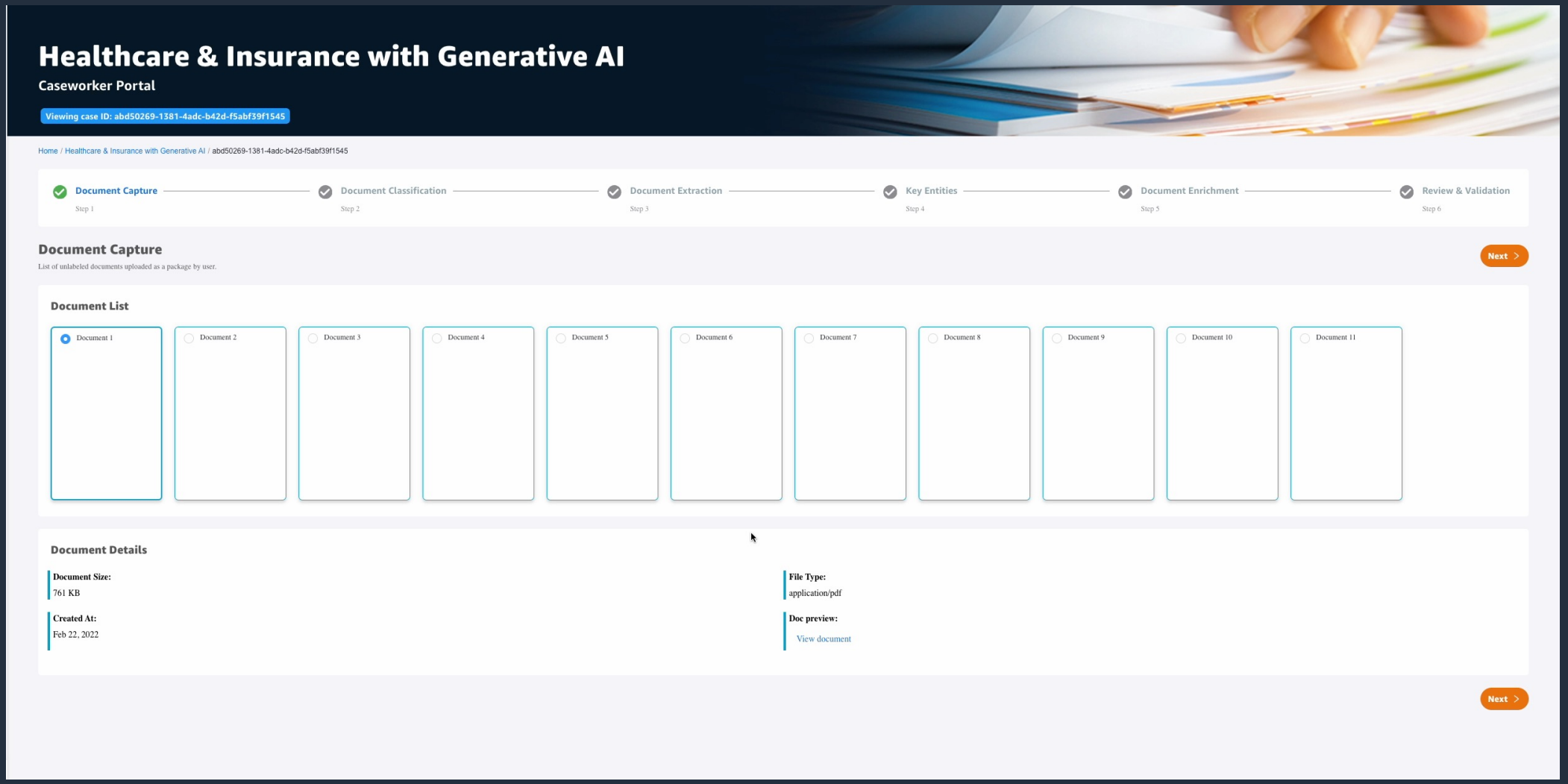
1–2 months for learning APIs and building a standalone pilot

Immersion day and partner recommendations

1–3 months

Integration tips and partner recommendations

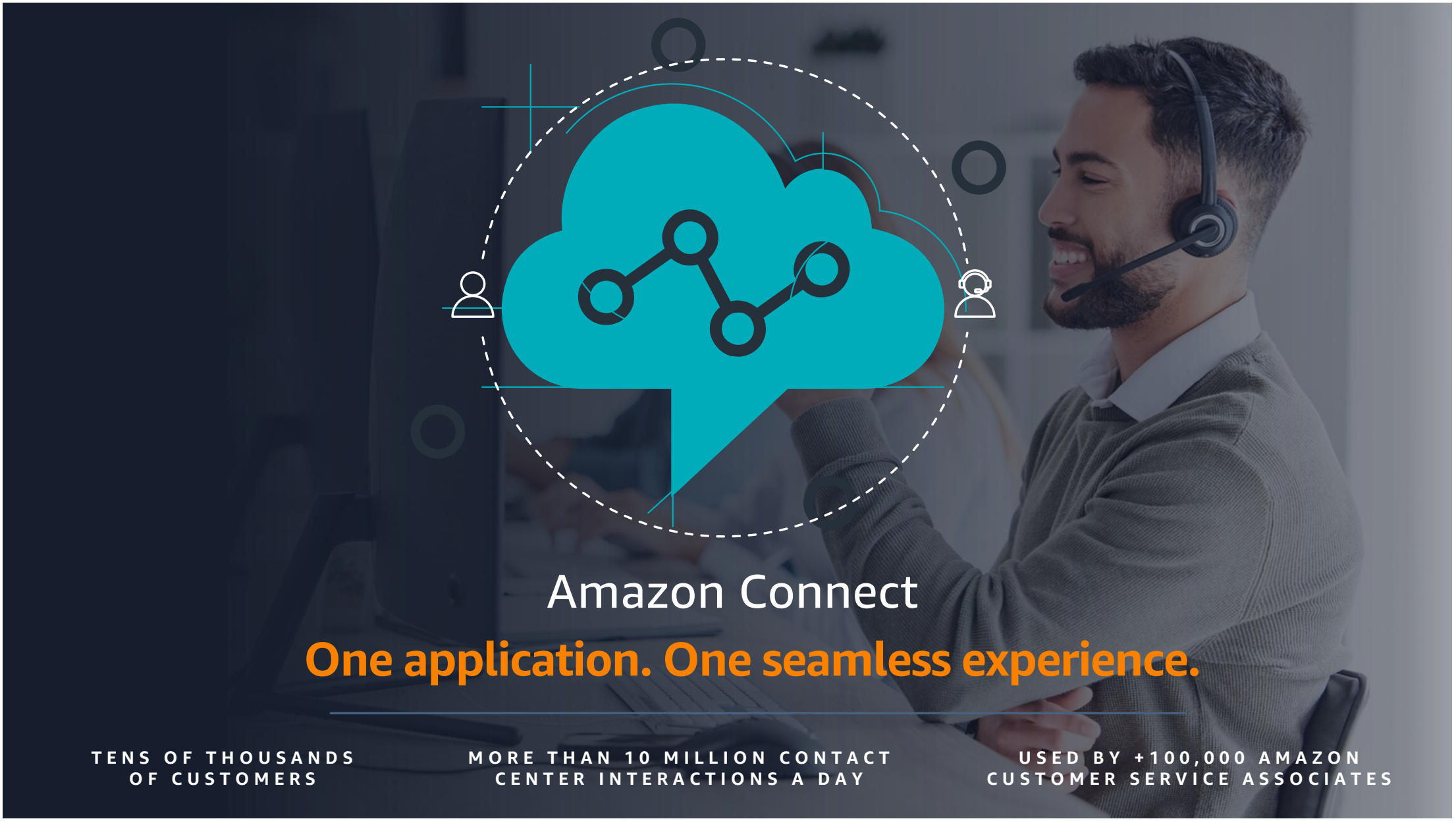
IDP
Demo



Amazon Connect

and

Generative AI


Amazon Connect
One application. One seamless experience.
TENS OF THOUSANDS OF CUSTOMERS
MORE THAN 10 MILLION CONTACT CENTER INTERACTIONS A DAY
USED BY +100,000 AMAZON CUSTOMER SERVICE ASSOCIATES

CUSTOMER

OMNICHANNEL CUSTOMER EXPERIENCE

- High quality voice
- Conversational IVR and chatbots
- Chat, SMS, and messaging
- In-app/web calling and video
- Outbound campaigns
- Voice authentication
- Task management

AGENT

AGENT EMPOWERMENT AND PRODUCTIVITY

- Agent workspace
- Knowledge and agent assist
- Unified customer view
- Case management
- Step-by-step guides

AMAZON CONNECT

SUPERVISOR

ANALYTICS, INSIGHTS, AND OPTIMIZATION

- Real-time and historical reports and dashboards
- Real-time conversational analytics
- Quality and performance management
- Forecasting, capacity planning, and scheduling

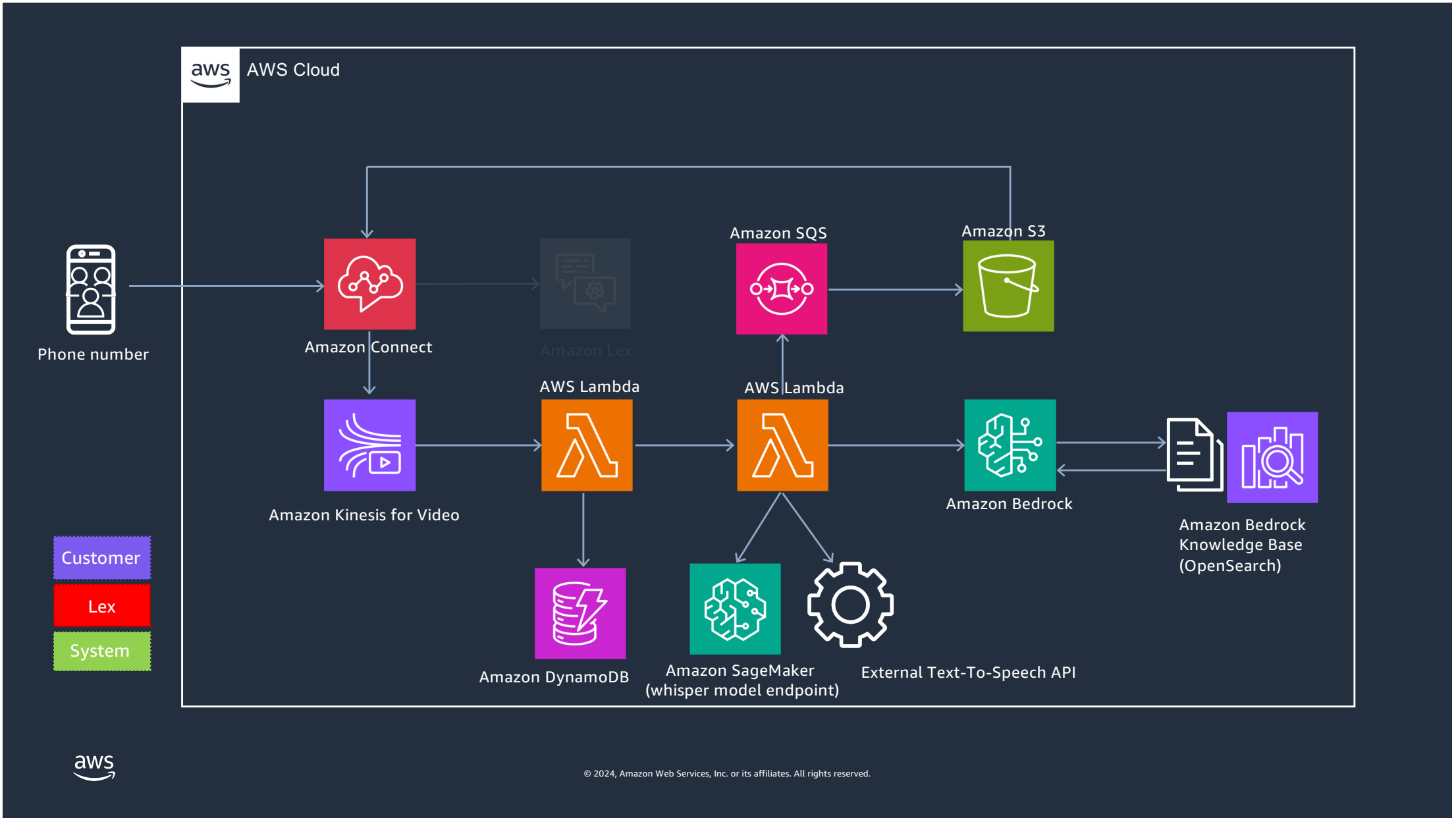
ADMINISTRATOR

CONFIGURATION AND FLEXIBILITY

- Simple, self-service UI; make changes in minutes, not months
- Secure, reliable, and scalable
- Modern API and SDKs for those who want to build

app built by aws

One application. One seamless experience.



Quiz

Thank you!

Arpad Csoke
arpadcs@amazon.de