AI Conclave

Online

AIOT101

Amazon Bedrock state of the union

Aparajithan Vaidyanathan

Principal Solutions Architect
AWS India

Agenda

- Introduction
- Choose the best model
- Customize with your data
- Apply safety and responsible AI checks
- Build and orchestrate agents
- Optimize for cost, latency and accuracy

Introduction

Amazon Bedrock

The easiest way to build and scale generative AI applications

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 | re:Invent launches

THE EASIEST AND FASTEST WAY TO BUILD AND SCALE GENERATIVE AI APPLICATIONS

CHOOSE THE BEST MODEL

amazon

Luma

poolside

Marketplace (GA)

Model evaluation LLM-as-a-judge (preview)

CUSTOMIZE WITH YOUR DATA

Data Automation (preview)

Knowledge Bases supports GraphRAG (preview)

Knowledge Bases supports structured data retrieval (GA)

Knowledge Bases supports multimodal data processing (GA)

Knowledge Bases supports streaming responses (GA)

Knowledge Bases supports real-time sync from custom data sources (GA)

Rerank API (GA)

Knowledge Bases supports RAG evaluation (preview)

Knowledge Bases provides auto-generated query filters for retrieval (GA)

APPLY SAFETY AND RESPONSBILE AI CHECKS

Guardrails support Automated Reasoning check (preview)

Guardrails supports multimodal toxicity detection (preview)

BUILD AND ORCHESTRATE AGENTS

Multi-agent collaboration (preview)

OPTIMIZE FOR COST, LATENCY, AND ACCURACY

Latency-optimized Inference Options (preview)

Prompt caching (preview)

Intelligent Prompt Routing (preview)

Model distillation (preview)

AWS security, privacy and reliability built-in

aws

Choose the best model

Amazon Bedrock

AI21labs

Effective reasoning & rapid analysis for long context windows

JAMBA

amazon

Frontier multimodal intelligence at low-latency, Agent & RAG Applications, high-quality image & video generation

AMAZON NOVA

New

ANTHROP\C

Advanced reasoning & coding capabilities, including computer use skills

CLAUDE

cohere

Multimodal search & advanced retrieval powering multilingual knowledge agents

COMMAND

EMBED

RERANK New

Luma

High-quality video generation from text & images

LUMA RAY 2

Coming soon

Meta

Advanced image & language reasoning

LLAMA

MISTRAL AI_

Knowledge summarization, expert agents, & code completion

MISTRAL

MIXTRAL

poolside

Software engineering AI for large enterprises

MALIBU

POINT

Coming soon

stability.ai

High-quality AI image generation, easily deployable at scale

STABLE DIFFUSION

STABLE IMAGE

aws

Amazon Nova foundation models

State-of-the-art foundation models that deliver frontier intelligence and industry leading price performance.

Understanding 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

Lower cost & latency ↔ Increasing intelligence

Creative content generation models

Amazon Nova Canvas

State-of-the-art image generation model

GENERALLY AVAILABLE

Amazon Nova Reel

State-of-the-art video generation model

GENERALLY AVAILABLE

Amazon Nova - Understanding models

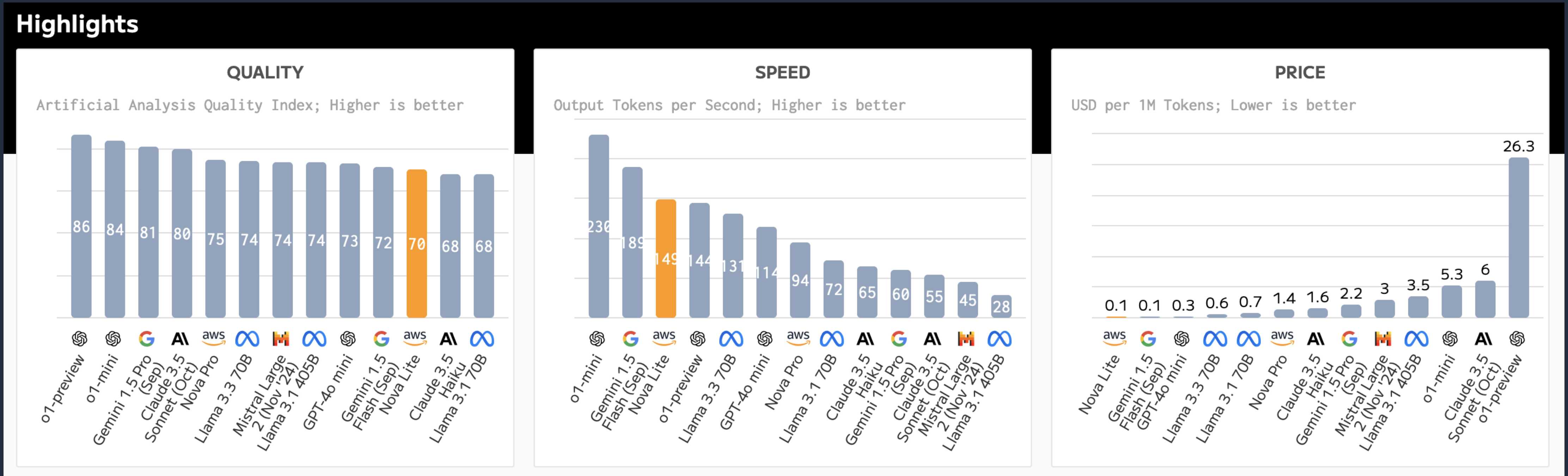
	Amazon Nova Micro	Amazon Nova Lite	Amazon Nova Pro	Amazon Nova Premier
Availability	GA	GA	GA	Coming soon
Context window	128K	300K	300K (5M coming soon)	Coming soon
Languages	200+ languages	200+ languages	200+ languages	Coming soon
Modalities supported	Text input; text output	Text, image, video input; text output	Text, image, video input; text output	Coming soon
Fine-tuning	Yes	Yes	Yes	Coming soon

Amazon Nova creative content generation models

	Amazon Nova Reel	Amazon Nova Canvas
Availability	GA	GA
Input characters	512	1024
Languages	EN	EN
Modalities supported	Text, image input; video output	Text, image input; image output
Duration	6 seconds (2 minutes coming soon)	N/A
Fine-tuning	Coming soon	Coming soon

3rd party penchmarks

Amazon Nova Lite



Highlights

QUALITY

Artificial Analysis Quality Index; Higher is better

SPEED

Output Tokens per Second; Higher is better

PRICE

USD per 1M Tokens; Lower is better

Source: artificialanalysis.ai (as of 12/9/24)

Amazon Bedrock

Marketplace

(GA)

Discover and use over 100 popular, emerging and specialized models in Amazon Bedrock

- Streamline development workflows with a unified console experience
- Deploy models on managed endpoints with custom scaling policies
- Leverage Amazon Bedrock’s APIs, tools, and security

Amazon Bedrock

LLM-as-a-Judge

(Preview)

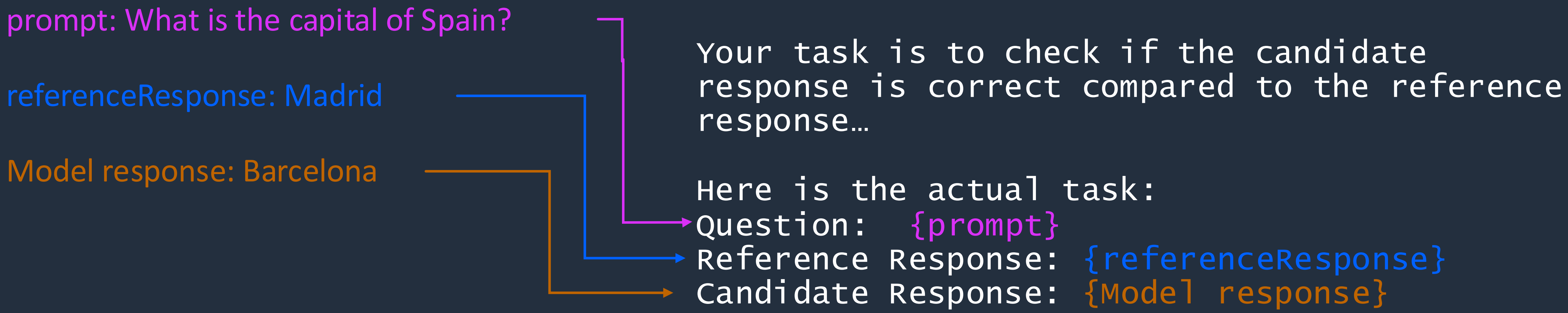
Get human-like evaluation quality at a much lower cost than full human-based evaluations, while saving weeks of time

- Ensure you have the right combination of evaluator models and models being evaluated
- Use curated metrics to evaluate objective facts or subjective evaluation of writing style and tone on your dataset
- Compare results across evaluation jobs to make decisions faster

How does LLM-as-a-judge work?

Example input

Judge prompt (simplified)



prompt: What is the capital of Spain?

referenceResponse: Madrid

Model response: Barcelona

```
You are a helpful assistant…

You are given a question, a candidate
response from an LLM, and reference response.

Your task is to check if the candidate
response is correct compared to the reference
response…

Here is the actual task:
Question:  {prompt}
Reference Response: {referenceResponse}
Candidate Response: {Model response}

Explain your response, followed by your
evaluation:
2) Correct
1) Partially correct
0) Incorrect
```

Note: Correctness can be with or without ground truth

Customize with your data

Amazon Bedrock Knowledge Bases

re:Invent 2024 launches

Structured data retrieval

Auto-generated query filters

Multi-modal data processing

Rerank API

GraphRAG

Real-time sync for custom data sources

Streaming responses

RAG evaluation

Amazon Bedrock Knowledge Bases

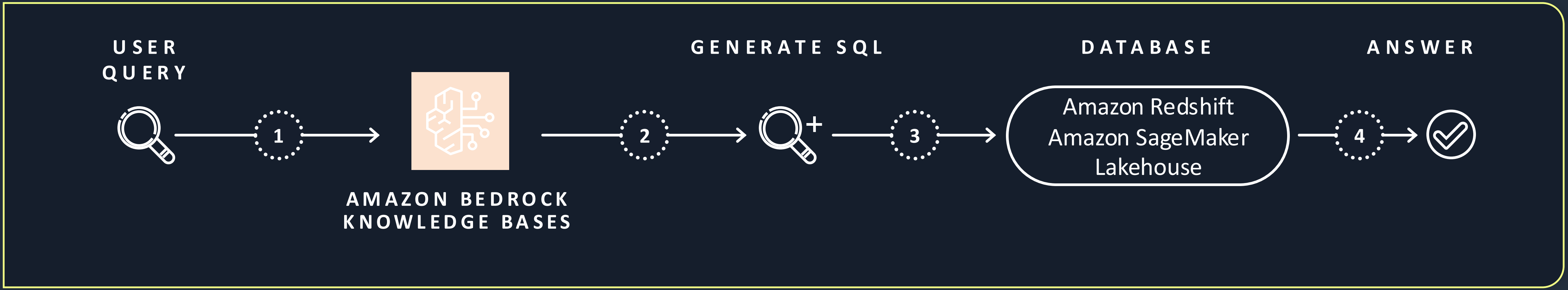
structured data retrieval

(GA)

Seamlessly integrate structured data for RAG

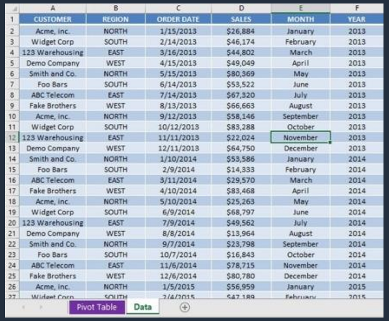
- Use data stored in Amazon SageMaker Lakehouse, Amazon Redshift and Amazon S3 Tables
- Reduce application development time from months to days
- Improve the accuracy of your queries with customization context

Structured data retrieval



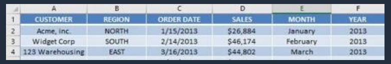
“What was the top selling product in November?”

```
SELECT customer_id, SUM(order_total) AS
total_revenue

FROM customer_orders

WHERE order_date BETWEEN '2024-01-01’
AND '2024-01-31’

GROUP BY customer_id ORDER BY
total_revenue DESC;
```



	A	B	C	D	E	F
1	CUSTOMER	REGION	ORDER DATE	SALES	MONTH	YEAR
2	Acme, inc.	NORTH	1/15/2013	$26,884	January	2013
3	Widget Corp	SOUTH	2/14/2013	$46,174	February	2013
4	123 Warehousing	EAST	3/16/2013	$44,802	March	2013
5	Demo Company	WEST	4/15/2013	$49,049	April	2013
6	Smith and Co.	NORTH	5/15/2013	$80,369	May	2013
7	Foo Bars	SOUTH	6/14/2013	$53,522	June	2013
8	ABC Telecom	EAST	7/14/2013	$67,320	July	2013
9	Fake Brothers	WEST	8/13/2013	$66,663	August	2013
10	Acme, inc.	NORTH	9/12/2013	$58,146	September	2013
11	Widget Corp	SOUTH	10/12/2013	$83,288	October	2013
12	123 Warehousing	EAST	11/11/2013	$22,024	November	2013
13	Demo Company	WEST	12/11/2013	$64,750	December	2013
14	Smith and Co.	NORTH	1/10/2014	$53,586	January	2014
15	Foo Bars	SOUTH	2/9/2014	$14,333	February	2014
16	ABC Telecom	EAST	3/11/2014	$29,570	March	2014
17	Fake Brothers	WEST	4/10/2014	$83,468	April	2014
18	Acme, inc.	NORTH	5/10/2014	$25,263	May	2014
19	Widget Corp	SOUTH	6/9/2014	$68,797	June	2014
20	123 Warehousing	EAST	7/9/2014	$49,562	July	2014
21	Demo Company	WEST	8/8/2014	$13,964	August	2014
22	Smith and Co.	NORTH	9/7/2014	$23,798	September	2014
23	Foo Bars	SOUTH	10/7/2014	$16,843	October	2014
24	ABC Telecom	EAST	11/6/2014	$78,715	November	2014
25	Fake Brothers	WEST	12/6/2014	$80,780	December	2014
26	Acme, inc.	NORTH	1/5/2015	$56,959	January	2015
27	Widget Corp	SOUTH	2/4/2015	$47,189	February	2015

“In November,
the top selling product is YYYY”



	A	B	C	D	E	F
1	CUSTOMER	REGION	ORDER DATE	SALES	MONTH	YEAR
2	Acme, inc.	NORTH	1/15/2013	$26,884	January	2013
3	Widget Corp	SOUTH	2/14/2013	$46,174	February	2013
4	123 Warehousing	EAST	3/16/2013	$44,802	March	2013

Amazon Bedrock Knowledge Bases

GraphRAG

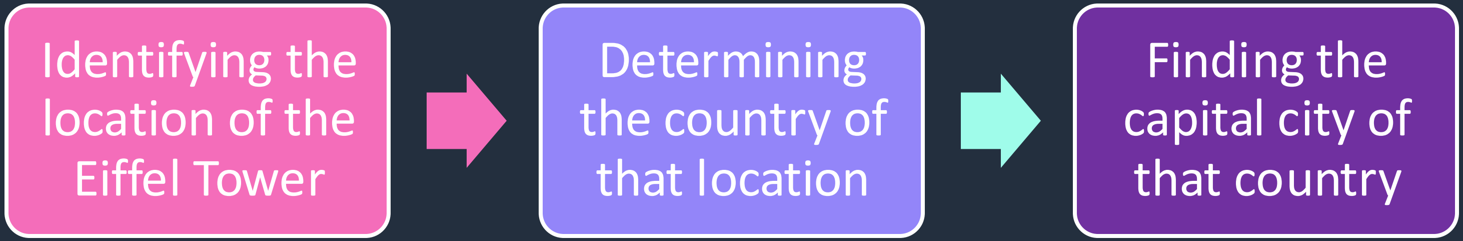
(preview)

Generate more relevant responses for RAG applications using knowledge graphs

- Generate knowledge graphs to link relationships across data sources
- Build more comprehensive, explainable generative AI applications
- Enhance transparency of source information for better fact verification

What is the name of the capital city of the country where Eiffel Tower is located?

Requires **Multi-step reasoning**



Identifying the location of the Eiffel Tower

Determining the country of that location

Finding the capital city of that country

Basic RAG

It doesn't inherently understand relationships between entities (Eiffel Tower, Paris, France)

Basic RAG might find information about the Eiffel Tower or about France's capital separately, but may not connect these pieces of information effectively.

GraphRAG

It understands the relationships: Eiffel Tower is in Paris, Paris is the capital of France

GraphRAG can establish connections between multiple entities: Eiffel Tower, Paris, France

Recognizes the hierarchy: Landmark -> City -> Country -> Capital City

Amazon Bedrock Knowledge Bases

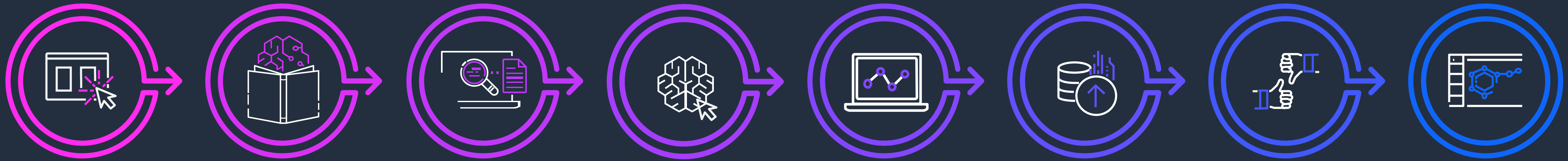
RAG evaluation

(preview)

Evaluate end-to-end RAG workflow

- Get actionable insights to improve your RAG system
- Ensure the generated content is correct, complete, limits hallucinations, and adheres to responsible AI principles
- Accelerate time to value for deploying RAG applications

RAG evaluation:



1. Choose evaluator model
2. Choose your knowledge base
3. Choose to evaluate retrieval only or retrieve and generate
4. Choose your generator model
5. Choose your metrics
6. Upload your prompt dataset
7. Inference and evaluation
8. View results

Amazon Bedrock Knowledge Bases

streaming responses

(GA)

Introduce a new RetrieveAndGenerateStream API to respond fast

- Reduce latency and improve user experience
- Maintain the same level of accuracy
- Adhere to all relevant data privacy and security regulations, including GDPR, HIPAA, and FedRAMP

Amazon Bedrock Knowledge Bases

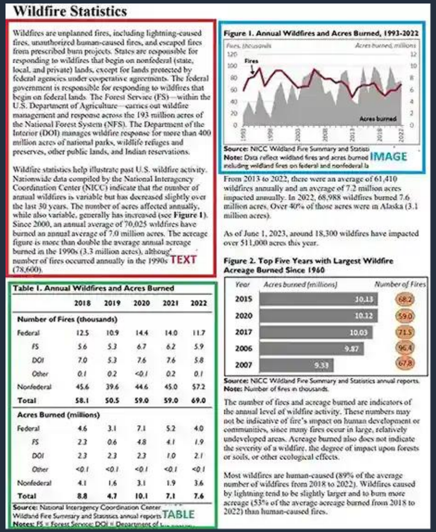
multimodal data processing

(GA)

Analyze and leverage insights from both textual and image data

- Get a fully-managed RAG workflow for visually-rich data
- Retrieve and generate answers to questions derived from text and visual data
- Improve accuracy, relevancy, and depth of responses

Multimodal data processing



Wildfire Statistics

Wildfires are unplanned fires, including lightning-caused fires, unauthorized human-caused fires, and escaped fires from prescribed burn projects. States are responsible for responding to wildfires that begin on nonfederal (state, local, and private) lands, except for lands protected by federal agencies under cooperative agreements. The federal government is responsible for responding to wildfires that begin on federal lands. The Forest Service (FS)—within the U.S. Department of Agriculture—carries out wildfire management and response across the 193 million acres of the National Forest System (NFS). The Department of the Interior (DOI) manages wildfire response for more than 400 million acres of national parks, wildlife refuges and preserves, other public lands, and Indian reservations.

Wildfire statistics help illustrate past U.S. wildfire activity. Nationwide data compiled by the National Interagency Coordination Center (NICC) indicate that the number of annual wildfires is variable but has decreased slightly over the last 30 years. The number of acres affected annually, while also variable, generally has increased (see **Figure 1**). Since 2000, an annual average of 70,025 wildfires have burned an annual average of 7.0 million acres. The acreage figure is more than double the average annual acreage burned in the 1990s (3.3 million acres), althoug' number of fires occurred annually in the 1990s (78,600).

TEXT

Figure 1. Annual Wildfires and Acres Burned, 1993-2022

Source: NICC Wildland Fire Summary and Statisti
Note: Data reflect wildland fires and acres burned including wildland fires on federal and nonfederal la

IMAGE

From 2013 to 2022, there were an average of 61,410 wildfires annually and an average of 7.2 million acres impacted annually. In 2022, 68,988 wildfires burned 7.6 million acres. Over 40% of those acres were in Alaska (3.1 million acres).

As of June 1, 2023, around 18,300 wildfires have impacted over 511,000 acres this year.

Table 1. Annual Wildfires and Acres Burned

	2018	2019	2020	2021	2022
Number of Fires (thousands)					
Federal	12.5	10.9	14.4	14.0	11.7
FS	*5.6*	*5.3*	*6.7*	*6.2*	*5.9*
DOI	*7.0*	*5.3*	*7.6*	*7.6*	*5.8*
Other	*0.1*	*0.2*	*<0.1*	*0.2*	*0.1*
Nonfederal	45.6	39.6	44.6	45.0	57.2
Total	**58.1**	**50.5**	**59.0**	**59.0**	**69.0**
Acres Burned (millions)					
Federal	4.6	3.1	7.1	5.2	4.0
FS	*2.3*	*0.6*	*4.8*	*4.1*	*1.9*
DOI	*2.3*	*2.3*	*2.3*	*1.0*	*2.1*
Other	*<0.1*	*<0.1*	*<0.1*	*<0.1*	*<0.1*
Nonfederal	4.1	1.6	3.1	1.9	3.6
Total	**8.8**	**4.7**	**10.1**	**7.1**	**7.6**

Source: National Interagency Coordination Center Wildland Fire Summary and Statistics annual reports.
Notes: FS = Forest Service; DOI = Department of [illegible]

TABLE

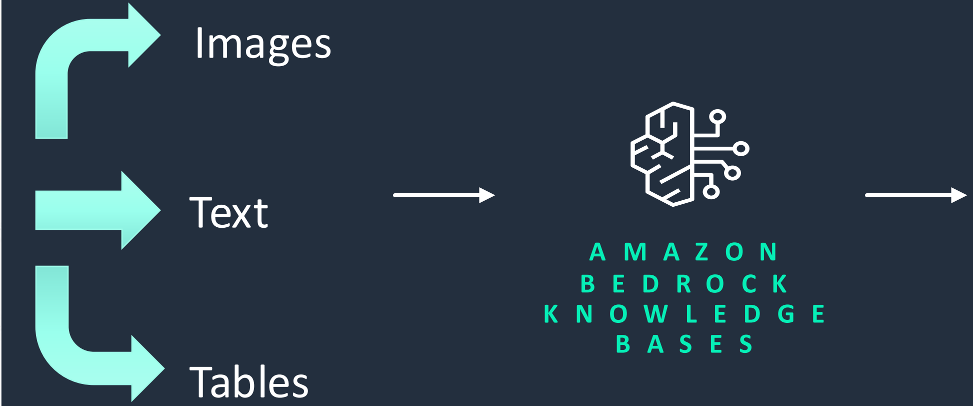
Figure 2. Top Five Years with Largest Wildfire Acreage Burned Since 1960

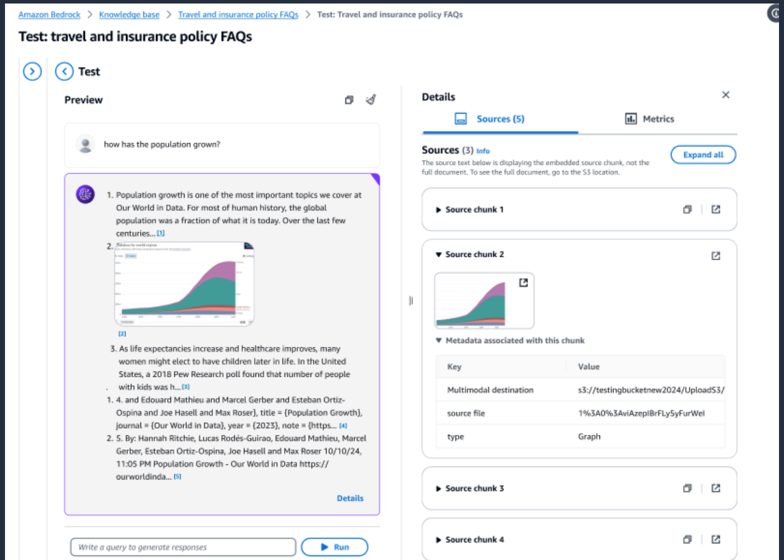
Source: NICC Wildland Fire Summary and Statistics annual reports.
Note: Number of fires in thousands.

The number of fires and acreage burned are indicators of the annual level of wildfire activity. These numbers may not be indicative of fire's impact on human development or communities, since many fires occur in large, relatively undeveloped areas. Acreage burned also does not indicate the severity of a wildfire, the degree of impact upon forests or soils, or other ecological effects.

Most wildfires are human-caused (89% of the average number of wildfires from 2018 to 2022). Wildfires caused by lightning tend to be slightly larger and to burn more acreage (53% of the average acreage burned from 2018 to 2022) than human-caused fires.

Documents





Amazon Bedrock > Knowledge base > Travel and insurance policy FAQs > Test: Travel and insurance policy FAQs

Test: travel and insurance policy FAQs

how has the population grown?

1. Population growth is one of the most important topics we cover at Our World in Data. For most of human history, the global population was a fraction of what it is today. Over the last few centuries... [1]
2. [2]
3. As life expectancies increase and healthcare improves, many women might elect to have children later in life. In the United States, a 2018 Pew Research poll found that number of people with kids was h... [3]

1. 4. and Edouard Mathieu and Marcel Gerber and Esteban Ortiz-Ospina and Joe Hasell and Max Roser}, title = {Population Growth}, journal = {Our World in Data}, year = {2023}, note = {https... [4]
2. 5. By: Hannah Ritchie, Lucas Rodés-Guirao, Edouard Mathieu, Marcel Gerber, Esteban Ortiz-Ospina, Joe Hasell and Max Roser 10/10/24, 11:05 PM Population Growth - Our World in Data https://ourworldinda... [5]

Sources (3) Info

The source text below is displaying the embedded source chunk, not the full document. To see the full document, go to the S3 location.

Key	Value
Multimodal destination	s3://testingbucketnew2024/UploadS3/
source file	1%3A0%3AviAzeplBrFLy5yFurWel
type	Graph

Choose either Amazon Bedrock Data Automation or foundation model for parsing

Ingest documents with images, tables, and text

Query your knowledge base

Get generated response back with images as source attribution

Amazon Bedrock Knowledge Bases

real-time sync from custom data sources

(GA)

Custom connector API to directly ingest content and manage individual documents easily

- Enables ingestion of content and metadata from different sources in real-time
- Enables selective updates and deletion

Amazon Bedrock Knowledge Bases

auto-generated query filters for retrieval

(GA)

Automatically generate the query filter expressions based on the metadata schema

- Improves user experience as the retrieval filter generator automatically generates the appropriate filter expressions based on user inputs
- Support metadata embedding to ensure metadata is considered in RAG process
- Highly configurable, allowing users to select metadata fields used for filtering, and allowing them to customize prompts

Amazon Bedrock

Rerank API

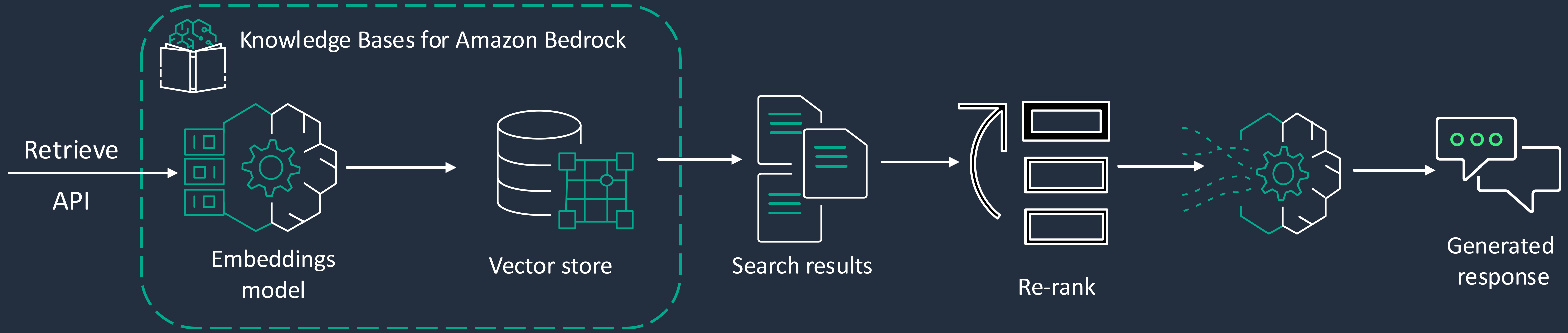
(GA)

Rerank the retrieved document chunks based on their relevance to the query

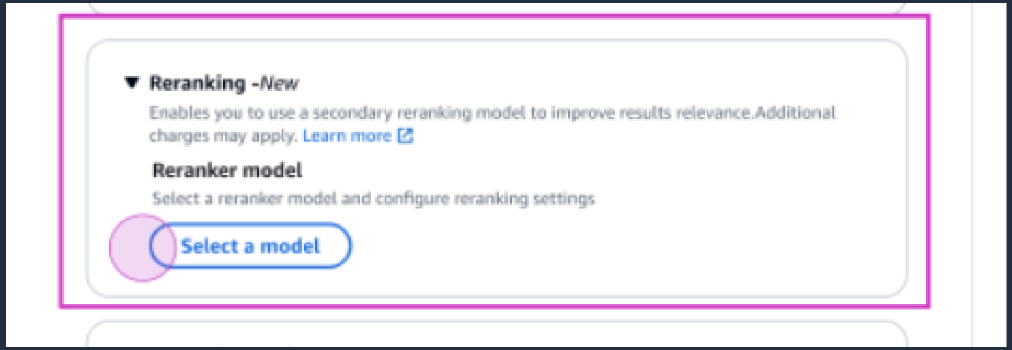
- Improve accuracy of documents by prioritizing the most important content to be passed to generation models
- Invoke the reRank API without any additional model deployment or code
- Provide model flexibility

Re-ranking

GENERALLY AVAILABLE



- Model selection – Amazon rerank, Cohere 3.5 rerank
- Accessible independently as well through Rerank API



▼ **Reranking** -*New*

Enables you to use a secondary reranking model to improve results relevance.Additional charges may apply. Learn more

Reranker model

Select a reranker model and configure reranking settings

Amazon Bedrock

Data Automation

(preview)

Transform unstructured multimodal data for generative AI applications and analytics

- Extract, transform, and generate structured data from multimodal content
- Generate customized outputs based on your requirements and business rules
- Streamline application workflows with a fully managed, single API experience

Apply safety and responsible AI checks

Amazon Bedrock Guardrails

Automated Reasoning checks

(preview)

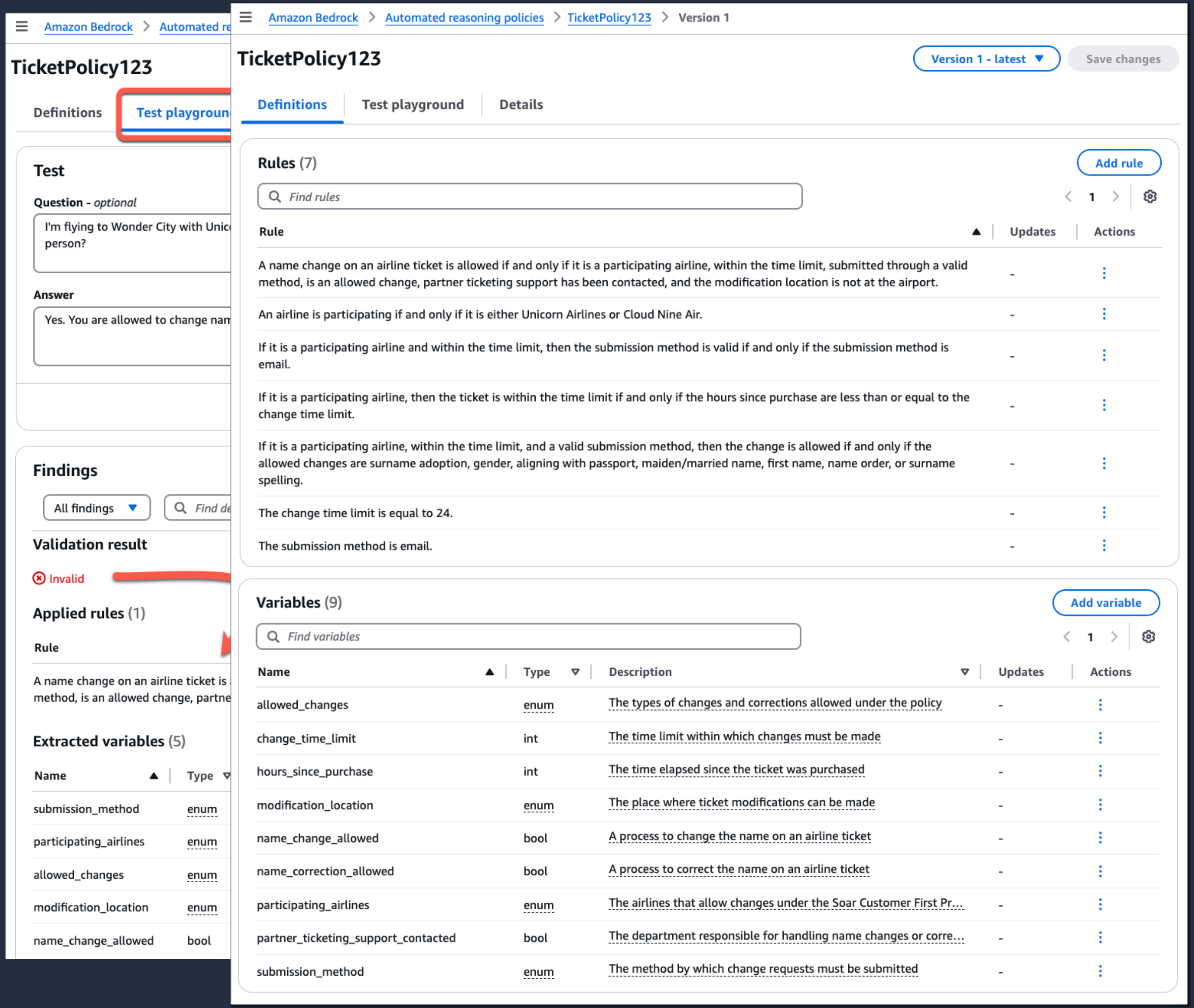
Prevent factual errors due to hallucinations

- Verify accuracy of model responses using mathematical proof
- Provides recommendations for correcting factual errors
- Enhance reliability of LLM responses for critical use cases

Automated reasoning checks – NEW!

DETECT FACTUAL ERRORS FROM HALLUCINATION WITH LOGICALLY ACCURATE AND VERIFIABLE REASONING

- **Accurate** Identifies and suggests corrections for inaccurate factual claims on supported knowledge
- **Sound** Uses formal logical reasoning to correctly determine accuracy
- **Transparent** Explains why a claim is accurate or not



TicketPolicy123

Test

Question - *optional*

I'm flying to Wonder City with Unic
person?

Answer

Yes. You are allowed to change nam

Findings

Validation result

Invalid

Applied rules (1)

Rule

A name change on an airline ticket is
method, is an allowed change, partne

Extracted variables (5)

Name ▲	Type ▼
submission_method	enum
participating_airlines	enum
allowed_changes	enum
modification_location	enum
name_change_allowed	bool

TicketPolicy123

Rules (7)

Rule ▲	Updates	Actions
A name change on an airline ticket is allowed if and only if it is a participating airline, within the time limit, submitted through a valid method, is an allowed change, partner ticketing support has been contacted, and the modification location is not at the airport.	-	⋮
An airline is participating if and only if it is either Unicorn Airlines or Cloud Nine Air.	-	⋮
If it is a participating airline and within the time limit, then the submission method is valid if and only if the submission method is email.	-	⋮
If it is a participating airline, then the ticket is within the time limit if and only if the hours since purchase are less than or equal to the change time limit.	-	⋮
If it is a participating airline, within the time limit, and a valid submission method, then the change is allowed if and only if the allowed changes are surname adoption, gender, aligning with passport, maiden/married name, first name, name order, or surname spelling.	-	⋮
The change time limit is equal to 24.	-	⋮
The submission method is email.	-	⋮

Variables (9)

Name ▲	Type ▼	Description ▼	Updates	Actions
allowed_changes	enum	The types of changes and corrections allowed under the policy	-	⋮
change_time_limit	int	The time limit within which changes must be made	-	⋮
hours_since_purchase	int	The time elapsed since the ticket was purchased	-	⋮
modification_location	enum	The place where ticket modifications can be made	-	⋮
name_change_allowed	bool	A process to change the name on an airline ticket	-	⋮
name_correction_allowed	bool	A process to correct the name on an airline ticket	-	⋮
participating_airlines	enum	The airlines that allow changes under the Soar Customer First Pr...	-	⋮
partner_ticketing_support_contacted	bool	The department responsible for handling name changes or corre...	-	⋮
submission_method	enum	The method by which change requests must be submitted	-	⋮

Amazon Bedrock Guardrails

Multimodal toxicity detection

(preview)

Configurable safeguards for image content

- Enhance security of multimodal generative AI applications
- Available for all foundation models in Amazon Bedrock with image support
- Enable consistent policy control

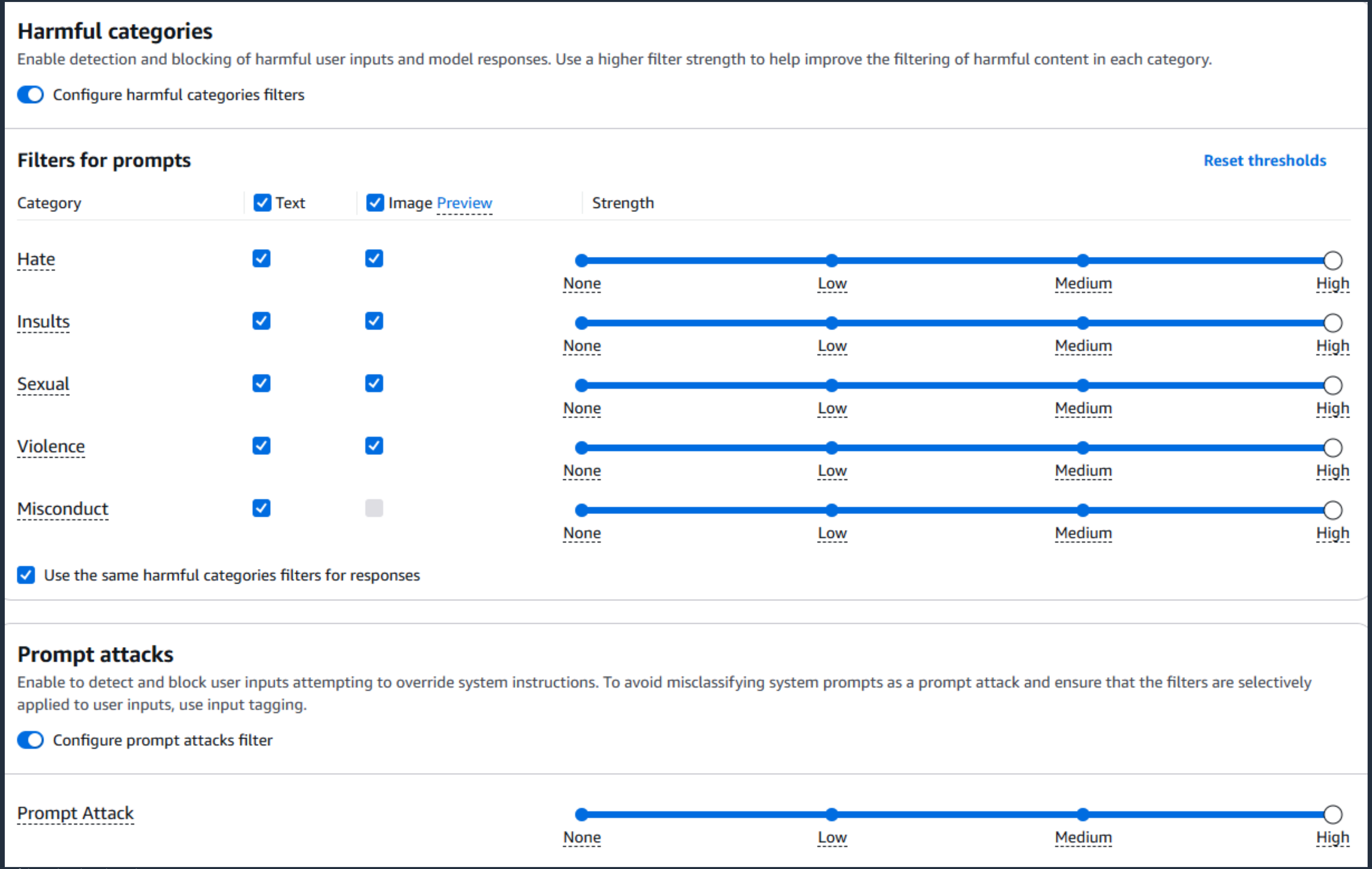
Content Filters – NEW: Image support!

CONFIGURE THRESHOLDS TO FILTER HARMFUL CONTENT

Filter harmful content across categories

- Hate
- Insults
- Sexual
- Violence
- Misconduct *
- Prompt attack *

* Text only



Harmful categories

Enable detection and blocking of harmful user inputs and model responses. Use a higher filter strength to help improve the filtering of harmful content in each category.

● Configure harmful categories filters

Filters for prompts Reset thresholds

Category	☑ Text	☑ Image Preview	Strength
Hate	☑	☑	None – Low – Medium – High ○
Insults	☑	☑	None – Low – Medium – High ○
Sexual	☑	☑	None – Low – Medium – High ○
Violence	☑	☑	None – Low – Medium – High ○
Misconduct	☑	□	None – Low – Medium – High ○

☑ Use the same harmful categories filters for responses

Prompt attacks

Enable to detect and block user inputs attempting to override system instructions. To avoid misclassifying system prompts as a prompt attack and ensure that the filters are selectively applied to user inputs, use input tagging.

● Configure prompt attacks filter

Prompt Attack	None – Low – Medium – High ○

Build and orchestrate agents

Amazon Bedrock

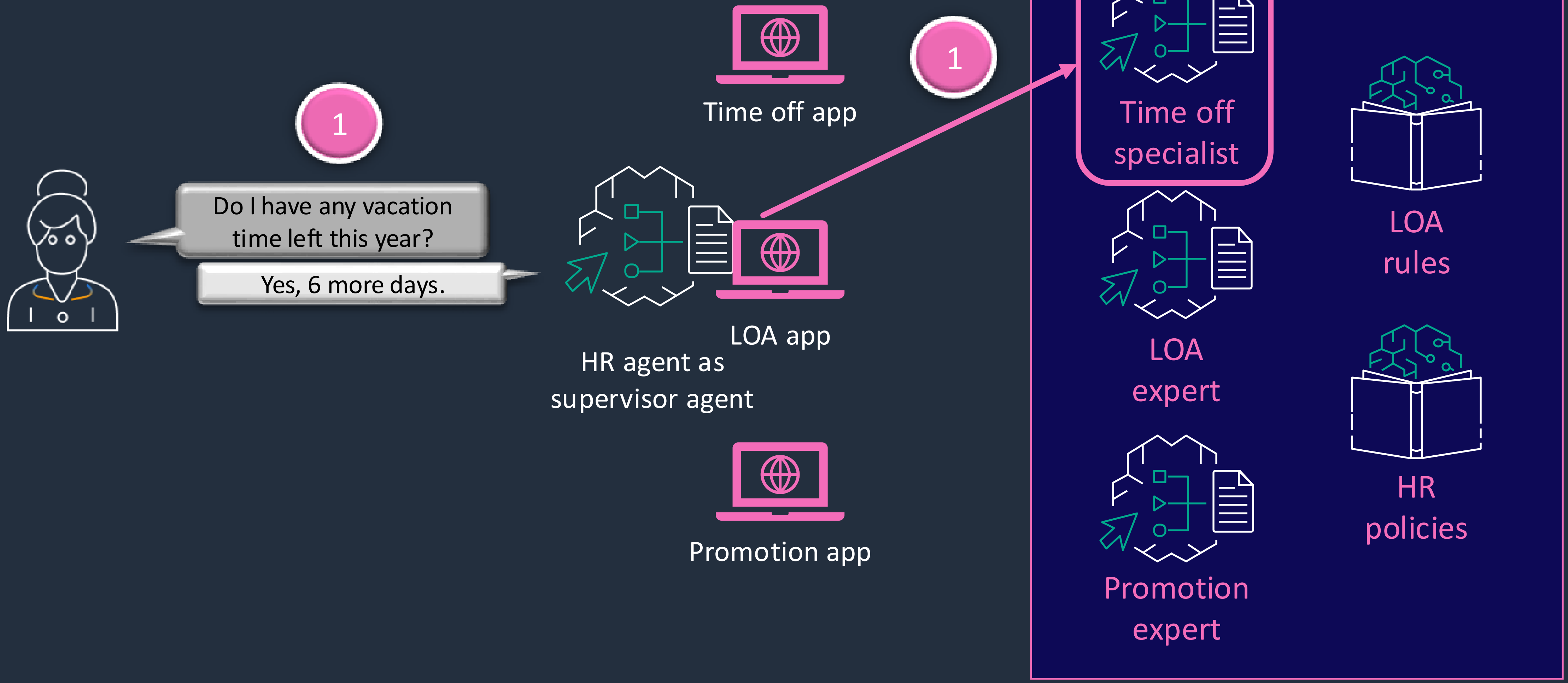
Multi-agent collaboration

(preview)

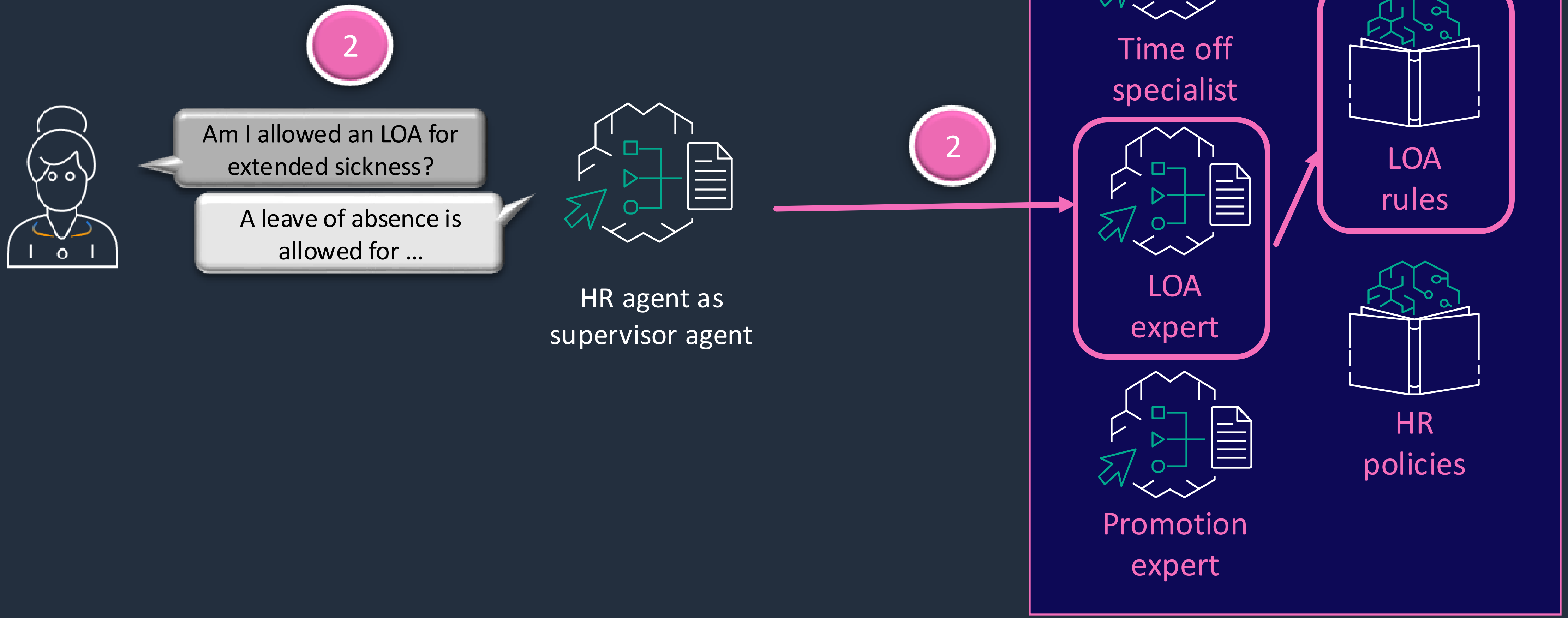
Easily build, deploy and orchestrate teams of agents that work together to handle complex, multi-step tasks

- Accelerate tasks with agents working in parallel
- Effortlessly orchestrate agents without complex coding

1. Unify customer experience



1. Unify customer experience


2
Am I allowed an LOA for extended sickness?
A leave of absence is allowed for …
HR agent as supervisor agent
2
Time off specialist
LOA expert
Promotion expert
LOA rules
HR policies

2. Automate complex processes with supervisor agents . . . .

Give me a complete marketing strategy for my new product XYZ . . .

Here’s the XYZ 10-page marketing strategy, including . . .

Orchestrate multi-step plan across agents

1

Generate plan dynamically

1. Conduct thorough market research, include competitors
2. Develop detailed project summary, target persona
3. Formulate comprehensive marketing strategy with goals, tactics, channels, KPIs
4. Create three innovative marketing campaign ideas
5. Develop detailed marketing copy for each campaign, including a video ad script, and a draft video
6. Produce final report and save interim results

Sub-agents and knowledge bases

Marketing strategist

Content creator

Style guide

Campaign feedback

Market analyst

Copy editor

Project storage agent

Video generation expert

2

Execute plan

Inline agents

Configure your agent dynamically at runtime

Quickly experiment with different agent's configurations

Enable subscription-based personalization for different customers

Select persona-based data sources at runtime

Dynamically select actions provided to your agent

Incorporate dynamic identity/behavior to your agent

Optimize for cost, latency and accuracy

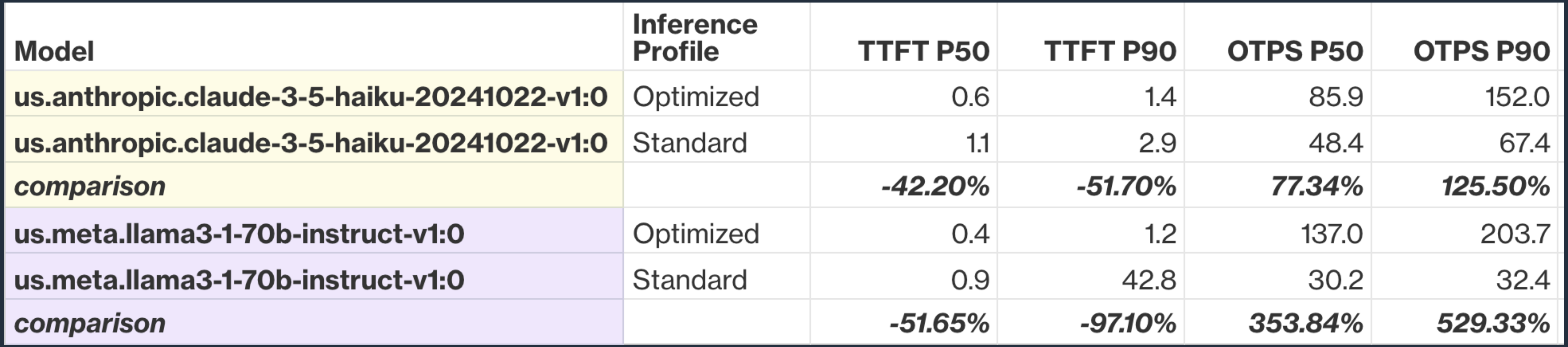
Internal testing of latency-optimized inference

3.5 Haiku

- Output tokens per second, 67 -> 152
- Time to first token, 1.1s -> 0.6s

Llama 3.1 70b

- Output tokens per second, 32 -> 203
- Time to first token, 0.9s -> 0.4s



Model	Inference Profile	TTFT P50	TTFT P90	OTPS P50	OTPS P90
us.anthropic.claude-3-5-haiku-20241022-v1:0	Optimized	0.6	1.4	85.9	152.0
us.anthropic.claude-3-5-haiku-20241022-v1:0	Standard	1.1	2.9	48.4	67.4
comparison		***-42.20%***	***-51.70%***	***77.34%***	***125.50%***
us.meta.llama3-1-70b-instruct-v1:0	Optimized	0.4	1.2	137.0	203.7
us.meta.llama3-1-70b-instruct-v1:0	Standard	0.9	42.8	30.2	32.4
comparison		***-51.65%***	***-97.10%***	***353.84%***	***529.33%***

Amazon Bedrock

supports prompt caching

(preview)

Cache repetitive context in prompts across multiple API calls

- Securely cache entire prompts
- Enhance accuracy through longer prompts
- Reduce cost by up to 90% and latency by up to 85% for supported models

Amazon Bedrock

Intelligent prompt routing

(preview)

Automatically route prompts to different foundation models to optimize response quality and lower costs

- Provides a single endpoint to efficiently route prompts
- Meets cost and latency thresholds with advanced prompt matching techniques
- Reduces application development costs by up to 30%

Amazon Bedrock

Model Distillation

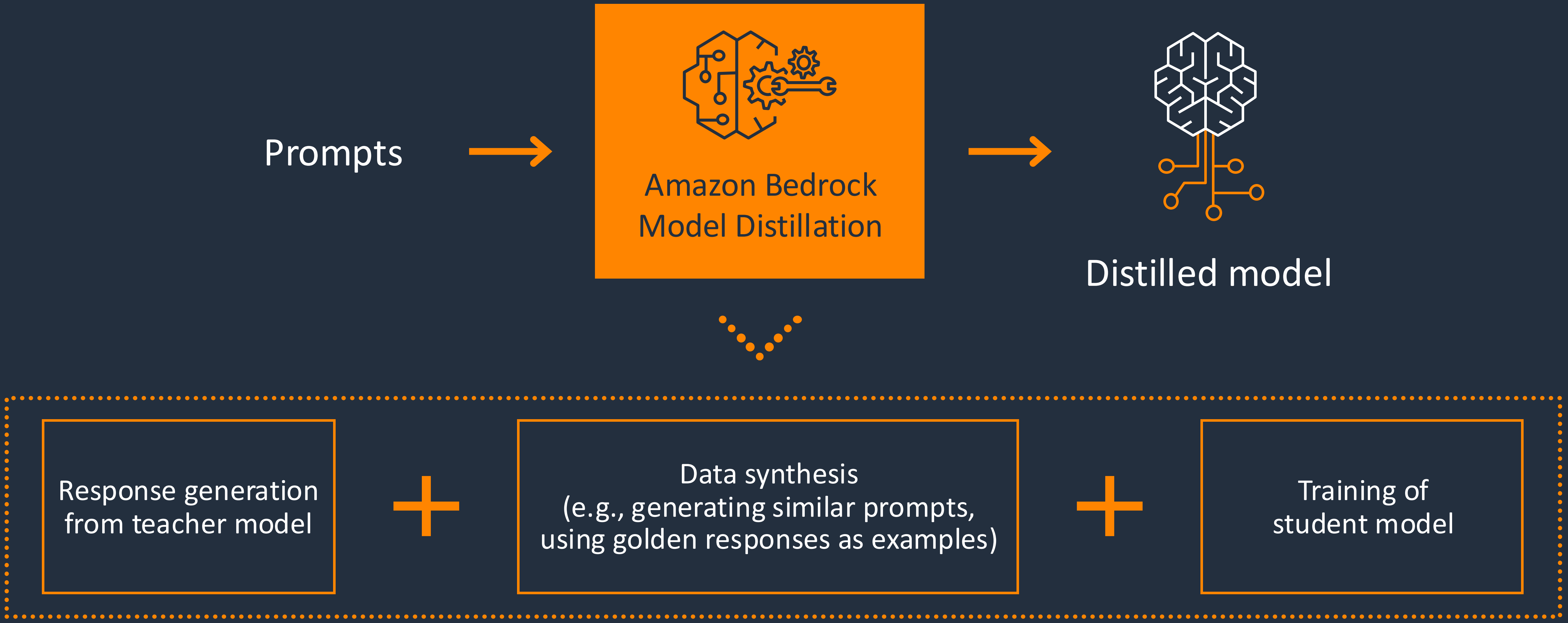
(preview)

Create smaller, faster, more cost-effective models

- Easily transfer knowledge from a large, complex model to a smaller one
- Distilled models up to 500% faster and up to 75% less expensive
- Anthropic, Meta, and Amazon models

Behind the scenes proprietary data synthesis

SINGLE WORKFLOW GENERATES TEACHER RESPONSES, ADDS PROPRIETARY DATA SYNTHESIS, AND TRAINS STUDENT MODEL, REMOVING ITERATIONS NEEDED TO IMPROVE DISTILED MODEL'S ACCURACY



Prompts

Amazon Bedrock Model Distillation

Distilled model

Response generation from teacher model

+

Data synthesis (e.g., generating similar prompts, using golden responses as examples)

+

Training of student model

Amazon Bedrock

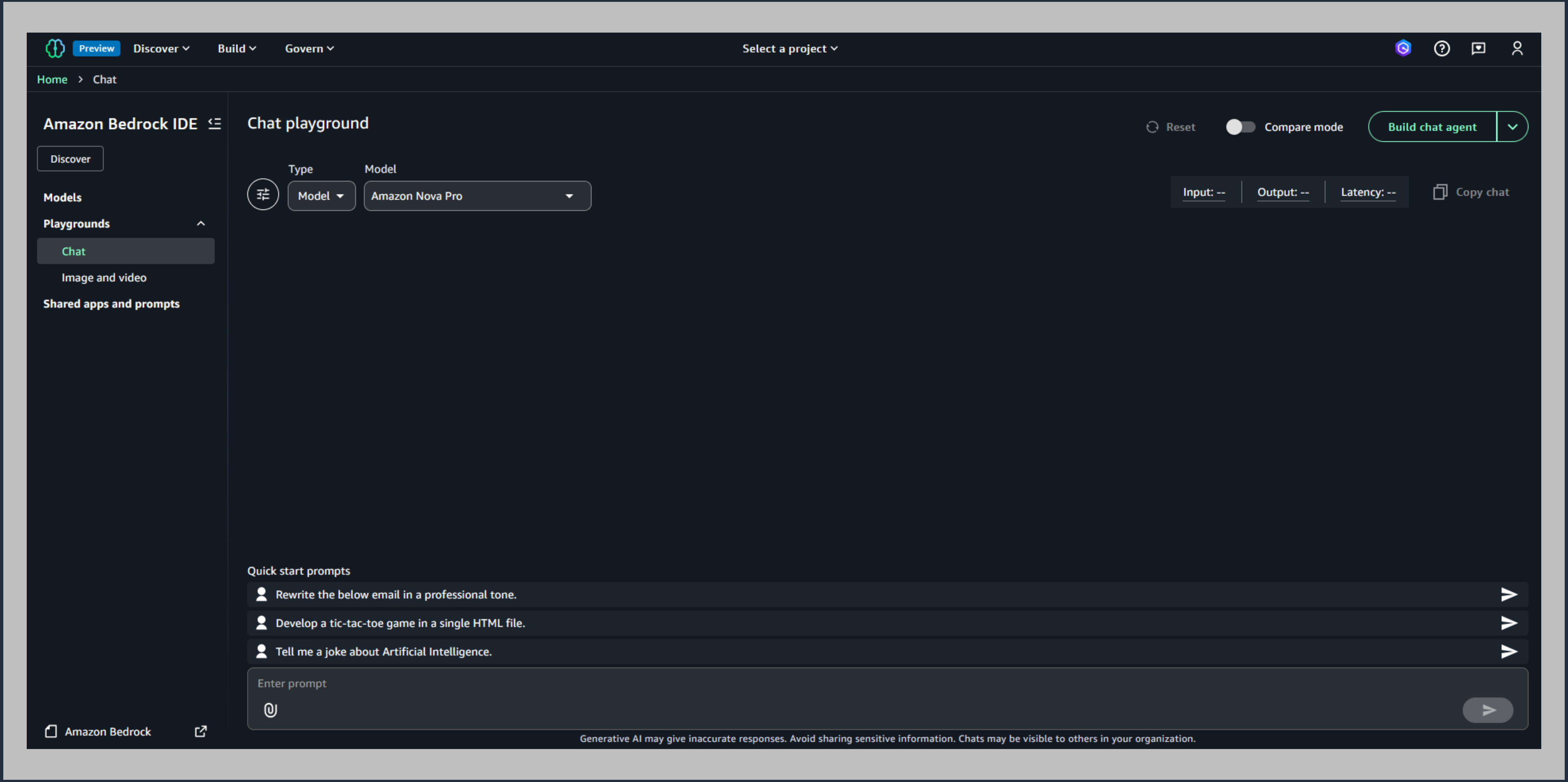
IDE

(preview)

Enable effortless generative AI development in a governed collaborative environment

- Facilitates custom AI application building with advanced features
- Promotes seamless collaboration among stakeholders
- Simplifies model evaluation and adoption through Playground experience

Amazon Bedrock IDE



Type: Model

Model: Amazon Nova Pro

Input: --

Output: --

Latency: --

Rewrite the below email in a professional tone.

Develop a tic-tac-toe game in a single HTML file.

Tell me a joke about Artificial Intelligence.

Generative AI may give inaccurate responses. Avoid sharing sensitive information. Chats may be visible to others in your organization.

Thank you!

Aparajithan Vaidyanathan

Principal Solutions Architect
AWS India