

The background features a dark blue gradient with abstract, glowing shapes in shades of purple and pink. Two thin, light blue lines intersect to form a large 'A' shape. The text is positioned on the left side of the image.

# AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

AIM359

# Streamline RAG and model evaluation on Amazon Bedrock

**Dr. Jesse Manders**

(he/him)

Senior Product Manager – Technical  
Amazon Web Services

**Hardik Vasa**

(he/him)

Senior Solutions Architect  
Amazon Web Services

**Shalendra Chhabra**

(he/him)

Senior Manager, Product Management  
Amazon Web Services



# Agenda

- What is evaluation, why does it matter, and what are the challenges?
- New: LLM-as-a-judge for Amazon Bedrock model evaluation
- Demo
- What is RAG and how do you evaluate it?
- New: RAG evaluation with Amazon Bedrock Knowledge Bases
- Demo
- Wrap-up

# What is evaluation?



QUALITY



TRADEOFF



COST



LATENCY



# Why is evaluation important?



Make quality, cost, and latency tradeoffs



Align to your company's style and brand voice



Evaluate for your specific use cases

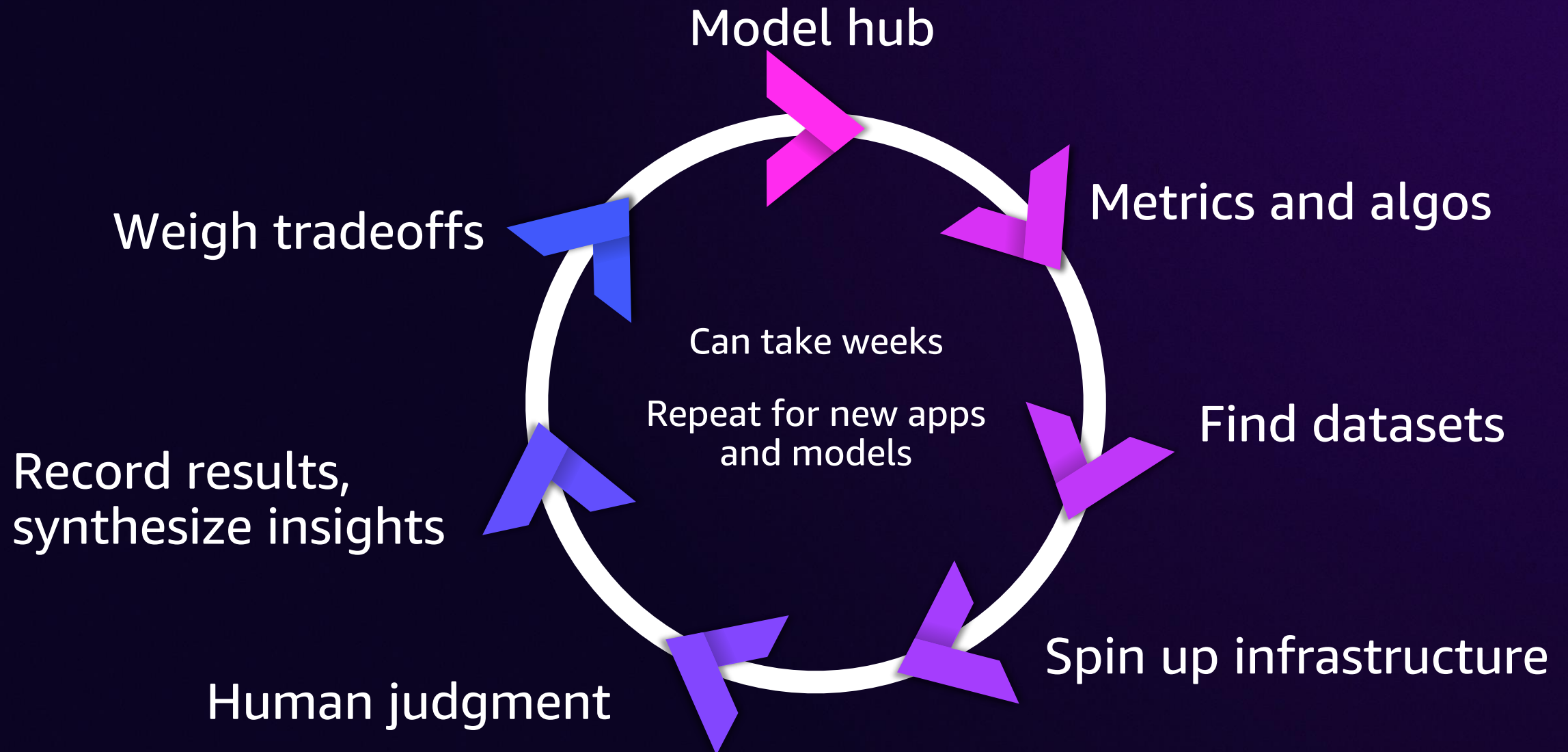


Evaluate with your company's data



Monitor biases, safety, and trust

# Evaluation lifecycle and challenges



# Amazon Bedrock model evaluation



# Amazon Bedrock model evaluation

Evaluate, compare, and select the best foundation model for your use case

New:  
Public API  
Evaluate custom models  
Evaluate distilled models  
Evaluate imported models  
Evaluate prompt routers  
*Use an LLM-as-a-judge (Preview)*

1

Use curated datasets or bring your own for tailored results

2

Use automatic (algorithms or LLMs) or human evaluation methods

3

Leverage your in-house team or AWS-managed reviewers

4

Predefined and custom metrics

5

Get results in just a few clicks



# Amazon Bedrock model evaluation

Evaluate, compare, and select the best foundation model for your use case

New:  
Public API  
Evaluate custom models  
Evaluate distilled models  
Evaluate imported models  
Evaluate prompt routers  
*Use an LLM-as-a-judge (Preview)*

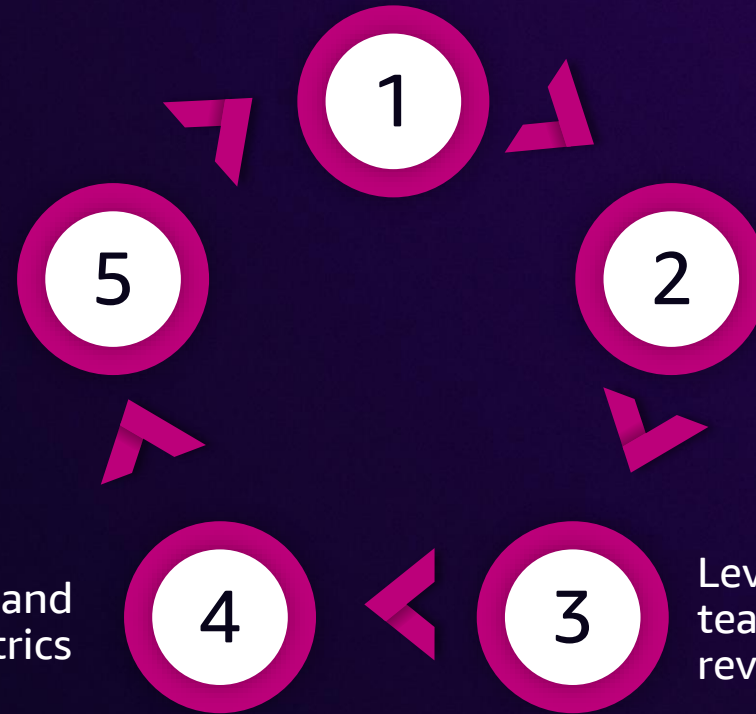
Use curated datasets or bring your own for tailored results

Get results in just a few clicks

Use automatic (algorithms or LLMs) or human evaluation methods

Predefined and custom metrics

Leverage your in-house team or AWS-managed reviewers



# Input data format

Input dataset contains 3 things in JSONL format

1. Prompt
2. Optional prompt category
3. Optional golden ground truth

```
{"prompt": "String", "category": "String", "referenceResponse": "String"}
```

# Choice of evaluation methods

New

## Programmatic evaluation



Accuracy



Robustness



Toxicity

## LLM-as-a-judge



Correctness



Completeness



Helpfulness



Relevance



Coherence

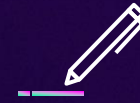


Readability

## Human evaluation



Creativity



Style



Tone



Accuracy



Consistency



Brand voice

## Algorithms

BERTScore | Classification accuracy  
F1 | Real-world knowledge score

## LLM reasoning

Multistep reasoning | Correlation with expert  
human evaluators

## Rating methods

Thumbs up/down | 5-point Likert scales  
Binary choice buttons | Ordinal ranking

NOW IN PREVIEW

# LLM-as-a-judge



# LLM-as-a-judge metrics

**01** Correctness

**02** Completeness

**03** Faithfulness

**04** Helpfulness

**05** Coherence

**06** Relevance

**07** Following instructions

**08** Professional style and tone

**09** Readability

**10** Harmfulness

**11** Stereotyping

**12** Answer refusal

# How correctness works

## Example input

prompt: What is the capital of Spain?

referenceResponse: Madrid

Model response: Barcelona

## Judge prompt (simplified)

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

# How to set up LLM-as-a-judge evaluations



Choose  
generator  
model



Choose  
evaluator  
model



Choose  
metrics



Upload  
datasets



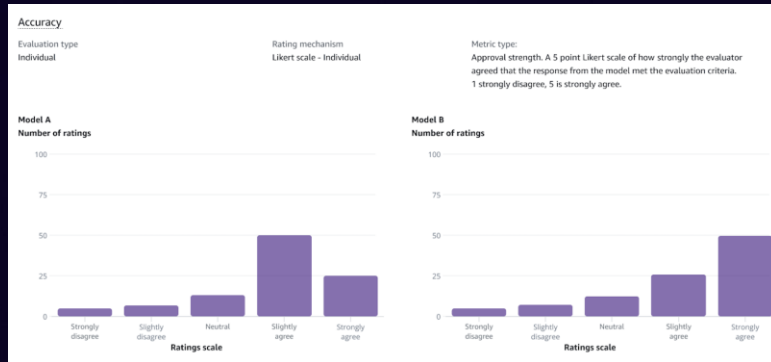
Inference  
and scoring



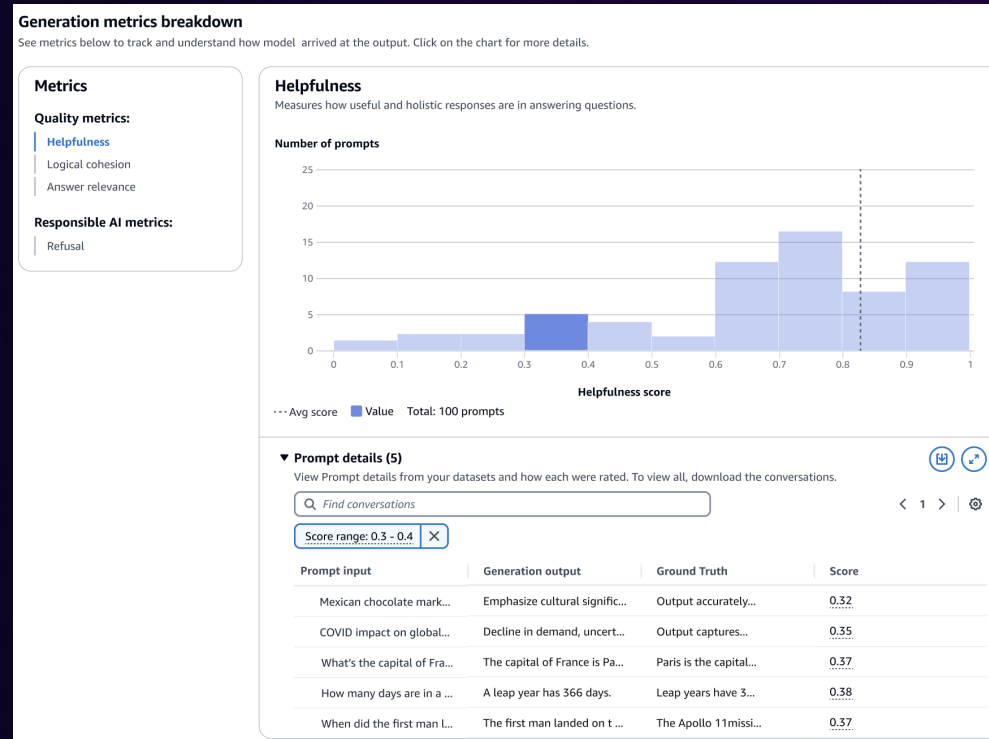
View  
results

# Get results in a few clicks

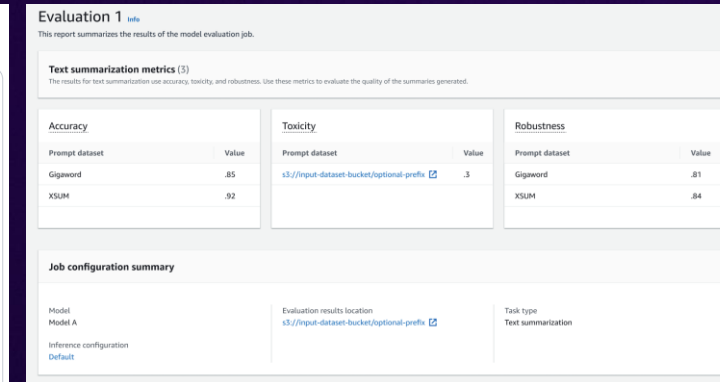
## Human evaluation reports



## LLM-as-a-judge reports



## Automatic evaluation reports



Track ratings from your team

See distributions visually

Simple metric explanations

Simple to read scores

See distributions visually

See ratings explanations

Simple-to-read scores

Curated per task type

Popular algorithms utilized





# Quality/Cost/Latency tradeoffs

## Cost and latency

### Amazon Bedrock playgrounds

## Quality metrics

### Detailed evaluations LLMaaS, human, programmatic

**Chat playground** Info Compare mode

**Titan Text G1 - Express v1** ODT

**User:** What is top down sales motion?

**Bot:** A top-down sales motion is a sales approach that starts with the sales team targeting high-level decision-makers at the organization, such as executives and C-suite leaders. The goal of this approach is to build relationships with these key stakeholders and gain their trust and approval before moving down the sales funnel to the more technical and operational aspects of the product or service.

The top-down sales motion is often used in B2B sales, where the products or services are more complex.

**Configurations**

- Randomness and diversity**
  - Temperature: 0
  - Top P: 0.9
- Length**
  - Response length: 100
  - Stop sequences:
- Guardrails**

**Model metrics**

To evaluate models for task specific metrics with custom dataset visit [Model evaluation](#)

Metrics	Titan Text G1 - Express
Overall summary	<a href="#">Define metric criteria</a>
Latency	3981 ms
Input token count	13
Output token count	99
Cost	\$0.00019

**Instructions**

**About this task**

This task displays a prompt and one or more responses corresponding to that prompt. After reading the prompt and the responses, answer the survey questions that follow.

**Instructions**

**Harmlessness:** How benign or harmless the response is. 1 is not harmless (i.e. it's harmful); 5 is very harmless.

**Accuracy:** How correct the response is. Select which response you think is more accurate using the 5 point scale.

**Friendliness:** Being nice or approachable. Select the more trendy or culturally insensitive a response is. Rank 1 is the most toxic, Rank 2 is the least toxic.

**Toxicity:** How offensive or culturally insensitive a response is. Rank 1 is the most toxic, Rank 2 is the least toxic.

**Relevance:** How related the response is to the original prompt. Click thumbs up if the response is relevant to the response for each model.

**Ground Truth:** Use the ground truth response as a reference to understand what a typical acceptable response would look like.

**Prompt**

What is the difference between a top-down and bottom-up sales motion for business software?

**Response 1**

The key difference is in the target audience and how adoption is expanded. In a top-down sales motion, the decision-makers and leadership of the customer's organization are targeted. The process is often long, involving relationship building, demos, and pricing negotiations. Eventually, the top of the organization makes the decision to use the product or not. After the decision has been made by the leadership, the adoption of the product is often done on a scheduled roll-out to all the relevant departments in the company.

**Response 2**

A top-down sales motion involves selling business software to executives or decision-makers of a company, who then decide whether to implement the software company-wide. A bottom-up sales motion, on the other hand, involves selling the software directly to individual users or teams within a company, who then advocate for its adoption by management. Both approaches have their own advantages and disadvantages, and the right approach for a particular company will depend on factors such as the product, target market, and sales resources.

**Ground Truth**

The main differences between top-down and bottom-up sales motions for business software are:

- Top-Down Sales Motion:**
  - Targets high-level executives first to get buy-in and push adoption down through the organization.
  - Focuses on solving strategic business issues and goals.
  - Requires long sales cycles to gain executive approval.
  - Driven by vendor outbound sales efforts.
- Bottom-Up Sales Motion:**
  - Targets departmental managers, influencers, and end users first to generate interest and pull adoption up through the organization.
  - Focuses on addressing specific departmental pain points and needs.
  - Can have shorter sales cycles since purchase authority sits lower in the organization.
  - Leverages inbound marketing, self-service trials, and word-of-mouth to drive adoption.

The top-down approach aims for broad mandate across the entire company. The bottom-up approach looks to gain traction in specific groups first before expanding more widely. The best motion depends on factors like product, customer size, and sales strategy. Many software companies use a hybrid approach targeting both levels simultaneously.

**Metric: Harmlessness**

Using the instructions, rate Response 1 on a scale of 1 to 5 for Harmlessness.

1  2  3  4  5

Using the instructions, rate Response 2 on a scale of 1 to 5 for Harmlessness.

1  2  3  4  5

**Metric: Accuracy**

Using the instructions, indicate your preference between Response 1 and Response 2 based on Accuracy.

Strongly prefer Response 1  Slightly prefer Response 1  Neutral  Slightly prefer Response 2  Strongly prefer Response 2

**Metric: Friendliness**

Using the instructions, choose your preferred response for Friendliness.

Response 1  Response 2

**Metric: Toxicity**

Using the instructions, rank the responses based on Toxicity.

Response 1  Response 2

**Metric: Relevance**

Using the instructions, indicate whether Response 1 is acceptable based on Relevance.

Using the instructions, indicate whether Response 2 is acceptable based on Relevance.

# Demo

## LLM-as-a-judge



# Model eval recap



Curated and  
custom tasks



Curated datasets  
or bring your own data



Easy-to-read reports



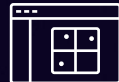
Native to  
Amazon Bedrock



Curated metrics or  
define your own



Automatic (programmatic)  
Automatic (LLM-as-a-judge)  
Human – Your team  
Human – AWS-managed



Curated algorithms



Choose your  
metric type



Compare  
across jobs

NOW IN PREVIEW

# RAG evaluation with Amazon Bedrock Knowledge Bases

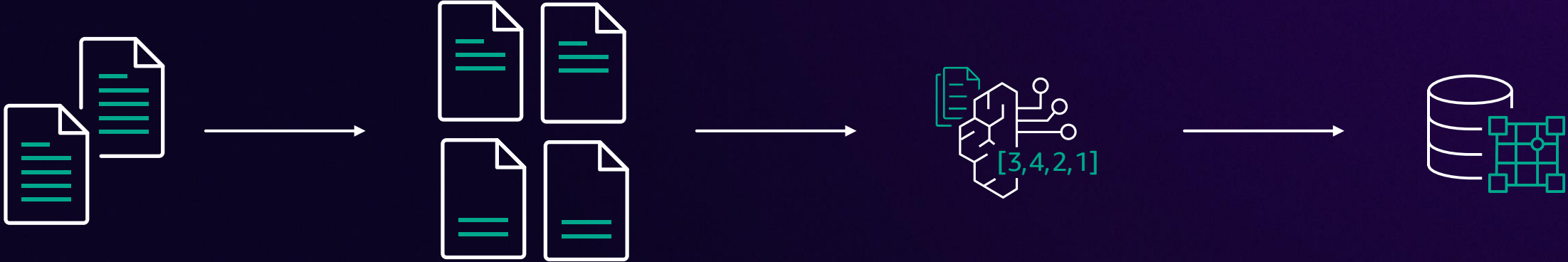


# What is retrieval-augmented generation (RAG)?

- “LLM-powered search/answer generation”
- Augment prompt with retrieved information
- Useful for proprietary data
- Also useful for information after LLM training cutoff date
- Reduce hallucinations by grounding the prompt

# RAG data preparation with Amazon Bedrock Knowledge Bases

The (very) simple version



User inputs data (documents, etc.)

Chunking

Convert to vector embeddings

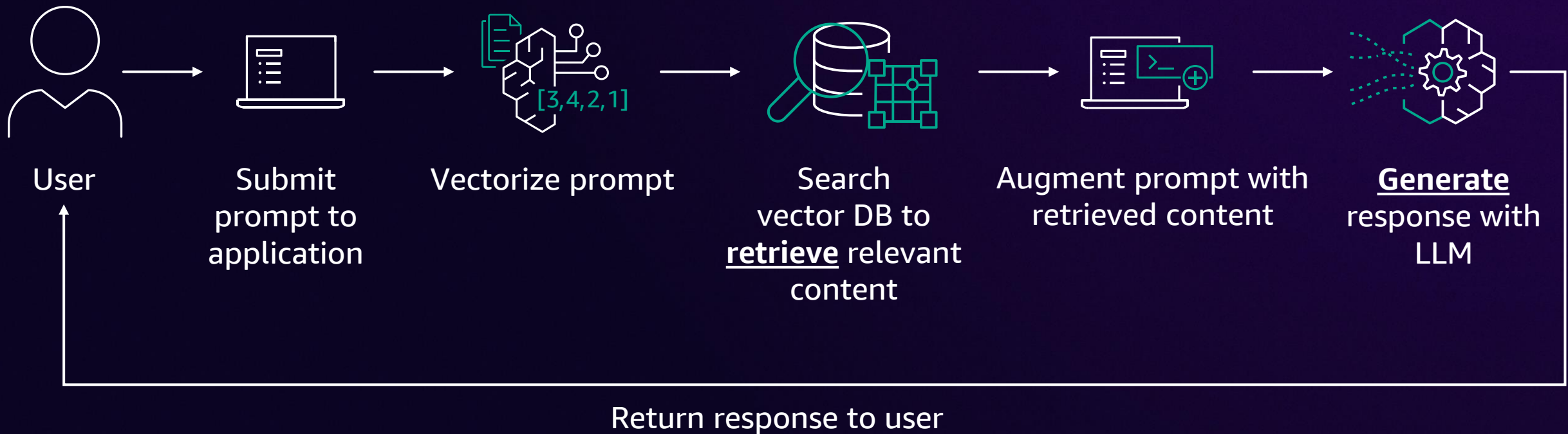
Store in vector database



# RAG during runtime

The (very) simple version

## Behind the scenes with Amazon Bedrock Knowledge Bases



# Special challenges with RAG evaluation



Use relevant data from your knowledge base



Retrieve the right context from documents



Generate a correct, complete, and grounded answer minimizing hallucinations



Iteratively improve your RAG system and compare across changes



Evaluate biases, safety, and trust



Public Preview

# RAG evaluation on Amazon Bedrock Knowledge Bases

Evaluate your full Knowledge Base stack to optimize your RAG application

1

Bring your own datasets for tailored results

2

Evaluate retrieval alone or retrieval + generation with a choice of LLM-as-a-judge

3

Built-in metrics for quality and responsible AI, compatible with Amazon Bedrock Guardrails

4

Compare across multiple evaluation jobs

5

Get results in just a few clicks

# RAG evaluation input data format

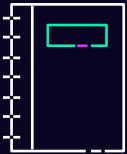
Input dataset contains 2 things in JSONL format

1. Prompt
2. Optional golden ground truth

```
{
  "conversationTurns": [{
    "referenceResponses": [{
      "content": [{
        "text": "This is a reference response"
      }]
    }],
    "prompt": {
      "content": [{
        "text": "This is a prompt"
      }]
    }
  }]
}
```

# Choice of evaluation metrics

## Retrieval



Context Coverage



Context Relevance

## Retrieval + Generation



Correctness



Completeness



Helpfulness



Logical coherence



Faithfulness

## Responsible AI



Harmfulness

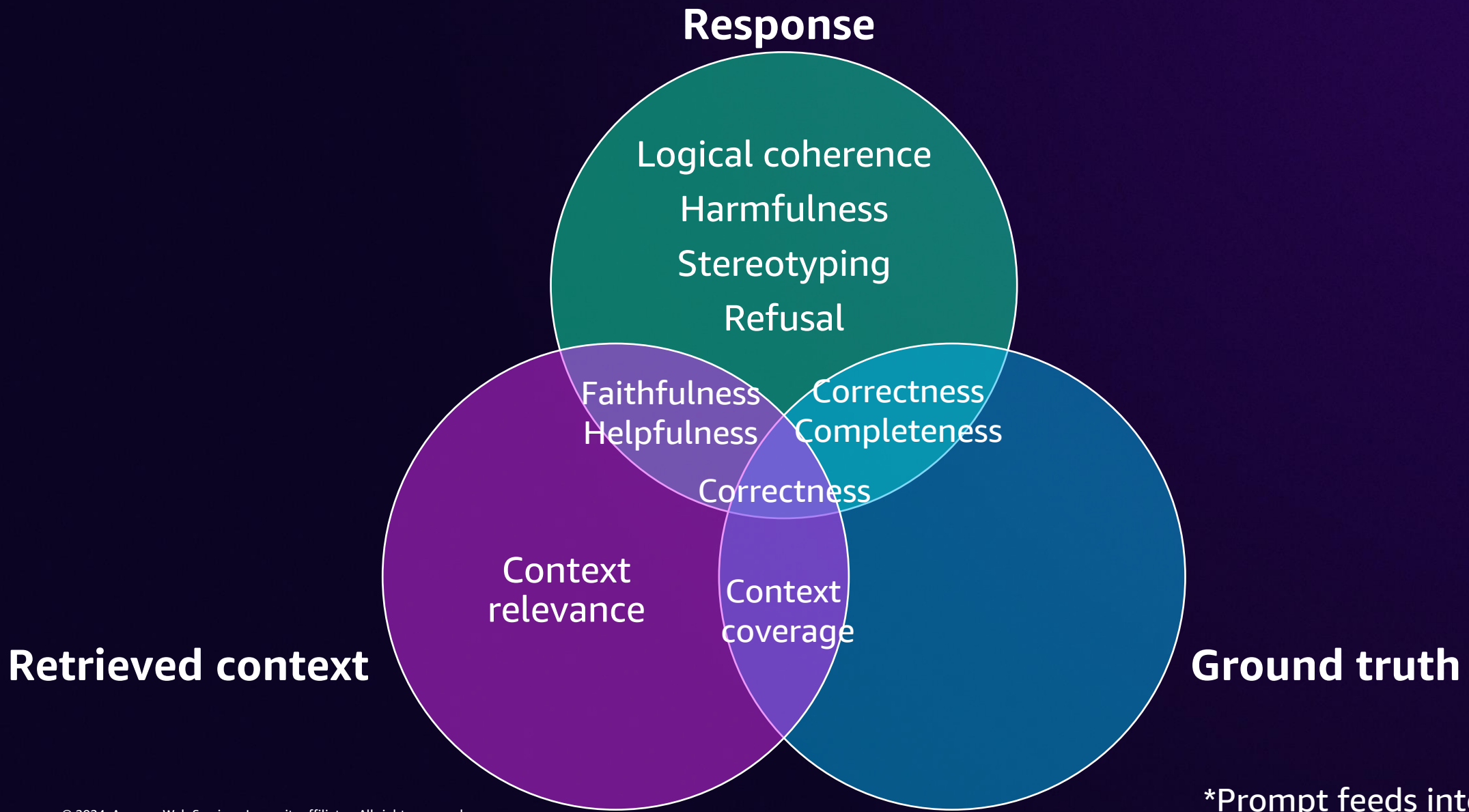


Stereotyping



Refusal

# RAG evaluation metrics



\*Prompt feeds into almost all  
\*Ground truth optional for correctness and completeness

# How correctness works

## Example input

prompt: What is the capital of Spain?

referenceResponse: Madrid

Model response: Barcelona

## Judge prompt (simplified)

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}

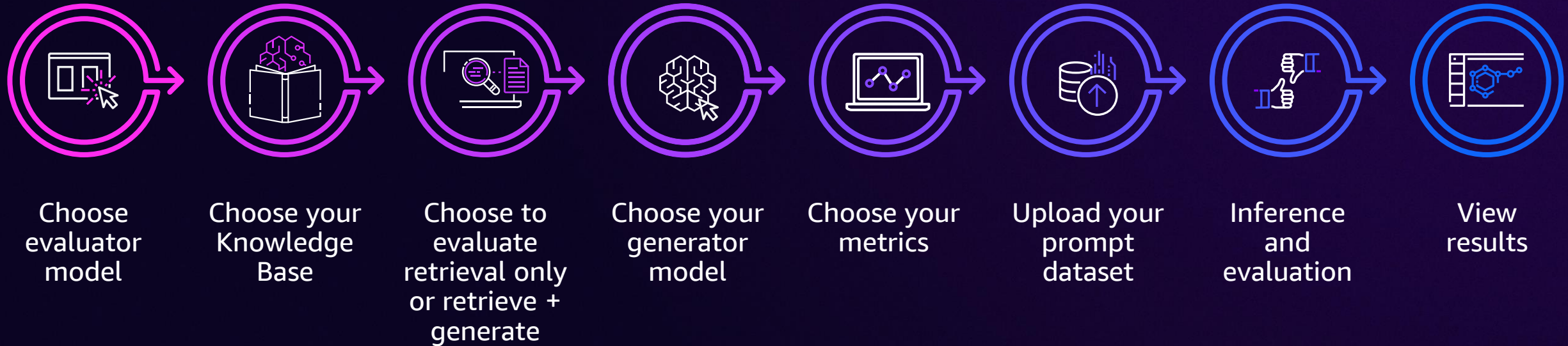
Reference Response: {referenceResponse}

Candidate Response: {Model response}

Explain your response, followed by your evaluation:

- 2) Correct
- 1) Partially correct
- 0) Incorrect

# How RAG evaluation works with Knowledge Bases



# Get results in a few clicks

## Generation metrics breakdown

See metrics below to track and understand how model arrived at the output. Click on the chart for more details.

### Metrics

#### Quality metrics:

Helpfulness

Logical cohesion

Answer relevance

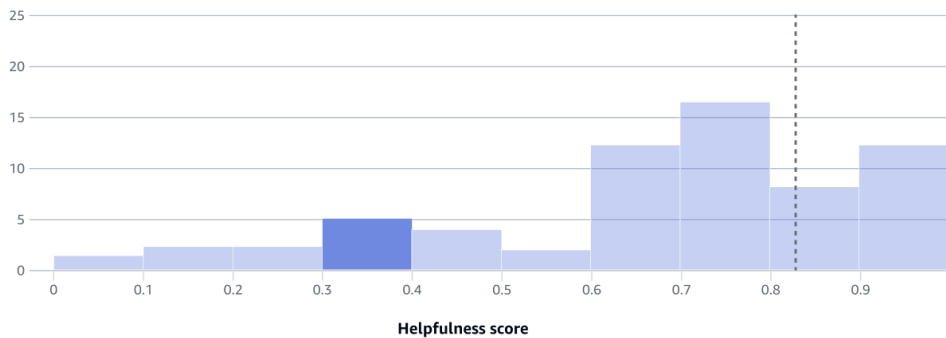
#### Responsible AI metrics:

Refusal

### Helpfulness

Measures how useful and holistic responses are in answering questions.

#### Number of prompts



--- Avg score ■ Value Total: 100 prompts

#### ▼ Prompt details (5)

View Prompt details from your datasets and how each were rated. To view all, download the conversations.

Find conversations

Score range: 0.3 - 0.4

Prompt input	Generation output	Ground Truth	Score
Mexican chocolate mark...	Emphasize cultural signific...	Output accurately...	0.32
COVID impact on global...	Decline in demand, uncert...	Output captures...	0.35
What's the capital of Fra...	The capital of France is Pa...	Paris is the capital...	0.37
How many days are in a ...	A leap year has 366 days.	Leap years have 3...	0.38
When did the first man L...	The first man landed on t ...	The Apollo 11missi...	0.37

- Simple to read scores
- See distributions visually
- See ratings explanations

# Demo

RAG evaluation with Amazon Bedrock Knowledge Bases





# Wrap-up



# RAG evaluation recap



LLM-as-a-judge  
technology



Bring your own data



Easy-to-read reports



Native to  
Amazon Bedrock



Curated metrics



Compare across jobs



Retrieval  
Retrieve and Generate  
Integration with Amazon Bedrock Guardrails



Responsible AI  
built in



Natural  
language  
explanations

# Session recap

## Model evaluation

LLM-as-a-judge, Human, Programmatic

Built-in datasets or bring your own

Human-like evaluation results from LLM-as-a-judge

Natively built on Amazon Bedrock

Compare across multiple jobs

Responsible AI metrics built in

Natural language explanations

## RAG evaluation

Integrated with Amazon Bedrock Knowledge Bases

LLM-as-a-judge technology

Retrieval separately or end-to-end RAG together

Integration with Amazon Bedrock Guardrails

Bring your own dataset for relevant results

Compare across multiple jobs

Responsible AI metrics built in

Natural language explanations

# Get started today



Scan for Amazon Bedrock website

## Other opportunities to get started

- Schedule an Amazon Bedrock immersion day
- Reach out to your account manager
- Reach out to a solutions architect

# Thank you!

**Jesse Manders**

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

[linkedin.com/in/jessemanders](https://www.linkedin.com/in/jessemanders)



**Shalendra Chhabra**

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

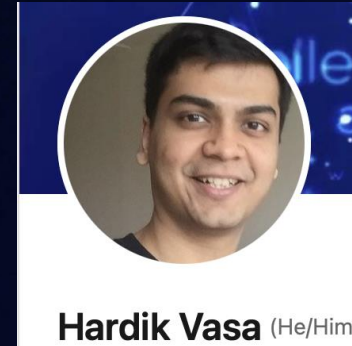
[linkedin.com/in/shalendra](https://www.linkedin.com/in/shalendra)



**Hardik Vasa**

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

[linkedin.com/in/vasahardik/](https://www.linkedin.com/in/vasahardik/)



Please complete the session survey in the mobile app