

Amazon Bedrock

***The easiest way to build and scale generative AI applications
with foundation models***

Michael Lin

Sr. Solutions Architect
Amazon Web Services



GENERALLY AVAILABLE

Amazon *Bedrock*

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



Access a range of leading FMs through a single API



Privately customize FMs using your organization's data



Build agents that execute complex business tasks by dynamically invoking APIs

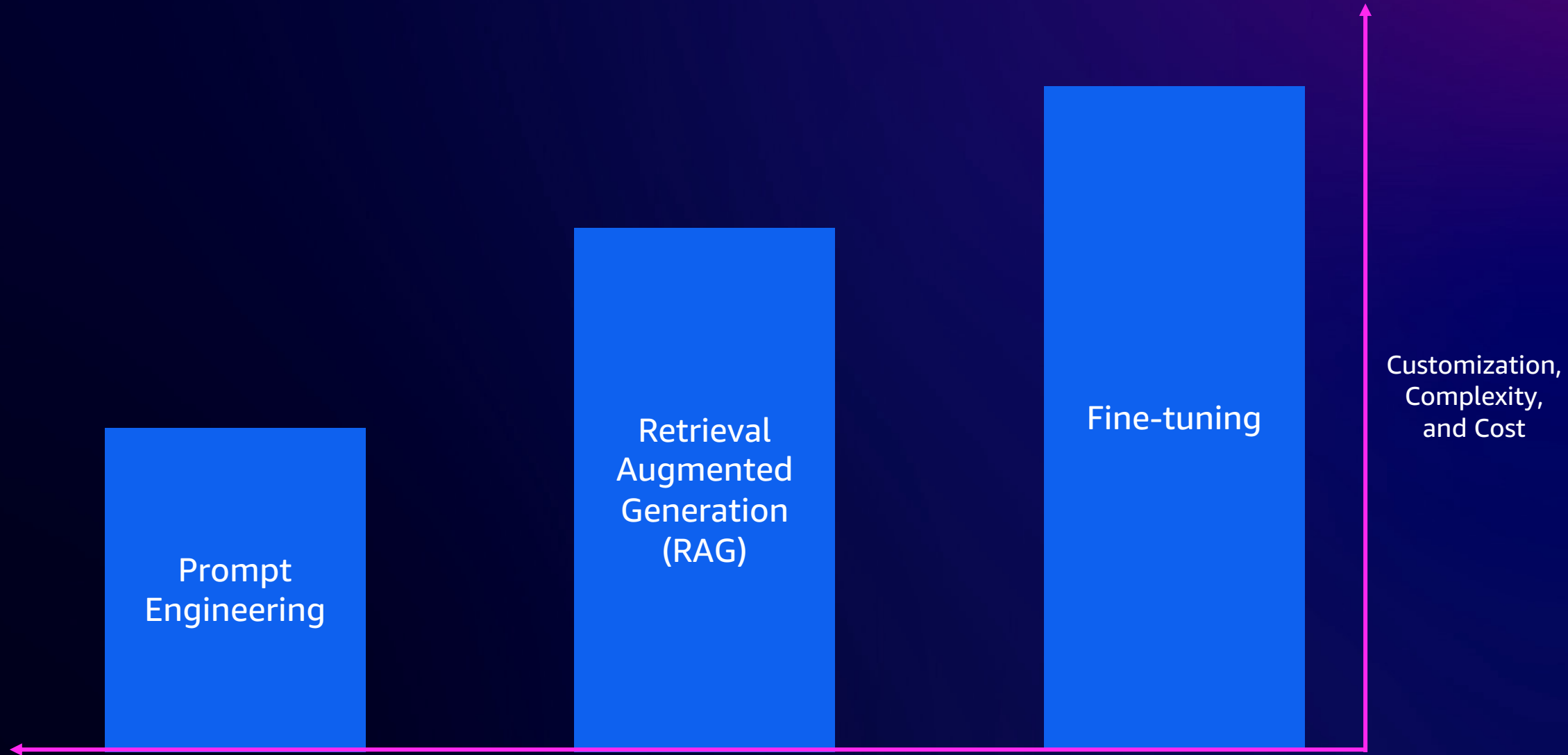


Extend the power of FMs with your data using Retrieval Augmented Generation (RAG)



Enable data security and compliance

Your customization journey



What is Prompt Engineering? – An Example

What is $10 + 10$?

$10 + 10 = 20$

$1 + 1$ is an addition problem.
 $1 - 1$ is a subtraction problem.
 1×1 is a multiplication problem.
 $1 / 1$ is a division problem.

What is $10 + 10$?

$10 + 10$ is an addition problem

Use examples

- Examples are probably the single most effective tool for getting Claude to behave as desired
- Make sure to give Claude examples of common edge cases
- Generally more examples = more reliable responses at the cost of latency and tokens

Example:

Human: I will give you some quotes. Please extract the author from the quote block.

Here is an example:

<example>

Quote:

“When the reasoning mind is forced to confront the impossible again and again, it has no choice but to adapt.”

— N.K. Jemisin, The Fifth Season

Author: N.K. Jemisin

</example>

Quote:

“Some humans theorize that intelligent species go extinct before they can expand into outer space. If they're correct, then the hush of the night sky is the silence of the graveyard.”

— Ted Chiang, Exhalation

Author:

Assistant Ted Chiang

Knowledge Bases for Amazon Bedrock

Gives FMs and agents contextual information from your private data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, accurate, and customized responses.



Fully managed support for end-to-end RAG workflow



Securely connect FMs and agents to data sources



Easily retrieve relevant data and augment prompts



Provide source attribution

Knowledge Base in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

TestDelete

Knowledge base overview

Edit

Knowledge base name

knowledge-base-cwa

Knowledge base ID

CJOMCDOKNC

Knowledge base description

—

Status

Ready

Service Role

AmazonBedrockExecutionRoleForKnowledgeBase_cwa

Created date

February 29, 2024, 20:13 (UTC+08:00)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key

Value

113年度新聞媒體從業人員氣象資訊推廣說明會，總共有幾場。

113年度新聞媒體從業人員氣象資訊推廣說明會總共有4場。^[1]

Show source details >

北區場次的時間，與地點。

113年度新聞媒體從業人員氣象資訊推廣說明會北區場次的時間是2月29日(星期四),地點是中央氣象署617會議室(臺北市中正區公園路64號6樓)。^[1]

Show source details >



Knowledge Base in Action

Amazon Bedrock > Knowledge base > knowledge-base-cwa

knowledge-base-cwa

TestDelete

Knowledge base overview

Edit

Knowledge base name

knowledge-base-cwa

Knowledge base description

—

Service Role

AmazonBedrockExecutionRoleForKnowledgeBase_cwa

Knowledge base ID

CJOMCDOKNC

Status

Ready

Created date

February 29, 2024, 20:13 (UTC+08:00)

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value
-----	-------

111年度預算執行情形。

111年度本署歲入預算共編列2,897萬元,決算數為2,929萬4千元,決算數占預算數101.12%。111年度本署歲出預算加計動支第一、二預備金共21億5,114萬5千元,決算數為21億4,675萬元,決算數占預算數99.8%。^[1]

Show source details >

113年度施政計畫。請提供摘要。

根據第3號搜索結果,113年度中央氣象署的4個主要施政計畫為:1)強化氣象觀測 2)精準預報技術發展 強化預報



Amazon Bedrock custom models

New!

Create custom models using the console or APIs

Maximize accuracy of FMs by providing labeled or raw unlabeled data

Once deployed, custom models are invoked the same way as base models (playground or API)

Customizations now supported for Amazon Titan and some third party FMs


The screenshot shows the Amazon Bedrock Custom models console. At the top, it says "Amazon Bedrock > Custom models". Below this is the heading "Custom models" with an "Info" link. A subtitle reads "Customize model with fine-tune or continual pre-training." A section titled "How it works" contains two steps: "Step 1. Customize a model" (with a fine-tuning icon) and "Step 2. Purchase Provisioned Throughput" (with a throughput icon). Below this, there are tabs for "Models" and "Training jobs". The "Training jobs" tab is active, showing "Training jobs (6)" and a search bar. At the bottom, there is a table with columns: Job Name, Status, Customization type, Model metrics, Source Model, Provider, and Custom mo.

Amazon Bedrock > Custom models

Custom models [Info](#)

Customize model with fine-tune or continual pre-training.


▼ How it works



Step 1. Customize a model [Info](#)

Fine-tuning - Improve model accuracy for specific tasks with labeled data, using smaller datasets and less training time.

Continued Pre-training - Updates some parameters. Retains prior learning and adds new knowledge. The model is regularly updated with new data, which is more efficient than fully retraining a large model.



Step 2. Purchase Provisioned Throughput [Info](#)

After a custom model is created, you need to purchase Provisioned Throughput to load custom models in playground.

[Purchase Provisioned Throughput](#)

Models | **Training jobs**

Training jobs (6)

Training jobs are initiated when you begin fine-tuning a model.

[Refresh](#) [Stop job](#) [Customize model ▼](#)

< 1 2 3 4 > [Settings](#)

Job Name ▼	Status ▼	Customization type ▼	Model metrics ▼	Source Model ▼	Provider ▼	Custom mo
------------	----------	----------------------	-----------------	----------------	------------	-----------

Fine-Tuning in Action: Text Summarization

```
prompt = """"
Summarize the simplest and most interesting part of the following conversation.

#Person1#: Hello. My name is John Sandals, and I've got a reservation.
#Person2#: May I see some identification, sir, please?
#Person1#: Sure. Here you are.
#Person2#: Thank you so much. Have you got a credit card, Mr. Sandals?
#Person1#: I sure do. How about American Express?
#Person2#: Unfortunately, at the present time we take only MasterCard or VISA.
#Person1#: No American Express? Okay, here's my VISA.
#Person2#: Thank you, sir. You'll be in room 507, nonsmoking, with a queen-size bed. Do you approve, sir?
#Person1#: Yeah, that'll be fine.
#Person2#: That's great. This is your key, sir. If you need anything at all, anytime, just dial zero.

Summary:
""""

body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}
```

Fine-Tuning in Action: Baseline Completion

```
response = bedrock_runtime.invoke_model(  
    modelId="meta.llama2-13b-chat-v1", # compare to chat model  
    body=json.dumps(body)  
)  
  
response_body = response["body"].read().decode('utf8')  
print(json.loads(response_body)["generation"])
```

A man named John Sandals checks into a hotel and provides identification and a credit card. The hotel only takes MasterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.

Fine-Tuning in Action: Improved Completion

```
response = bedrock_runtime.invoke_model(  
    modelId=provisioned_model_arn, # custom fine-tuned model  
    body=json.dumps(body)  
)  
  
response_body = response["body"].read().decode('utf8')  
print(json.loads(response_body)["generation"])
```

John Sandals checks in the hotel with VISA and is assigned room 507, nonsmoking, with a queen-size bed.



Agents for Amazon Bedrock

Enable generative AI applications
to execute multi-step business tasks
using natural language

Generally available

Features

- Uses power of LLM's to prompt and respond using natural language
- Breaks down and orchestrates tasks
- Completes tasks by dynamically invoking APIs
- Securely and privately accesses company data
- Surfaces chain-of-thought trace and underlying agent prompts

Agent in Action

Hi, I am looking for shoes

Could you please provide your name so I can look up your account?
[Show trace >](#)

Jane Smith

I found two hiking boots in stock - ShoeID 2: Durable leather hiking boots in Brown or Black for \$240 and ShoeID 5: Waterproof hiking boots in Tan or Olive for \$99.99. Which would you like to order?
[Show trace >](#)

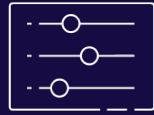
Place an order for Shoe ID 5

I have placed an order for ShoeID 5 for you. The waterproof hiking boots in tan or olive green should be on their way soon! Let me know if you need any other assistance with finding shoes.
[Show trace >](#)

PREVIEW

Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies



Apply guardrails to multiple foundation models and Agents for Amazon Bedrock



Configure harmful content filtering based on your responsible AI policies



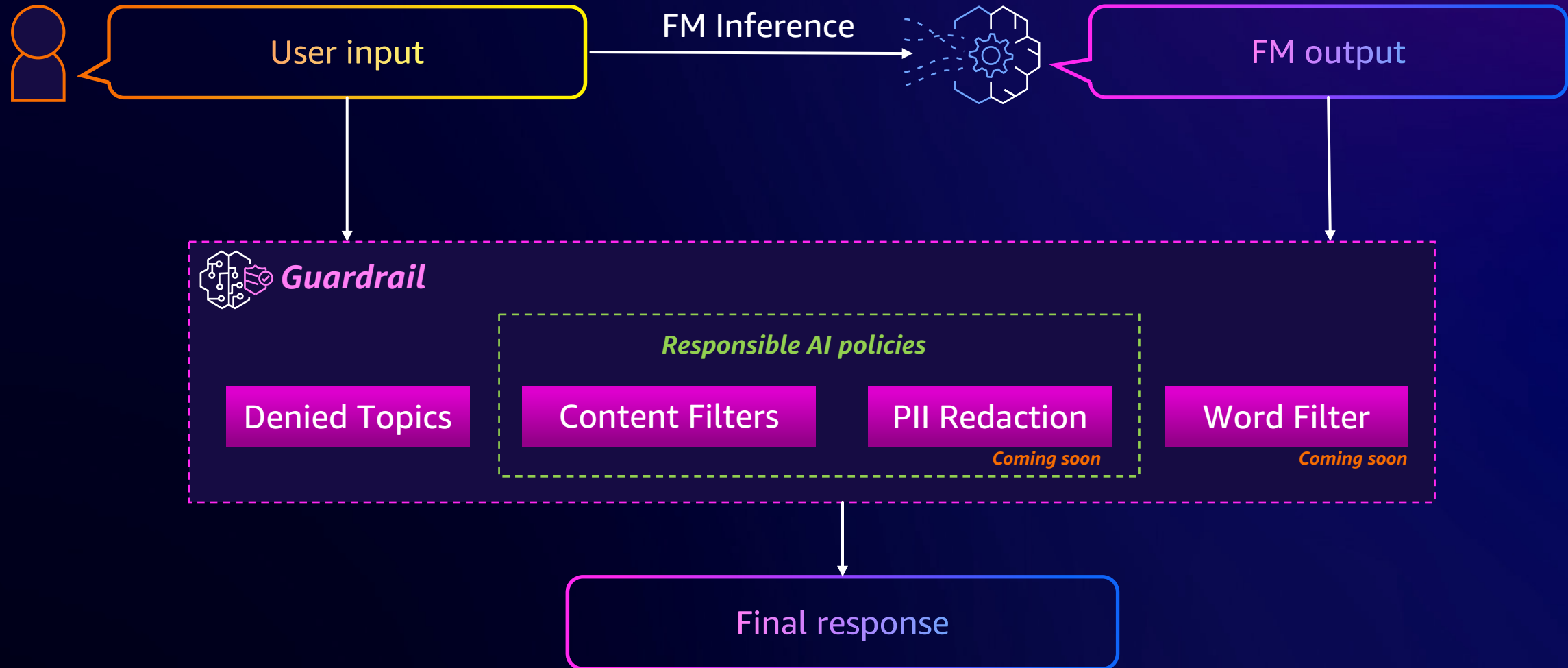
Define and disallow denied topics with short natural language descriptions



COMING SOON

Redact sensitive PII information in FM responses

How it works: Guardrails for Amazon Bedrock



New

PREVIEW

Model evaluation on Amazon Bedrock

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

1

Use curated datasets or bring your own for tailored results

2

Use automatic 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



Model Evaluation in Action

Human evaluation reports

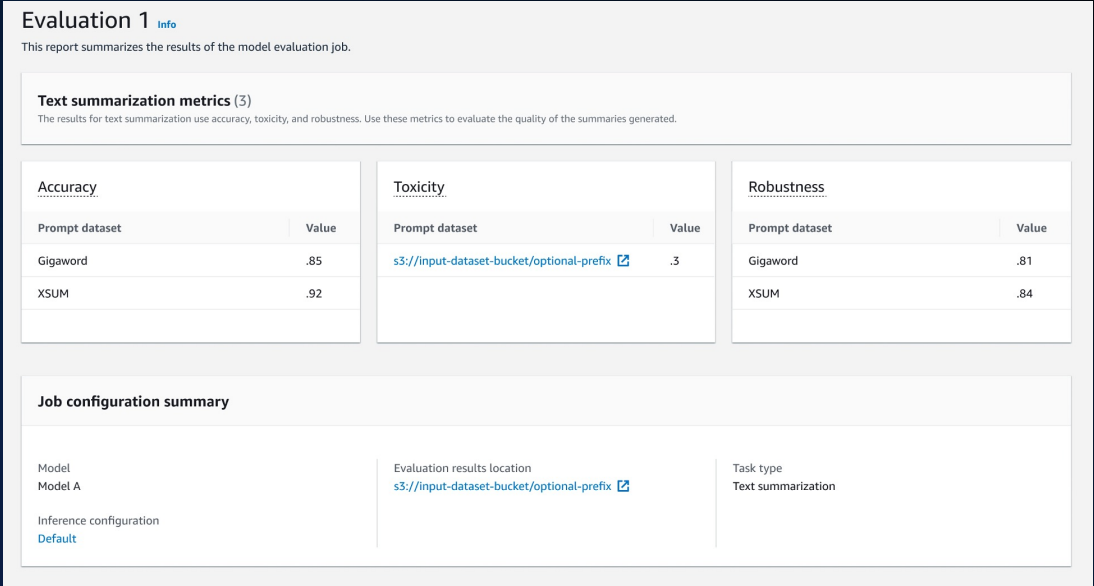


Track ratings from your team

See distributions visually

Simple metric explanations

Automatic evaluation reports



Simple-to-read scores

Curated per task type

Popular algorithms utilized

Thank you!

Michael Lin

linmicht@amazon.com



Please complete the session
survey in the mobile app

