**Sử dụng chỉ số tùy chỉnh để đánh giá ứng dụng AI tạo sinh với Amazon Bedrock**

Bởi Shreyas Subramanian, Adewale Akinfaderin, Ishan Singh, và Jesse Manders | ngày 06 tháng 5 năm 2025

Với [Amazon Bedrock Evaluations](https://aws.amazon.com/bedrock/evaluations/), bạn có thể đánh giá các mô hình nền tảng (FMs) và hệ thống Retrieval Augmented Generation (RAG), dù được lưu trữ trên [Amazon Bedrock](https://aws.amazon.com/bedrock/) hay một mô hình hoặc hệ thống RAG khác được lưu trữ ở nơi khác, bao gồm [Amazon Bedrock Knowledge Bases](https://aws.amazon.com/bedrock/knowledge-bases/) hoặc các triển khai đa đám mây và tại chỗ. Chúng tôi [gần đây đã công bố sự có sẵn chung](https://aws.amazon.com/blogs/machine-learning/evaluate-models-or-rag-systems-using-amazon-bedrock-evaluations-now-generally-available/) của kỹ thuật mô hình ngôn ngữ lớn (LLM)-as-a-judge trong đánh giá mô hình và công cụ đánh giá RAG mới, cũng được hỗ trợ bởi LLM-as-a-judge đằng sau hậu trường.

Kỹ thuật LLM-as-a-judge hỗ trợ các đánh giá này cho phép đánh giá tự động, chất lượng đánh giá giống con người ở quy mô lớn, sử dụng FMs để đánh giá chất lượng và các chiều hướng AI có trách nhiệm mà không cần can thiệp thủ công. Với các chỉ số tích hợp như tính chính xác (độ chính xác thực tế), tính đầy đủ (tính toàn diện của phản hồi), tính trung thực (phát hiện ảo giác), và các chỉ số AI có trách nhiệm như tính có hại và từ chối trả lời, bạn và đội ngũ của bạn có thể đánh giá các mô hình được lưu trữ trên Amazon Bedrock và cơ sở tri thức một cách tự nhiên, hoặc sử dụng phản hồi BYOI từ các hệ thống tùy chỉnh của bạn.

Giờ đây với Amazon Bedrock, bạn có thể phát triển các chỉ số đánh giá tùy chỉnh cho cả đánh giá mô hình và RAG. Khả năng này mở rộng khung LLM-as-a-judge điều khiển Amazon Bedrock Evaluations.

Trong bài đăng này, chúng tôi trình bày cách sử dụng chỉ số tùy chỉnh trong Amazon Bedrock Evaluations để đo lường và cải thiện hiệu suất của các ứng dụng AI tạo sinh theo các yêu cầu kinh doanh cụ thể và tiêu chí đánh giá của bạn.

## Tổng quan

Chỉ số tùy chỉnh trong Amazon Bedrock Evaluations cung cấp các tính năng sau:

* Trải nghiệm bắt đầu đơn giản – Các mẫu khởi động được xây dựng sẵn có sẵn trên [AWS Management Console](http://aws.amazon.com/console) dựa trên các chỉ số tích hợp đã được kiểm tra trong ngành của chúng tôi, với các tùy chọn tạo từ đầu cho các tiêu chí đánh giá cụ thể.
* Hệ thống chấm điểm linh hoạt – Hỗ trợ có sẵn cho cả chấm điểm định lượng (số) và định tính (phân loại) để tạo chỉ số thứ tự, chỉ số danh nghĩa, hoặc thậm chí sử dụng các công cụ đánh giá cho các nhiệm vụ phân loại.
* Quản lý quy trình làm việc hợp lý – Bạn có thể lưu các chỉ số tùy chỉnh để tái sử dụng qua nhiều công việc đánh giá hoặc nhập các chỉ số đã được định nghĩa trước đó từ các tệp JSON.
* Tích hợp nội dung động – Với các biến mẫu tích hợp (ví dụ, {{prompt}}, {{prediction}}, và {{context}}), bạn có thể tiêm nội dung tập dữ liệu và đầu ra mô hình vào các prompt đánh giá một cách mượt mà.
* Kiểm soát đầu ra có thể tùy chỉnh – Bạn có thể sử dụng lược đồ đầu ra được khuyến nghị của chúng tôi để có kết quả nhất quán, với các tùy chọn nâng cao để định nghĩa các định dạng đầu ra tùy chỉnh cho các trường hợp sử dụng chuyên biệt.

Chỉ số tùy chỉnh cho bạn quyền kiểm soát chưa từng có về cách bạn đo lường hiệu suất hệ thống AI, vì vậy bạn có thể căn chỉnh các đánh giá với các yêu cầu kinh doanh cụ thể và các trường hợp sử dụng của bạn. Dù đánh giá tính thực tế, tính mạch lạc, tính hữu ích, hay các tiêu chí cụ thể theo lĩnh vực, chỉ số tùy chỉnh trong Amazon Bedrock cho phép những hiểu biết đánh giá có ý nghĩa và có thể hành động hơn.

## Các định dạng dữ liệu được hỗ trợ

### Tải lên prompt của thẩm phán

Để tải lên các chỉ số tùy chỉnh đã lưu trước đó vào một công việc đánh giá, hãy tuân theo định dạng JSON trong các ví dụ sau.

Mã sau minh họa một định nghĩa với thang số:

{

"customMetricDefinition": {

"metricName": "my\_custom\_metric",

"instructions": "Prompt chỉ số tùy chỉnh hoàn chỉnh của bạn bao gồm ít nhất một {{biến đầu vào}}",

"ratingScale": [

{

"definition": "định nghĩa đánh giá đầu tiên",

"value": {

"floatValue": 3

}

},

{

"definition": "định nghĩa đánh giá thứ hai",

"value": {

"floatValue": 2

}

},

{

"definition": "định nghĩa đánh giá thứ ba",

"value": {

"floatValue": 1

}

}

]

}

}

## Định dạng tập dữ liệu đánh giá mô hình

Khi sử dụng LLM-as-a-judge, chỉ một mô hình có thể được đánh giá mỗi công việc đánh giá. Do đó, bạn phải cung cấp một mục duy nhất trong danh sách modelResponses cho mỗi đánh giá, mặc dù bạn có thể chạy nhiều công việc đánh giá để so sánh các mô hình khác nhau. Sau đây là định dạng JSONL đầu vào cho LLM-as-a-judge trong đánh giá mô hình:

{

"prompt": string

"referenceResponse"?: string

"category"?: string

"modelResponses"?: [

{

"response": string

"modelIdentifier": string

}

]

}

## Định dạng tập dữ liệu đánh giá RAG

Chúng tôi đã cập nhật định dạng tập dữ liệu đầu vào công việc đánh giá để trở nên linh hoạt hơn cho đánh giá RAG. Giờ đây, bạn có thể mang referenceContexts, là các đoạn văn được truy xuất mong đợi, vì vậy bạn có thể so sánh các ngữ cảnh được truy xuất thực tế với các ngữ cảnh được truy xuất mong đợi của bạn.

## Biến để tiêm dữ liệu vào prompt thẩm phán

Để đảm bảo rằng dữ liệu của bạn được tiêm vào prompt thẩm phán ở đúng vị trí, hãy sử dụng các biến từ bảng sau. Bảng sau tóm tắt các biến cho đánh giá mô hình:

|  |  |  |  |
| --- | --- | --- | --- |
| Tên Thông Thường | Biến | Khóa JSONL Tập Dữ Liệu Đầu Vào | Bắt Buộc hoặc Tùy Chọn |
| Prompt | {{prompt}} | prompt | Tùy chọn |
| Phản hồi | {{prediction}} | Cho công việc BYOI: modelResponses.response | Bắt buộc |
| Phản hồi sự thật cơ bản | {{ground\_truth}} | referenceResponse | Tùy chọn |

## Điều kiện tiên quyết

Để sử dụng các tính năng đánh giá mô hình LLM-as-a-judge và đánh giá RAG với BYOI, bạn phải có các điều kiện tiên quyết sau:

* Tài khoản AWS và quyền truy cập mô hình:
  + Một [tài khoản AWS](https://signin.aws.amazon.com/signin?redirect_uri=https%3A%2F%2Fportal.aws.amazon.com%2Fbilling%2Fsignup%2Fresume&client_id=signup) đang hoạt động
  + Các mô hình đánh giá và tạo sinh được chọn được kích hoạt trong Amazon Bedrock
  + Đã xác nhận [AWS Regions](https://docs.aws.amazon.com/glossary/latest/reference/glos-chap.html#region) nơi các mô hình [có sẵn và hạn ngạch của chúng](https://docs.aws.amazon.com/bedrock/latest/userguide/models-regions.html)
* Cấu hình [AWS Identity and Access Management](https://aws.amazon.com/iam/) (IAM) và [Amazon Simple Storage Service](http://aws.amazon.com/s3) (Amazon S3):
  + Đã hoàn thành thiết lập IAM và [quyền](https://docs.aws.amazon.com/bedrock/latest/userguide/judge-service-roles.html) cho cả đánh giá mô hình và RAG
  + Đã cấu hình [bucket S3](https://docs.aws.amazon.com/AmazonS3/latest/userguide/access-policy-language-overview.html) với quyền thích hợp
  + Đã bật [CORS](https://docs.aws.amazon.com/bedrock/latest/userguide/model-evaluation-security-cors.html) trên bucket S3 của bạn

## Tạo công việc đánh giá mô hình với chỉ số tùy chỉnh sử dụng Amazon Bedrock Evaluations

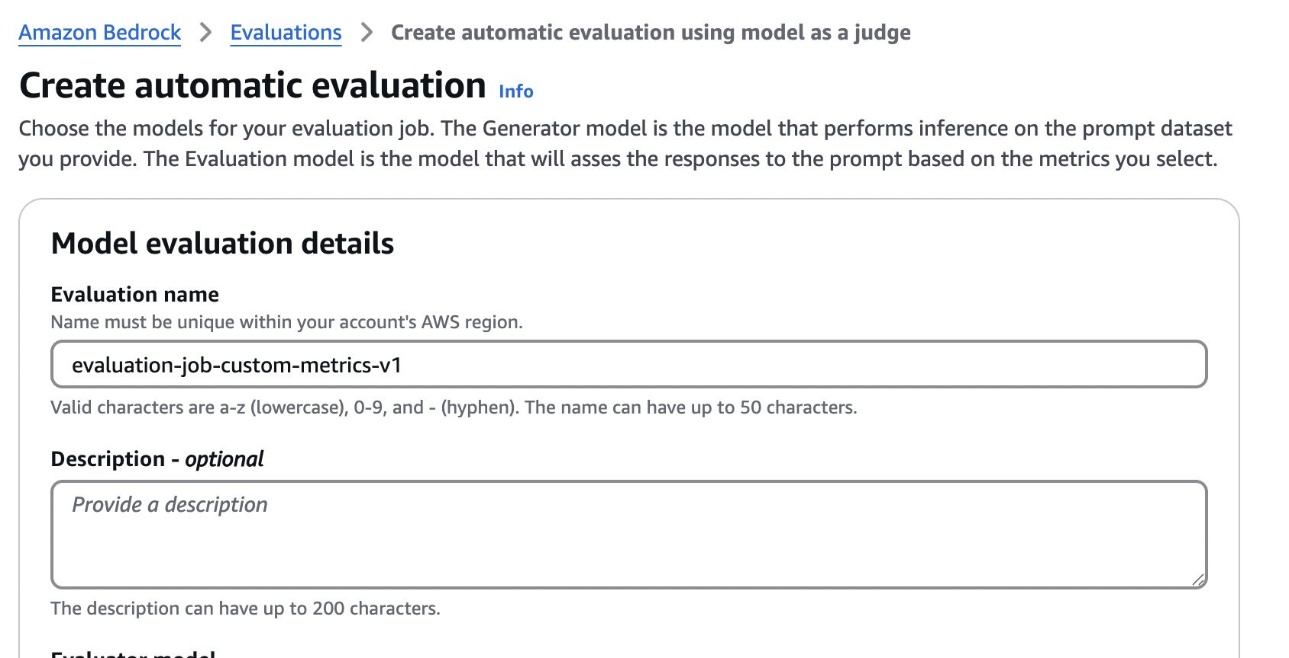
Hoàn thành các bước sau để tạo một công việc với đánh giá mô hình và chỉ số tùy chỉnh sử dụng Amazon Bedrock Evaluations:

1. Trên bảng điều khiển Amazon Bedrock, chọn **Evaluations** trong ngăn điều hướng và chọn **Models**
2. Trong phần **đánh giá mô hình**, trên menu thả xuống Create, **chọn Automatic: model as a judge.**

A screenshot of a computer

AI-generated content may be incorrect.

1. Cho chi tiết đánh giá mô hình, nhập tên đánh giá và mô tả tùy chọn.



1. **Cho mô hình đánh giá**, chọn mô hình bạn muốn sử dụng cho đánh giá tự động.
2. Cho **nguồn suy luận**, chọn nguồn và chọn mô hình bạn muốn đánh giá.
3. Trong phần Chỉ số tùy chỉnh, chúng ta tạo một chỉ số mới gọi là "Comprehensiveness."

Sau đây là chỉ số chúng tôi đã định nghĩa đầy đủ:

Vai trò của bạn là đánh giá tính toàn diện của một câu trả lời dựa trên câu hỏi và

dự đoán. Đánh giá chất lượng, độ chính xác và tính hữu ích của phản hồi mô hình ngôn ngữ,

và sử dụng những điều này để đánh giá mức độ toàn diện của phản hồi. Trao điểm cao hơn cho các phản hồi

chi tiết và sâu sắc.

Đánh giá cẩn thận tính toàn diện của phản hồi LLM cho truy vấn đã cho (prompt)

dựa trên tất cả các tiêu chí được chỉ định. Gán một điểm tổng thể duy nhất đại diện tốt nhất cho

tính toàn diện, và cung cấp một giải thích ngắn gọn biện minh cho đánh giá của bạn, tham chiếu

các điểm mạnh và điểm yếu cụ thể được quan sát.

Khi đánh giá chất lượng phản hồi, hãy xem xét các tiêu chí sau:

- Độ chính xác: Tính chính xác thực tế của thông tin được cung cấp

- Tính đầy đủ: Bao phủ các khía cạnh quan trọng của truy vấn

- Tính rõ ràng: Tổ chức và trình bày thông tin rõ ràng

- Tính hữu ích: Tiện ích thực tế của phản hồi cho người dùng

Đánh giá những điều sau:

Truy vấn:

{{prompt}}

Phản hồi để đánh giá:

{{prediction}}

Tạo công việc đánh giá mô hình với chỉ số tùy chỉnh sử dụng Python SDK và APIs

Để sử dụng Python SDK để tạo một công việc đánh giá mô hình với chỉ số tùy chỉnh, hãy làm theo các bước này (hoặc tham khảo [notebook ví dụ](https://github.com/aws-samples/amazon-bedrock-samples/tree/main/evaluation-observe/bedrock-eval-custom-metrics) của chúng tôi):

Thiết lập các cấu hình cần thiết:

import boto3

import time

from datetime import datetime

# Cấu hình cơ sở tri thức và cài đặt mô hình

evaluator\_model = "anthropic.claude-3-5-sonnet-20240620-v1:0"

generator\_model = "amazon.nova-lite-v1:0"

custom\_metrics\_evaluator\_model = "anthropic.claude-3-5-sonnet-20240620-v1:0"

role\_arn = "arn:aws:iam:::role/"

BUCKET\_NAME = ""

# Chỉ định vị trí S3

input\_data = f"s3://{BUCKET\_NAME}/evaluation\_data/input.jsonl"

output\_path = f"s3://{BUCKET\_NAME}/evaluation\_output/"

# Tạo client Bedrock

bedrock\_client = boto3.client('bedrock', region\_name='us-east-1')

Tạo đánh giá hệ thống RAG với chỉ số tùy chỉnh sử dụng Amazon Bedrock Evaluations

Trong ví dụ này, chúng tôi hướng dẫn qua một đánh giá hệ thống RAG với sự kết hợp của các chỉ số tích hợp và chỉ số đánh giá tùy chỉnh trên bảng điều khiển Amazon Bedrock.

1. Trên bảng điều khiển Amazon Bedrock, chọn Evaluations trong ngăn điều hướng.
2. Trên tab RAG, chọn Create.
3. Cho chi tiết đánh giá RAG, nhập tên đánh giá và mô tả tùy chọn.
4. Cho mô hình đánh giá, chọn mô hình bạn muốn sử dụng cho đánh giá tự động.
5. Cho nguồn suy luận, chọn nguồn.
6. Trong phần Chỉ số tùy chỉnh, chọn mô hình đánh giá của bạn.
7. Tạo chỉ số mới của bạn cho đánh giá RAG gọi là information\_comprehensiveness.

## Cân nhắc và thực hành tốt nhất

Khi sử dụng lược đồ đầu ra của các chỉ số tùy chỉnh, lưu ý những điều sau:

* Nếu bạn sử dụng lược đồ đầu ra tích hợp (được khuyến nghị), đừng thêm thang điểm của bạn vào prompt thẩm phán chính.
* Dịch vụ đánh giá sẽ tự động nối các hướng dẫn prompt thẩm phán của bạn với thang đánh giá lược đồ đầu ra đã định nghĩa của bạn.
* Các prompt thẩm phán được nối đầy đủ hiển thị trong cửa sổ Preview nếu bạn đang sử dụng bảng điều khiển Amazon Bedrock.
* Vì các LLM thẩm phán vốn dĩ ngẫu nhiên, có thể có một số phản hồi chúng tôi không thể phân tích và hiển thị trên bảng điều khiển.

## Dọn dẹp

Để tránh phát sinh phí trong tương lai, hãy xóa bucket S3, các instance notebook và các tài nguyên khác đã được triển khai như một phần của bài đăng.

## Kết luận

Việc bổ sung chỉ số tùy chỉnh vào Amazon Bedrock Evaluations trao quyền cho các tổ chức định nghĩa tiêu chí đánh giá riêng cho các hệ thống AI tạo sinh. Bằng cách mở rộng khung LLM-as-a-judge với chỉ số tùy chỉnh, các doanh nghiệp giờ đây có thể đo lường những gì quan trọng cho các trường hợp sử dụng cụ thể của họ cùng với các chỉ số tích hợp. Với sự hỗ trợ cho cả hệ thống chấm điểm số và phân loại, các chỉ số tùy chỉnh này cho phép đánh giá nhất quán phù hợp với các tiêu chuẩn và mục tiêu tổ chức.

Khi AI tạo sinh ngày càng được tích hợp vào các quy trình kinh doanh, khả năng đánh giá đầu ra dựa trên các tiêu chí được định nghĩa tùy chỉnh là thiết yếu để duy trì chất lượng và thúc đẩy cải tiến liên tục. Chúng tôi khuyến khích bạn khám phá các khả năng mới này thông qua bảng điều khiển Amazon Bedrock và các ví dụ API được cung cấp, và khám phá cách [khung đánh giá cá nhân hóa](https://aws.amazon.com/bedrock/evaluations/) có thể nâng cao hiệu suất và tác động kinh doanh của hệ thống AI của bạn.

**Về các tác giả**

## Shreyas Subramanian

Shreyas Subramanian là Nhà khoa học dữ liệu chính và giúp khách hàng bằng cách sử dụng AI tạo sinh và deep learning để giải quyết các thách thức kinh doanh của họ bằng cách sử dụng các dịch vụ AWS. Shreyas có nền tảng trong tối ưu hóa quy mô lớn và ML và trong việc sử dụng ML và reinforcement learning để tăng tốc các nhiệm vụ tối ưu hóa.

## Adewale Akinfaderin

Adewale Akinfaderin là Nhà khoa học dữ liệu cấp cao–AI tạo sinh, Amazon Bedrock, nơi anh đóng góp vào các đổi mới tiên tiến trong các mô hình nền tảng và ứng dụng AI tạo sinh tại AWS. Chuyên môn của anh là trong các phương pháp AI/ML có thể tái tạo và end-to-end, triển khai thực tế, và giúp khách hàng toàn cầu xây dựng và phát triển các giải pháp có thể mở rộng cho các vấn đề liên ngành.

## Jesse Manders

Jesse Manders là Quản lý sản phẩm cấp cao trên Amazon Bedrock, dịch vụ phát triển AI tạo sinh AWS. Anh làm việc tại giao điểm của AI và tương tác con người với mục tiêu tạo ra và cải thiện các sản phẩm và dịch vụ AI tạo sinh để đáp ứng nhu cầu của chúng ta.

## Ishan Singh

Ishan Singh là Nhà khoa học dữ liệu AI tạo sinh cấp cao tại Amazon Web Services, nơi anh giúp khách hàng xây dựng các giải pháp và sản phẩm AI tạo sinh sáng tạo và có trách nhiệm. Với nền tảng vững chắc trong AI/ML, Ishan chuyên về xây dựng các giải pháp AI tạo sinh tạo ra giá trị kinh doanh.