# Best Practices để tạo và đánh giá Ground Truth cho Generative AI Question-Answering với FMEval by Samantha Stuart, Ivan Cui, Philippe Duplessis-Guindon, and Rahul Jani on 05 MAR 2025 in [Amazon SageMaker](https://aws.amazon.com/blogs/machine-learning/category/artificial-intelligence/sagemaker/), [Best Practices](https://aws.amazon.com/blogs/machine-learning/category/post-types/best-practices/), [Expert (400)](https://aws.amazon.com/blogs/machine-learning/category/learning-levels/expert-400/), [Generative AI](https://aws.amazon.com/blogs/machine-learning/category/artificial-intelligence/generative-ai/), [Learning Levels](https://aws.amazon.com/blogs/machine-learning/category/learning-levels/), [Thought Leadership](https://aws.amazon.com/blogs/machine-learning/category/post-types/thought-leadership/)

Các ứng dụng hỏi đáp sử dụng Generative AI đang mở rộng ranh giới năng suất doanh nghiệp. Những trợ lý này có thể được vận hành bởi nhiều kiến trúc backend khác nhau bao gồm Retrieval Augmented Generation (RAG), quy trình agentic, các large language model (LLM) đã được fine-tune, hoặc kết hợp các kỹ thuật này. Tuy nhiên, việc xây dựng và triển khai các trợ lý AI đáng tin cậy đòi hỏi một ground truth và framework đánh giá mạnh mẽ.

Ground truth trong AI đề cập đến dữ liệu được biết là chính xác, đại diện cho kết quả use case mong đợi của hệ thống đang được mô hình hóa. Bằng cách cung cấp kết quả mong đợi để đối chiếu, ground truth cho phép đánh giá chất lượng hệ thống một cách xác định. Việc chạy đánh giá xác định các trợ lý generative AI dựa trên dữ liệu ground truth của use case cho phép tạo ra các benchmark tùy chỉnh. Những benchmark này rất cần thiết để theo dõi sự thay đổi hiệu suất theo thời gian và so sánh thống kê nhiều trợ lý trong việc hoàn thành cùng một nhiệm vụ. Ngoài ra, chúng cho phép định lượng các thay đổi hiệu suất như một hàm của các cải tiến đối với trợ lý cơ bản, tất cả trong môi trường được kiểm soát. Với các quy trình đánh giá xác định như các chỉ số Factual Knowledge và QA Accuracy của [FMEval](https://github.com/aws/fmeval/tree/main), việc tạo ground truth và triển khai các chỉ số đánh giá được kết nối chặt chẽ với nhau. Để đảm bảo đo lường chất lượng cao nhất cho ứng dụng hỏi đáp của bạn so với ground truth, việc triển khai chỉ số đánh giá phải định hướng việc tổng hợp ground truth.

Trong bài viết này, chúng tôi thảo luận về các phương pháp tốt nhất để áp dụng LLM để tạo ground truth nhằm đánh giá các trợ lý hỏi đáp với FMEval ở quy mô doanh nghiệp. [FMEval](https://github.com/aws/fmeval/tree/main) là một bộ đánh giá toàn diện từ [Amazon SageMaker Clarify](https://aws.amazon.com/sagemaker/clarify/), cung cấp các triển khai tiêu chuẩn hóa của các chỉ số để đánh giá chất lượng và trách nhiệm. Để tìm hiểu thêm về FMEval, hãy xem "[Evaluate large language models for quality and responsibility of LLMs](https://aws.amazon.com/blogs/machine-learning/evaluate-large-language-models-for-quality-and-responsibility/)". Ngoài ra, hãy xem [Generative AI Security Scoping Matrix](https://aws.amazon.com/blogs/security/securing-generative-ai-an-introduction-to-the-generative-ai-security-scoping-matrix/) để được hướng dẫn về việc kiểm duyệt thông tin bảo mật và thông tin nhận dạng cá nhân (PII) như một phần của giải pháp generative AI của bạn.

Bằng cách tuân theo những hướng dẫn này, các nhóm dữ liệu có thể triển khai việc tạo ground truth độ tin cậy cao để đánh giá use case hỏi đáp với FMEval. Để biết các phương pháp tốt nhất về tổng hợp ground truth cho đánh giá hỏi đáp với FMEval mà bạn có thể sử dụng để thiết kế các mẫu prompt ground truth FMEval, hãy xem "[Ground truth curation and metric interpretation best practices for evaluating generative AI question answering using FMEval.](https://aws.amazon.com/blogs/machine-learning/ground-truth-curation-and-metric-interpretation-best-practices-for-evaluating-generative-ai-question-answering-using-fmeval/)".

## Tạo dữ liệu ground truth cho đánh giá hỏi đáp FMEval

Một cách để bắt đầu tạo ground truth là chuẩn bị thủ công một tập dữ liệu câu hỏi-câu trả lời nhỏ. Tập dữ liệu được chuẩn bị thủ công này nên nhỏ gọn (dựa trên băng thông), có tín hiệu chất lượng cao và tốt nhất là được chuẩn bị bởi các chuyên gia trong lĩnh vực (SME). Việc tạo tập dữ liệu này buộc phải thực hiện căn chỉnh dữ liệu ngay từ đầu quá trình đánh giá, đặt ra những câu hỏi và cuộc thảo luận quan trọng giữa các bên liên quan về việc những câu hỏi nào cần được đo lường theo thời gian cho doanh nghiệp. Kết quả của việc này có ba phần:

* Sự đồng thuận của các bên liên quan về N câu hỏi quan trọng nhất
* Nhận thức của các bên liên quan về quy trình đánh giá
* Một tập dữ liệu ground truth chất lượng cao ban đầu cho việc đánh giá proof of concept đầu tiên

Mặc dù việc chuẩn bị ground truth bởi SME là một khởi đầu tốt, ở quy mô cơ sở kiến thức doanh nghiệp, việc tạo ground truth hoàn toàn bởi SME sẽ tốn quá nhiều thời gian và tài nguyên. Để mở rộng quy mô tạo và chuẩn bị ground truth, bạn có thể áp dụng phương pháp tiếp cận dựa trên rủi ro kết hợp với chiến lược dựa trên prompt sử dụng LLM. Cần lưu ý rằng ground truth được tạo bởi LLM không thay thế được sự tham gia của SME. Ví dụ, nếu ground truth được tạo bởi LLM trước khi có sự tham gia của SME, vẫn cần SME để xác định những câu hỏi nào là cơ bản đối với doanh nghiệp và sau đó căn chỉnh ground truth với giá trị kinh doanh như một phần của quy trình human-in-the-loop.

Để minh họa, chúng tôi cung cấp hướng dẫn từng bước sử dụng thư gửi cổ đông năm 2023 của Amazon làm dữ liệu nguồn.

Theo các “[ground truth curation best practices](https://aws.amazon.com/blogs/machine-learning/ground-truth-curation-and-metric-interpretation-best-practices-for-evaluating-generative-ai-question-answering-using-fmeval/)” cho đánh giá hỏi đáp FMEval, ground truth được chuẩn bị dưới dạng bộ ba câu hỏi-câu trả lời-sự kiện. Câu hỏi và câu trả lời được chuẩn bị để phù hợp với phản hồi lý tưởng của trợ lý hỏi đáp về nội dung, độ dài và phong cách. Sự kiện là biểu diễn tối thiểu của câu trả lời ground truth, bao gồm một hoặc nhiều thực thể chủ thể của câu hỏi.

Ví dụ, hãy xem xét cách chuyển đổi đoạn tài liệu nguồn sau đây từ thư gửi cổ đông năm 2023 của Amazon thành ground truth hỏi đáp.

Kính gửi các Cổ đông:

Năm ngoái vào thời điểm này, tôi đã chia sẻ sự nhiệt tình và lạc quan của mình về tương lai của Amazon. Hôm nay, tôi còn lạc quan hơn nữa. Có nhiều lý do, nhưng bắt đầu từ tiến bộ chúng tôi đã đạt được trong kết quả tài chính và trải nghiệm khách hàng, và mở rộng đến việc tiếp tục đổi mới và những cơ hội đáng chú ý phía trước. Năm 2023, tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước ("YoY") từ 514 tỷ USD lên 575 tỷ USD. Theo phân khúc, doanh thu Bắc Mỹ tăng 12% YoY từ 316 tỷ USD lên 353 tỷ USD, doanh thu Quốc tế tăng 11% YoY từ 118 tỷ USD lên 131 tỷ USD, và doanh thu AWS tăng 13% YoY từ 80 tỷ USD lên 91 tỷ USD. Hơn nữa, thu nhập từ hoạt động và Dòng tiền tự do ("FCF") của Amazon đã cải thiện đáng kể. Thu nhập từ hoạt động năm 2023 tăng 201% YoY từ 12,2 tỷ USD (biên lợi nhuận hoạt động 2,4%) lên 36,9 tỷ USD (biên lợi nhuận hoạt động 6,4%).

Để chuyển đổi đoạn tài liệu nguồn thành ground truth, chúng tôi cung cấp một mẫu prompt LLM cơ bản. Trong mẫu, chúng tôi hướng dẫn LLM áp dụng cách tiếp cận dựa trên sự kiện để diễn giải đoạn văn bản bằng cách sử dụng logic chuỗi suy nghĩ. Trong ví dụ của chúng tôi, chúng tôi làm việc với Claude LLM của Anthropic trên [Amazon Bedrock](https://aws.amazon.com/bedrock/). Mẫu này tương thích và có thể được điều chỉnh cho các LLM khác, chẳng hạn như các LLM được lưu trữ trên [Amazon Sagemaker Jumpstart](https://aws.amazon.com/sagemaker-ai/jumpstart/) và tự lưu trữ trên cơ sở hạ tầng AWS. Để điều chỉnh prompt cho các LLM khác, có thể cần một cách tiếp cận khác để đánh dấu các phần prompt thay vì thẻ XML. Ví dụ, các mô hình Meta Llama sử dụng các thẻ như <s> [INST] và <<SYS>>. Để biết thêm thông tin, [documentation on LLM prompt design](https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-meta.html) về thiết kế prompt LLM và [tài liệu FMEval.](https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-foundation-model-evaluate-whatis.html)

LLM được gán một persona để thiết lập quan điểm của nó khi thực hiện nhiệm vụ. Trong hướng dẫn, LLM xác định các sự kiện như các thực thể từ đoạn tài liệu nguồn. Đối với mỗi sự kiện, một bộ ba câu hỏi-câu trả lời-sự kiện được tạo dựa trên sự kiện được phát hiện và ngữ cảnh xung quanh. Trong prompt, chúng tôi cung cấp các ví dụ chi tiết để kiểm soát nội dung của các câu hỏi ground truth. Các ví dụ tập trung vào các câu hỏi về kiến thức kinh doanh theo từng đoạn trong khi bỏ qua metadata không liên quan có thể có trong đoạn. Bạn có thể tùy chỉnh các ví dụ prompt để phù hợp với trường hợp sử dụng ground truth của bạn.

Chúng tôi tiếp tục hướng dẫn [LLM áp dụng các best practices](https://aws.amazon.com/blogs/machine-learning/ground-truth-curation-and-metric-interpretation-best-practices-for-evaluating-generative-ai-question-answering-using-fmeval/) về chuẩn bị ground truth cho FMEval, chẳng hạn như tạo nhiều biến thể của các sự kiện để phù hợp với nhiều cách biểu thị đơn vị có thể có. Các yếu tố chuẩn bị bổ sung tùy thuộc vào nhiệm vụ cần thực hiện - như ngôn ngữ thương hiệu và giọng điệu - có thể được đưa vào prompt tạo ground truth. Với mẫu sau đây, chúng tôi đã xác minh rằng Claude Sonnet 3.5 của Anthropic có thể tạo các thuộc tính ground truth tùy chỉnh phù hợp với các tính năng FMEval, chẳng hạn như dấu phân cách <OR> để biểu thị các câu trả lời thay thế có thể chấp nhận được cho một sự kiện ground truth.

```Bạn là một chuyên gia về việc tạo dữ liệu ground truth để đánh giá ứng dụng AI tạo sinh trên AWS.

Hãy làm theo hướng dẫn trong thẻ XML <instructions> để tạo ra các bộ ba câu hỏi-câu trả lời-sự thật từ một đoạn trích tài liệu nguồn.

<instructions>

- Hãy giải quyết từng bước một để đảm bảo chúng ta có câu trả lời đúng.

- Xem xét đoạn trích tài liệu nguồn được cung cấp trong thẻ XML <document> bên dưới

- Với mỗi sự thật có ý nghĩa trong miền trong <document>, trích xuất một bộ câu hỏi-câu trả lời-sự thật rõ ràng ở định dạng JSON bao gồm một cặp câu hỏi và câu trả lời thể hiện sự thật dưới dạng một câu ngắn, theo sau là một sự thật được trích xuất ngắn gọn từ câu trả lời.

<domain\_knowledge\_focus>

- CHỈ tập trung vào kiến thức miền cụ thể có trong nội dung tài liệu

- Bỏ qua tất cả các phần tử metadata và cấu trúc bao gồm nhưng không giới hạn:

- Ngày tháng, phiên bản, số trang của tài liệu

- Số thứ tự hoặc tiêu đề phần

- Cấu trúc bảng hoặc vị trí hàng/cột

- Vị trí hoặc thứ tự danh sách

- Câu hỏi phải tham chiếu đến các thực thể miền cụ thể thay vì các phần tử tài liệu chung

</domain\_knowledge\_focus>

<context\_specification\_requirements>

Xác định Nguồn Tài liệu

- Luôn tham chiếu đến tài liệu nguồn cụ thể và ngày/phiên bản của nó

- Ví dụ: "Theo [Tên Tài liệu + Ngày], [câu hỏi cụ thể] là gì?"

Ngăn chặn Tham chiếu Chéo

- Mỗi câu hỏi phải có thể trả lời được từ phần tài liệu hiện tại

- Không tạo câu hỏi yêu cầu thông tin từ nhiều tài liệu

- Ví dụ: "Trong [Tên Tài liệu] này, [yêu cầu cụ thể] là gì?"

Chỉ định Phòng ban/LOB

- Luôn chỉ định phòng ban, lĩnh vực kinh doanh hoặc đơn vị tổ chức liên quan

- Ví dụ: "Yêu cầu của [Tên Phòng ban] cho [quy trình cụ thể] là gì?"

Nhắm mục tiêu Phần Tài liệu

- Tham chiếu các phần cụ thể khi vị trí thông tin có liên quan

- Ví dụ: "Trong Phần [X] của [Tên Tài liệu], các bước cho [quy trình cụ thể] là gì?"

Ngữ cảnh Dựa trên Vai trò

- Chỉ định vai trò, trách nhiệm hoặc cấp độ thẩm quyền liên quan

- Ví dụ: "[vai trò cụ thể] nào được ủy quyền để [hành động cụ thể]?"

Phần tử Kiểm soát Phiên bản

- Bao gồm thông tin phiên bản hoặc sửa đổi liên quan

- Ví dụ: "Những thay đổi nào được thực hiện trong bản sửa đổi [Tháng Năm] của [Tài liệu]?"

Số Chính sách/Thủ tục

- Bao gồm số tham chiếu chính sách hoặc thủ tục cụ thể

- Ví dụ: "Theo Chính sách [Số], yêu cầu cho [hành động cụ thể] là gì?"

Tham chiếu Khung Quy định

- Chỉ định khung quy định hoặc yêu cầu tuân thủ liên quan

- Ví dụ: "Yêu cầu tuân thủ [Quy định] nào được chỉ định cho [quy trình cụ thể]?"

Chỉ định Hệ thống/Nền tảng

- Đặt tên các hệ thống, nền tảng hoặc công cụ cụ thể

- Ví dụ: "Các bước nào được yêu cầu trong [Tên Hệ thống] để [hành động cụ thể]?"

Phân loại Loại Tài liệu

- Chỉ định loại tài liệu (SOP, Chính sách, Hướng dẫn, v.v.)

- Ví dụ: "Trong [Loại Tài liệu + Số], [thông tin cụ thể] được lưu trữ ở đâu?"

Hiệu lực Thời gian

- Bao gồm ngày có hiệu lực hoặc khoảng thời gian

- Ví dụ: "Quy trình nào có hiệu lực từ [Ngày] theo [Tài liệu]?"

Thẩm quyền Địa lý

- Chỉ định khu vực địa lý hoặc thẩm quyền liên quan

- Ví dụ: "Yêu cầu nào áp dụng cho [Khu vực] theo [Tài liệu]?"

Chủ sở hữu Quy trình Kinh doanh

- Xác định chủ sở hữu quy trình hoặc bên chịu trách nhiệm liên quan

- Ví dụ: "Theo [Tài liệu], ai sở hữu quy trình cho [hành động cụ thể]?"

Cấp độ Phân loại

- Bao gồm phân loại bảo mật hoặc bảo mật liên quan

- Ví dụ: "Yêu cầu cho dữ liệu [Cấp độ Phân loại] là gì?"

Phạm vi Bên liên quan

- Chỉ định các bên liên quan hoặc thẩm quyền phê duyệt liên quan

- Ví dụ: "[cấp độ bên liên quan] nào phải phê duyệt [hành động cụ thể]?"

</context\_specification\_requirements>

<question\_quality\_criteria>

- Câu hỏi phải đủ cụ thể để cơ sở dữ liệu vector có thể khớp chúng với phần tài liệu liên quan

- Câu hỏi nên bao gồm các thuật ngữ, tên và ngữ cảnh quan trọng

- Câu hỏi nên nhắm đến thông tin cụ thể, có thể thực hiện được

- Câu trả lời nên cung cấp ngữ cảnh đầy đủ mà không cần tham chiếu lại các phần tử tài liệu

</question\_quality\_criteria>

<output\_format>

Bộ câu hỏi-câu trả lời-sự thật mỗi cái nên là một chuỗi ngắn ở định dạng JSON với các khóa: "question", "ground\_truth\_answer", "fact"

</output\_format>

<best\_practices>

- Câu hỏi, câu trả lời và sự thật không nên đề cập đến thực thể chủ thể là "nó" hoặc "họ", và thay vào đó đề cập trực tiếp bằng tên

- Câu hỏi, câu trả lời và sự thật nên là duy nhất cho phần tài liệu, sao cho dựa trên câu hỏi, một lệnh gọi mới đến bộ truy xuất sẽ xử lý đúng phần tài liệu khi đặt câu hỏi ground truth

- Sự thật nên được thể hiện trong 3 từ hoặc ít hơn mô tả một thực thể trong <document>

- Nếu có đơn vị trong sự thật, mục "fact" phải cung cấp nhiều phiên bản của sự thật sử dụng <OR> làm dấu phân cách. Xem <unit\_variations> để biết ví dụ.

<unit\_variations>

- Tương đương Đơn vị Đô la: `1,234 triệu<OR>1.234 tỷ`

- Tương đương Định dạng Ngày: `2024-01-01<OR>1 tháng 1 năm 2024`

- Tương đương Số: `1<OR>một`

</unit\_variations>

</best\_practices>

- Bắt đầu phản hồi của bạn ngay lập tức với bộ JSON câu hỏi-câu trả lời-sự thật, và phân tách mỗi bản ghi JSON được trích xuất bằng một dòng mới.

</instructions>

<document>

{context\_document}

</document>

Bây giờ, hãy trích xuất các cặp câu hỏi câu trả lời và sự thật từ đoạn trích tài liệu theo hướng dẫn của bạn, bắt đầu ngay lập tức với JSON và không có phần mở đầu.```

Đầu ra của quá trình tạo được cung cấp dưới dạng các bản ghi JSONLines theo từng sự kiện với định dạng sau, trong đó các phần tử trong dấu ngoặc vuông đại diện cho các giá trị từ một dòng trong Bảng 1.

{

"question": "[Câu hỏi]",

"ground\_truth\_answer": "[Câu trả lời thực tế]",

"fact": "[Sự kiện]"

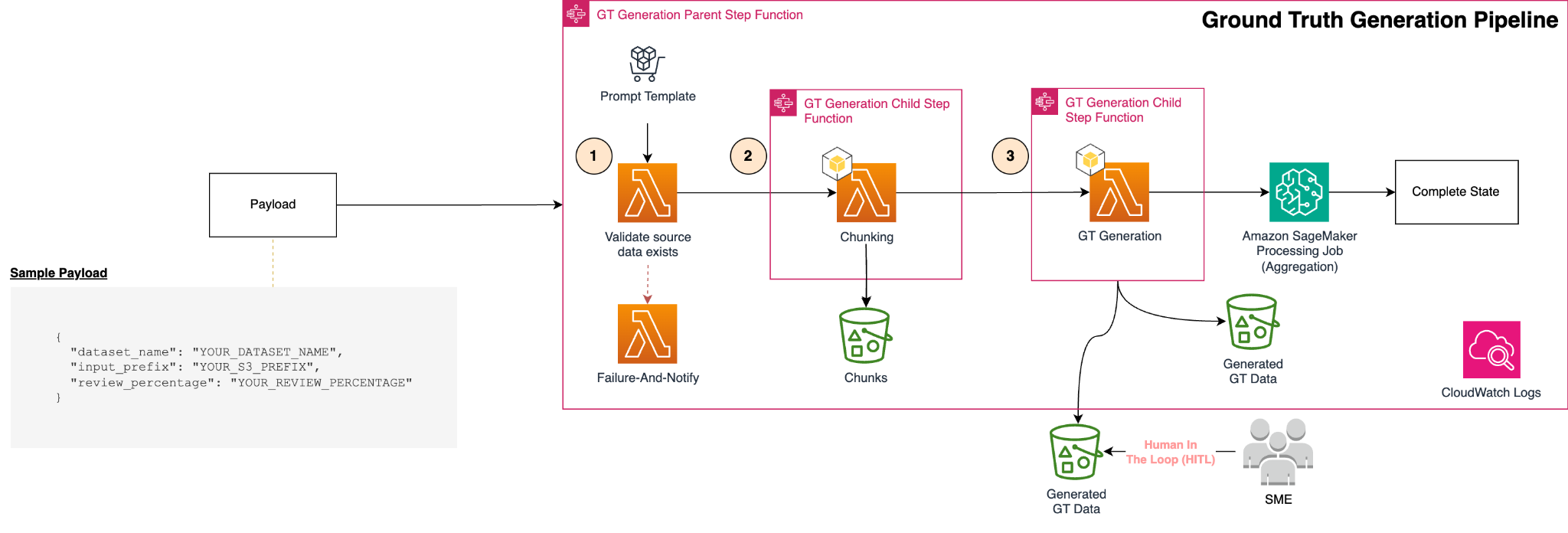
}

Dưới đây là một số ví dụ về ground truth được tạo ra:

| Câu hỏi | Câu trả lời thực tế | Sự kiện |
| --- | --- | --- |
| Tổng doanh thu của Amazon tăng bao nhiêu trong năm 2023? | Tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước từ 514 tỷ USD lên 575 tỷ USD trong năm 2023. | 12%<OR>$514B to $575B |
| Doanh thu khu vực Bắc Mỹ tăng bao nhiêu trong năm 2023? | Doanh thu khu vực Bắc Mỹ tăng 12% so với cùng kỳ năm trước từ 316 tỷ USD lên 353 tỷ USD. | 12%<OR>$316B to $353B |
| Mức tăng trưởng doanh thu quốc tế của Amazon trong năm 2023 là bao nhiêu? | Doanh thu quốc tế tăng 11% so với cùng kỳ năm trước từ 118 tỷ USD lên 131 tỷ USD. | 11%<OR>$118B to $131B |
| Doanh thu AWS tăng bao nhiêu trong năm 2023? | Doanh thu AWS tăng 13% so với cùng kỳ năm trước từ 80 tỷ USD lên 91 tỷ USD. | 13%<OR>$80B to $91B |
| Thu nhập hoạt động của Amazon cải thiện bao nhiêu trong năm 2023? | Thu nhập hoạt động năm 2023 tăng 201% so với cùng kỳ năm trước từ 12,2 tỷ USD lên 36,9 tỷ USD. | 201%<OR>$12.2B to $36.9B |
| Tỷ suất lợi nhuận hoạt động của Amazon trong năm 2023 là bao nhiêu? | Tỷ suất lợi nhuận hoạt động của Amazon trong năm 2023 là 6,4%. | 6.4% |

## Mở rộng quy mô tạo ground truth với pipeline

Để tự động hóa việc tạo ground truth, chúng tôi cung cấp một kiến trúc pipeline xử lý hàng loạt serverless, được minh họa trong hình dưới đây. Ở mức tổng quan, pipeline [AWS Step Functions](https://aws.amazon.com/step-functions/) nhận dữ liệu nguồn từ [Amazon Simple Storage Service (Amazon S3)](https://aws.amazon.com/s3/), và điều phối các [AWS Lambda](https://aws.amazon.com/pm/lambda/?gclid=CjwKCAjwg-24BhB_EiwA1ZOx8oY86uMo0f5u9iXnr3B_u8DalveaiTOP8KK-nrq23SaS66OZZXMPgxoCRegQAvD_BwE&trk=d87368f2-b0ac-4e30-804b-b10e2d25d291&sc_channel=ps&ef_id=CjwKCAjwg-24BhB_EiwA1ZOx8oY86uMo0f5u9iXnr3B_u8DalveaiTOP8KK-nrq23SaS66OZZXMPgxoCRegQAvD_BwE:G:s&s_kwcid=AL!4422!3!651612781100!e!!g!!aws%20lambda!19836398320!150095228874) function để thực hiện việc tiếp nhận, phân đoạn, và gửi prompt đến Amazon Bedrock nhằm tạo ra ground truth dạng JSONLines theo từng fact.

Có ba đầu vào từ người dùng cho step function:

* Tên tùy chỉnh cho tập dữ liệu ground truth
* Tiền tố Amazon S3 đầu vào cho dữ liệu nguồn
* Tỷ lệ phần trăm mẫu cần xem xét

Các cấu hình bổ sung được thiết lập bởi các biến môi trường Lambda, như S3 source bucket và Amazon Bedrock Model ID để gọi khi tạo.

{

"dataset\_name": "TÊN\_TẬP\_DỮ\_LIỆU\_CỦA\_BẠN",

"input\_prefix": "TIỀN\_TỐ\_ĐẦU\_VÀO\_CỦA\_BẠN",

"review\_percentage": "TỶ\_LỆ\_XEM\_XÉT"

}

Sau khi payload ban đầu được truyền vào, một hàm xác thực sẽ tập hợp cấu trúc payload sự kiện toàn cục theo đầu vào hệ thống và đầu vào người dùng.

{

"system\_input": {

"run\_id": "<ID thực thi AWS Step Function>",

"input\_bucket": "<Bucket Amazon S3 dữ liệu đầu vào>",

"output\_bucket": "<Bucket Amazon S3 dữ liệu đầu ra>",

"output\_document\_chunks\_prefix": "<Tiền tố Bucket Amazon S3 để lưu trữ các chunk>",

"chunk\_size": "<Kích thước chunk tài liệu>",

"chunk\_overlap": "<Số token chồng lấp giữa các chunk liên tiếp>"

},

"user\_input": {

"dataset\_name": "<Tên tập dữ liệu>",

"input\_prefix": "<Tiền tố bucket Amazon S3 cho dữ liệu đầu vào tạo ground truth>",

"review\_percentage": "<Phần trăm bản ghi cần đánh dấu để xem xét thủ công>"

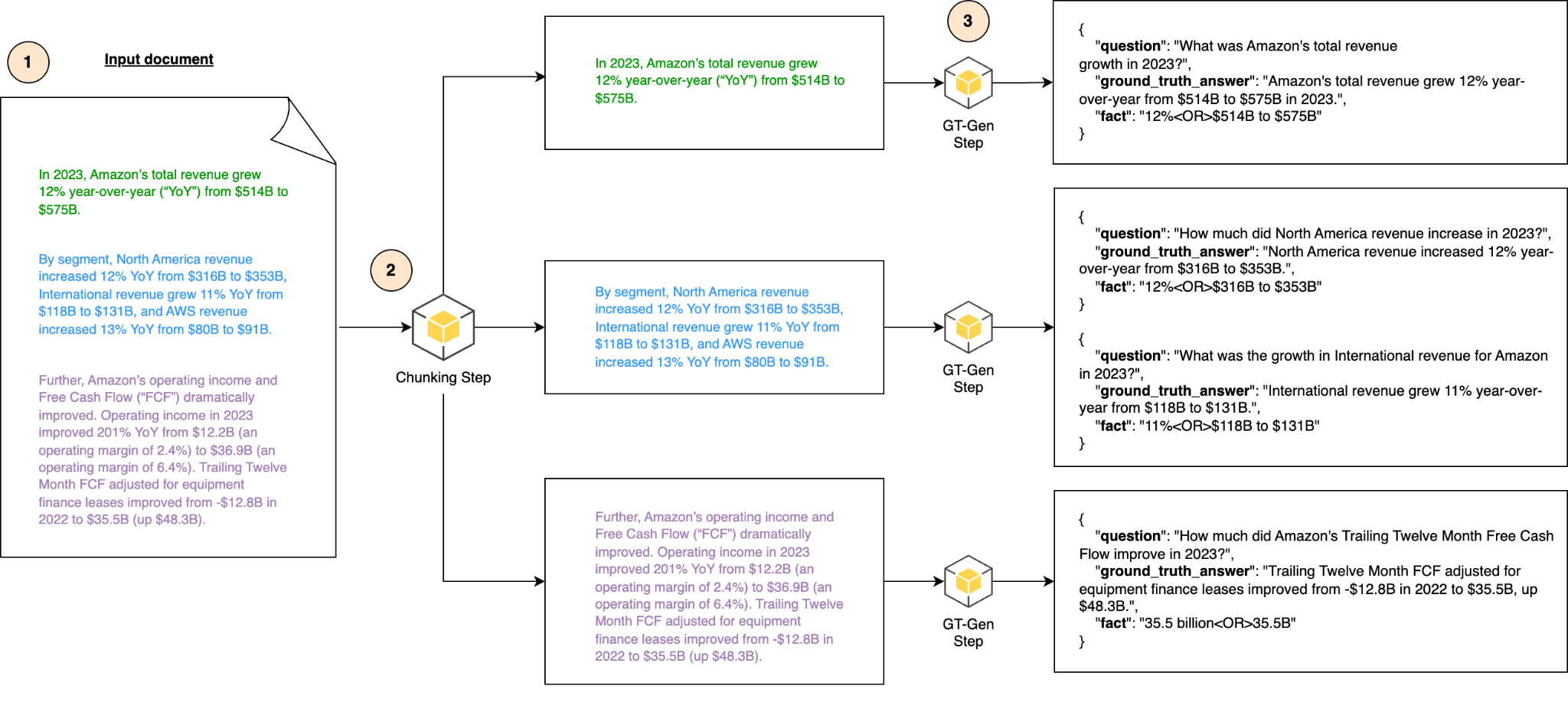
}

}

Sau khi xác thực, [trạng thái distributed map](https://aws.amazon.com/blogs/aws/step-functions-distributed-map-a-serverless-solution-for-large-scale-parallel-data-processing/) đầu tiên lặp qua các tệp trong input bucket để bắt đầu quá trình tiếp nhận và phân đoạn tài liệu với khả năng mở rộng theo chiều ngang. Các chunk kết quả được lưu trữ trong một bucket S3 trung gian.

Distributed map thứ hai là phần cốt lõi của pipeline để tạo dữ liệu. Mỗi chunk được tạo bởi map trước đó được đưa vào làm đầu vào cho prompt tạo ground truth trên Amazon Bedrock. Đối với mỗi chunk, một tệp JSONLines chứa các bộ ba câu hỏi-câu trả lời-sự kiện được xác thực và lưu trữ trong bucket S3 tại tiền tố đầu ra.

Hình sau đây cho thấy cấu trúc dữ liệu và mối quan hệ từ các đoạn văn tài liệu đến chunk ground truth cuối cùng qua các trạng thái chunking và generation map. Các số giữa hai hình chỉ ra cấu trúc dữ liệu hiện diện tại mỗi điểm trong pipeline. Cuối cùng, các tệp JSONLines được tổng hợp trong một [Amazon SageMaker Processing Job](https://docs.aws.amazon.com/sagemaker/latest/dg/processing-job.html), bao gồm việc gán một mẫu ngẫu nhiên để xem xét thủ công dựa trên đầu vào của người dùng.



Bước cuối cùng của pipeline là bước tổng hợp sử dụng [SageMaker Processing](https://docs.aws.amazon.com/sagemaker/latest/dg/processing-job.html) job. Bước tổng hợp này bao gồm việc nối các bản ghi JSONLines được tạo ra bởi mỗi lần thực thi con của generation map thành một tệp đầu ra ground truth duy nhất. Một tỷ lệ phần trăm được chọn ngẫu nhiên từ các bản ghi trong tệp đầu ra sẽ được lấy mẫu và đánh dấu để xem xét như một phần của quy trình human-in-the-loop.

## Đánh giá ground truth cho việc đánh giá câu trả lời trong FMEval

Trong phần này, chúng ta thảo luận về hai thành phần chính trong việc đánh giá chất lượng ground truth: con người trong quy trình (human in the loop) và sử dụng LLM làm trọng tài. Đo lường chất lượng ground truth là một thành phần thiết yếu trong vòng đời đánh giá.

### Con người trong quy trình (Human-in-the-loop)

Mức độ rà soát ground truth cần thiết được xác định bởi rủi ro của việc có ground truth không chính xác và những tác động tiêu cực của nó. Việc rà soát ground truth bởi các chuyên gia use case có thể xác minh liệu logic nghiệp vụ quan trọng có được thể hiện phù hợp trong ground truth hay không. Quy trình rà soát ground truth bởi con người được gọi là human-in-the-loop (HITL), và một ví dụ về quy trình HITL được thể hiện trong hình dưới đây.

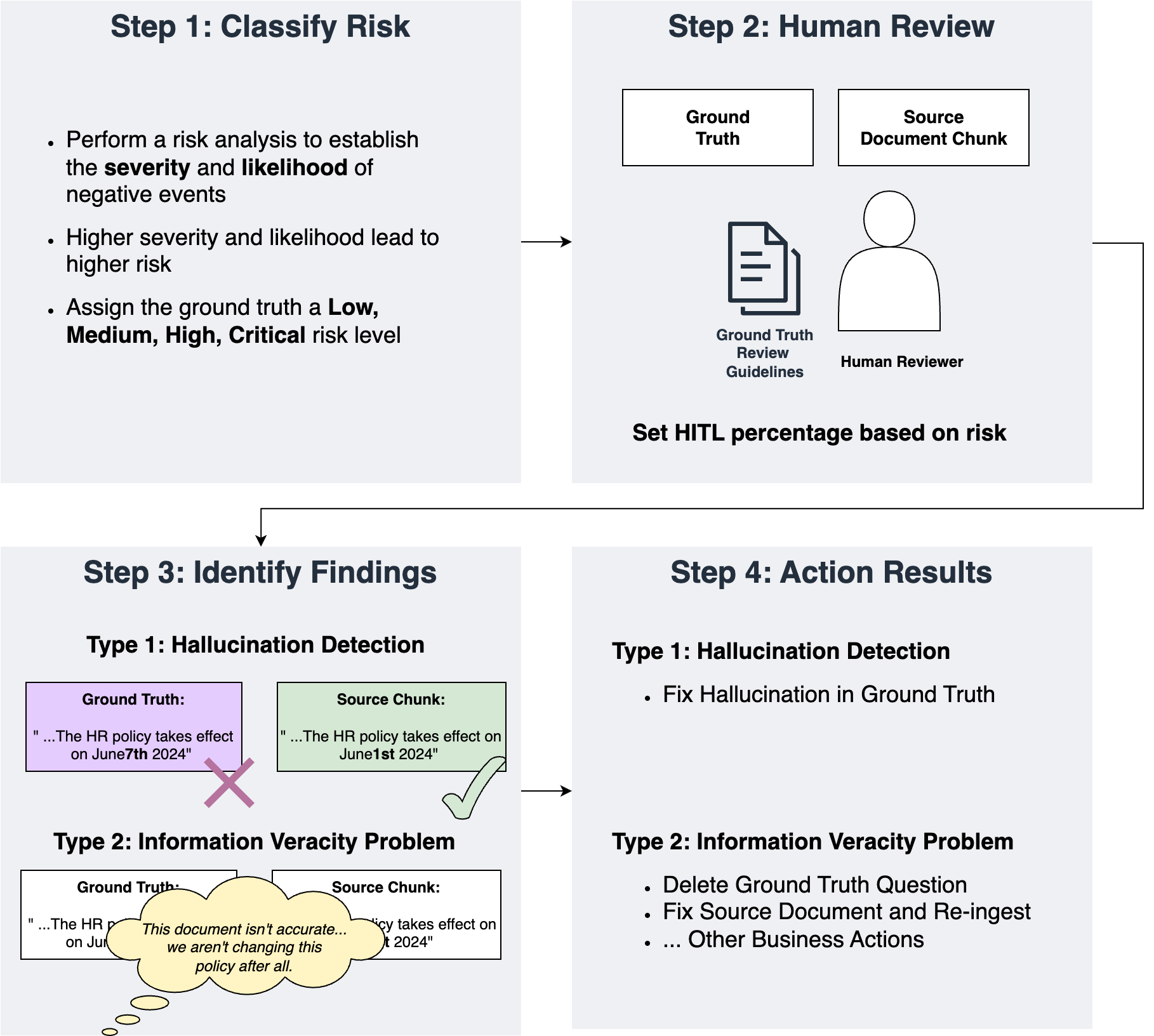
Các bước của HITL bao gồm:

1. **Phân loại rủi ro**: thực hiện phân tích rủi ro sẽ xác định mức độ nghiêm trọng và khả năng xảy ra các sự kiện tiêu cực do việc sử dụng ground truth không chính xác trong đánh giá use case AI tạo sinh. Dựa trên kết quả phân tích, gán mức độ rủi ro cho tập dữ liệu ground truth: Thấp, Trung bình, Cao hoặc Nghiêm trọng. Bảng dưới đây phác thảo mối quan hệ giữa mức độ nghiêm trọng của sự kiện, khả năng xảy ra và mức độ rủi ro. Xem thêm "[Learn how to assess the risk of AI systems](https://aws.amazon.com/blogs/machine-learning/learn-how-to-assess-risk-of-ai-systems/) " để tìm hiểu sâu về đánh giá rủi ro AI.

2. **Rà soát của con người**: Dựa trên mức độ rủi ro được gán, các chuyên gia rà soát use case sẽ kiểm tra một lượng tương ứng ground truth của use case. Các tổ chức có thể thiết lập ngưỡng chấp nhận được cho tỷ lệ can thiệp HITL dựa trên khả năng chấp nhận rủi ro của họ. Tương tự, nếu một tập dữ liệu ground truth được nâng cấp từ rủi ro thấp lên rủi ro trung bình, sẽ cần mức độ can thiệp HITL cao hơn.

3. **Xác định phát hiện**: Người rà soát có thể xác định bất kỳ hallucination nào so với dữ liệu nguồn, thách thức về tính xác thực thông tin theo chuyên môn của họ, hoặc các tiêu chí khác do tổ chức đặt ra. Trong bài viết này, chúng tôi tập trung vào phát hiện hallucination và tính xác thực thông tin.

4. **Hành động dựa trên kết quả**: Người rà soát có thể thực hiện các hành động nghiệp vụ dựa trên đánh giá của họ, như cập nhật và xóa bản ghi, hoặc viết lại các tài liệu nguồn liên quan. Việc đưa các chuyên gia LLMOps vào để áp dụng các Best Practices về quản lý dữ liệu cũng có thể là một kết quả.



Mức độ nghiêm trọng và khả năng xảy ra rủi ro đối với bộ dữ liệu ground truth dùng để xác thực một chatbot đang hoạt động với lượng người dùng cao sẽ lớn hơn so với bộ dữ liệu đánh giá nội bộ được các nhà phát triển sử dụng để cải tiến prototype.

|  | Khả năng xảy ra | | | | |
| --- | --- | --- | --- | --- | --- |
| Mức độ nghiêm trọng | Hiếm | Không chắc | Có thể | Nhiều khả năng | Thường xuyên |
| Cực kỳ nghiêm trọng | Thấp | Trung bình | Cao | Nguy hiểm | Nguy hiểm |
| Nghiêm trọng | Rất thấp | Thấp | Trung bình | Cao | Nguy hiểm |
| Trung bình | Rất thấp | Thấp | Trung bình | Trung bình | Cao |
| Thấp | Rất thấp | Rất thấp | Thấp | Thấp | Trung bình |
| Rất thấp | Rất thấp | Rất thấp | Rất thấp | Rất thấp | Thấp |

Tiếp theo, chúng ta sẽ đi qua quy trình từng bước để thực hiện đánh giá thủ công nhằm phát hiện hallucination và xác minh thông tin. Đánh giá thủ công được thực hiện bằng cách so sánh đoạn dữ liệu ground truth đầu vào của prompt LLM với bộ ba câu hỏi-câu trả lời-sự thật được tạo ra. Bảng dưới đây minh họa điều này.

| Đoạn dữ liệu nguồn | Bộ ba ground truth |
| --- | --- |
| Kính gửi các Cổ đông:  Năm ngoái vào thời điểm này, tôi đã chia sẻ sự nhiệt tình và lạc quan của mình về tương lai của Amazon. Hôm nay, tôi thậm chí còn lạc quan hơn. Có nhiều lý do, nhưng bắt đầu từ tiến bộ chúng tôi đã đạt được trong kết quả tài chính và trải nghiệm khách hàng, và mở rộng đến việc tiếp tục đổi mới cùng những cơ hội đáng chú ý phía trước. Trong năm 2023, tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước ("YoY") từ 514 tỷ USD lên 575 tỷ USD. Theo phân khúc, doanh thu Bắc Mỹ tăng 12% YoY từ 316 tỷ USD lên 353 tỷ USD, doanh thu Quốc tế tăng 11% YoY từ 118 tỷ USD lên 131 tỷ USD, và doanh thu AWS tăng 13% YoY từ 80 tỷ USD lên 91 tỷ USD. | {"question": "Tổng doanh thu của Amazon tăng bao nhiêu trong năm 2023?", "ground\_truth\_answer": "Tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước từ 514 tỷ USD lên 575 tỷ USD trong năm 2023.", "fact": "12%<OR>514 tỷ USD lên 575 tỷ USD"}  {"question": "Doanh thu Bắc Mỹ tăng bao nhiêu trong năm 2023?", "ground\_truth\_answer": "Doanh thu Bắc Mỹ tăng 12% so với cùng kỳ năm trước từ 316 tỷ USD lên 353 tỷ USD.", "fact": "12%<OR>316 tỷ USD lên 353 tỷ USD"}  {"question": "Doanh thu Quốc tế của Amazon tăng bao nhiêu trong năm 2023?", "ground\_truth\_answer": "Doanh thu Quốc tế tăng 11% so với cùng kỳ năm trước từ 118 tỷ USD lên 131 tỷ USD.", "fact": "11%<OR>118 tỷ USD lên 131 tỷ USD"} |

Người đánh giá sau đó xác định và thực hiện hành động dựa trên các phát hiện để sửa chữa hệ thống. Hallucination của LLM là hiện tượng khi LLM tạo ra thông tin nghe có vẻ hợp lý nhưng thực tế không chính xác hoặc vô nghĩa, được trình bày một cách tự tin như thể là sự thật. Các tổ chức có thể đưa ra thêm các tiêu chí đánh giá và chấm điểm ground truth, phù hợp với mức độ rủi ro và yêu cầu của từng trường hợp sử dụng.

Trong việc phát hiện hallucination, người đánh giá tìm kiếm để xác định văn bản đã được LLM tạo ra không chính xác. Bảng dưới đây minh họa một ví dụ về hallucination và cách khắc phục. Người đánh giá sẽ nhận thấy trong dữ liệu nguồn rằng tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước, tuy nhiên câu trả lời ground truth đã hallucinate thành con số 15%. Trong quá trình khắc phục, người đánh giá có thể sửa lại thành 12%.

| Đoạn dữ liệu nguồn | Ví dụ về hallucination | Ví dụ về khắc phục hallucination |
| --- | --- | --- |
| Trong năm 2023, tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước ("YoY") từ 514 tỷ USD lên 575 tỷ USD. | {"question": "Tổng doanh thu của Amazon tăng bao nhiêu trong năm 2023?", "ground\_truth\_answer": "Tổng doanh thu của Amazon tăng **15%** so với cùng kỳ năm trước từ 514 tỷ USD lên 575 tỷ USD trong năm 2023.", "fact": "12%<OR>514 tỷ USD lên 575 tỷ USD"} | {"question": "Tổng doanh thu của Amazon tăng bao nhiêu trong năm 2023?", "ground\_truth\_answer": "Tổng doanh thu của Amazon tăng 12% so với cùng kỳ năm trước từ 514 tỷ USD lên 575 tỷ USD trong năm 2023.", "fact": "12%<OR>514 tỷ USD lên 575 tỷ USD"} |

Trong đánh giá của chuyên gia về tính xác thực, người đánh giá tìm cách xác nhận xem ground truth có thực sự đúng hay không. Ví dụ, dữ liệu nguồn được sử dụng cho prompt tạo ground truth có thể đã lỗi thời hoặc không chính xác. Bảng dưới đây cho thấy góc nhìn của đánh giá HITL bởi một chuyên gia trong lĩnh vực.

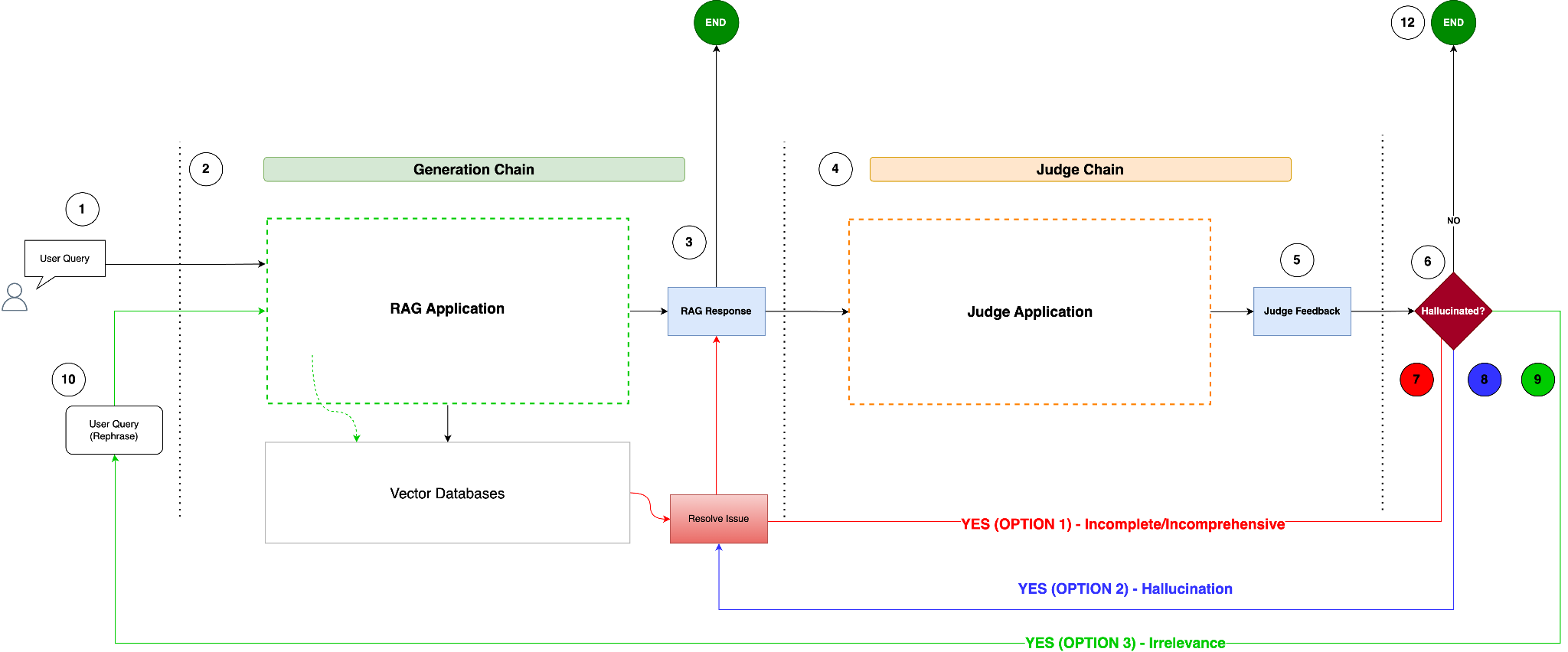
| Đoạn dữ liệu nguồn | Ví dụ về đánh giá của chuyên gia | Ví dụ về các biện pháp khắc phục |
| --- | --- | --- |
| Có hiệu lực từ ngày 1 tháng 6 năm 2023, AnyCompany vui mừng thông báo việc thực hiện "Casual Friday" như một phần của chính sách trang phục cập nhật. Vào thứ Sáu, nhân viên được phép mặc trang phục business casual, bao gồm quần jean gọn gàng, áo polo và giày kín mũi thoải mái. | "Với tư cách là Chuyên gia HR, tôi thấy điều này không chính xác. Chúng tôi đã không thực hiện chính sách Casual Friday tại AnyCompany - dữ liệu nguồn cho ground truth này chắc đã lỗi thời." | Xóa Ground Truth không chính xác  Cập nhật Tài liệu Dữ liệu nguồn  Các hành động khác tùy theo trường hợp sử dụng cụ thể |

Các ứng dụng machine learning truyền thống cũng có thể cung cấp thông tin cho thiết kế quy trình HITL. Để xem ví dụ về HITL cho machine learning truyền thống, hãy tham khảo [Human-in-the-loop review of model explanations with Amazon SageMaker Clarify and Amazon A2I](https://aws.amazon.com/blogs/machine-learning/human-in-the-loop-review-of-model-explanations-with-amazon-sagemaker-clarify-and-amazon-a2i/).

## LLM-as-a-judge

Khi mở rộng quy mô HITL, các trình đánh giá LLM có thể thực hiện [phát hiện và khắc phục hiện tượng ảo giác](https://arxiv.org/abs/2305.11747). Ý tưởng này được gọi là [RAG tự phản ánh](https://arxiv.org/abs/2310.11511?ref=blog.langchain.dev) và có thể được sử dụng để giảm - nhưng không loại bỏ hoàn toàn - mức độ nỗ lực của con người trong quá trình phát hiện ảo giác. Như một phương tiện để mở rộng quy mô đánh giá LLM-as-a-judge, Amazon Bedrock giờ đây cung cấp khả năng sử dụng trình đánh giá LLM và thực hiện kiểm tra lập luận tự động với Amazon Bedrock Guardrails để tự xác thực chính xác về mặt toán học dựa trên các chính sách được xác định trước. Để biết thêm thông tin về việc triển khai, hãy xem "New RAG evaluation and LLM-as-a-judge capabilities in Amazon Bedrock" và "[Prevent factual errors from LLM hallucinations with mathematically sound Automated Reasoning checks (preview)](https://aws.amazon.com/blogs/aws/new-rag-evaluation-and-llm-as-a-judge-capabilities-in-amazon-bedrock/)".

Hình dưới đây cho thấy một ví dụ sơ đồ cấp cao về mô hình RAG tự phản ánh. Một ứng dụng AI tạo sinh dựa trên RAG tạo ra các phản hồi được gửi đến ứng dụng đánh giá. Ứng dụng đánh giá xem xét liệu các phản hồi có không đầy đủ, ảo giác hay không liên quan. Dựa trên đánh giá, dữ liệu được định tuyến theo phương pháp khắc phục tương ứng.



Một sơ đồ thể hiện chuỗi tạo sinh theo sau bởi chuỗi đánh giá, định tuyến thông minh các yêu cầu quay lại nếu cần để xếp hạng lại

Quy tắc vàng trong việc triển khai HITL hoặc LLM-as-a-judge như một phần của quá trình tạo ground truth là đảm bảo quy trình đánh giá của tổ chức phù hợp với mức độ rủi ro được chấp nhận cho tập dữ liệu ground truth.

## Kết luận

Trong bài viết này, chúng tôi đã cung cấp hướng dẫn về việc tạo và đánh giá ground truth để đánh giá các ứng dụng hỏi đáp sử dụng FMEval. Chúng tôi đã khám phá các phương pháp tốt nhất để áp dụng LLM nhằm mở rộng quy mô tạo ground truth trong khi vẫn duy trì chất lượng và độ chính xác. Kiến trúc pipeline batch serverless mà chúng tôi trình bày cung cấp giải pháp có thể mở rộng để tự động hóa quy trình này trên các cơ sở kiến thức doanh nghiệp lớn. Chúng tôi cung cấp một prompt tạo ground truth mà bạn có thể sử dụng để bắt đầu đánh giá trợ lý tri thức bằng các chỉ số đánh giá Kiến thức Thực tế và Độ chính xác QA của FMEval.

Bằng cách tuân theo các hướng dẫn này, các tổ chức có thể tuân thủ các phương pháp tốt nhất về AI có trách nhiệm để tạo ra các tập dữ liệu ground truth chất lượng cao cho việc đánh giá xác định các trợ lý hỏi đáp. Các đánh giá dành riêng cho trường hợp sử dụng được hỗ trợ bởi ground truth được quản lý tốt đóng vai trò quan trọng trong việc phát triển và triển khai các giải pháp AI đáp ứng các tiêu chuẩn cao nhất về chất lượng và trách nhiệm.

Cho dù bạn đang phát triển một công cụ nội bộ, một trợ lý ảo phục vụ khách hàng, hay khám phá tiềm năng của AI tạo sinh cho tổ chức của mình, chúng tôi khuyến khích bạn áp dụng những phương pháp tốt nhất này. Hãy bắt đầu triển khai quy trình tạo và đánh giá ground truth mạnh mẽ cho các đánh giá hỏi đáp AI tạo sinh của bạn ngay hôm nay với FMEval.

Về các tác giả



**Samantha Stuart** là một Nhà khoa học Dữ liệu tại AWS Professional Services, và đã thực hiện nhiều dự án về AI tạo sinh, MLOps và ETL cho khách hàng. Samantha có bằng thạc sĩ nghiên cứu về kỹ thuật từ Đại học Toronto, nơi cô đã công bố một số ấn phẩm về AI tập trung vào dữ liệu cho thiết kế hệ thống phân phối thuốc. Ngoài công việc, cô thường dành thời gian chơi nhạc, gặp gỡ bạn bè và gia đình, tập yoga hoặc khám phá Toronto.



**Philippe Duplessis-Guindon** là chuyên gia tư vấn đám mây tại AWS, nơi anh đã làm việc trên nhiều dự án AI tạo sinh. Anh đã tham gia vào hầu hết các khía cạnh của những dự án này, từ cơ sở hạ tầng và DevOps đến phát triển phần mềm và AI/ML. Sau khi tốt nghiệp cử nhân kỹ thuật phần mềm và thạc sĩ về thị giác máy tính và học máy từ Polytechnique Montreal, Philippe gia nhập AWS để áp dụng chuyên môn của mình phục vụ khách hàng. Khi không làm việc, bạn có thể tìm thấy Philippe ngoài trời - hoặc đang leo núi hoặc chạy bộ.



**Rahul Jani** là Kiến trúc sư Dữ liệu tại AWS Professional Service. Anh hợp tác chặt chẽ với các khách hàng doanh nghiệp xây dựng nền tảng dữ liệu hiện đại, ứng dụng AI tạo sinh và MLOps. Anh chuyên về thiết kế và triển khai các ứng dụng dữ liệu lớn và phân tích trên nền tảng AWS. Ngoài công việc, anh trân trọng thời gian bên gia đình và nắm bắt cơ hội đi du lịch.



**Ivan Cui** là Trưởng nhóm Khoa học Dữ liệu tại AWS Professional Services, nơi anh giúp khách hàng xây dựng và triển khai các giải pháp sử dụng ML và AI tạo sinh trên AWS. Anh đã làm việc với khách hàng từ nhiều ngành công nghiệp đa dạng, bao gồm phần mềm, tài chính, dược phẩm, chăm sóc sức khỏe, IoT, và giải trí và truyền thông. Trong thời gian rảnh, anh thích đọc sách, dành thời gian bên gia đình và du lịch.