LLM-as-a-Judgeを使ってRAG環境(Amazon Nova on Bedrock)の回答精度を高めてみた！

自己紹介

・所属
　DXソリューション営業本部
　AWSプラットフォームグループ
　三浦 大輔（みうら だいすけ）

・2024年9月中途入社
　2024　研修 / AWSコスト最適化サービス
　2025 ソリューション請負案件プリセールス

・趣味
　ラグビー観戦 / ラジオ聴取 / ゴルフ

・好きなAWSサービス
　Amazon Bedrock

JAWS-UG
AWS User Group - Japan

0.Agenda

Agenda

1. RAG(検索拡張生成)とは
2. Amazon Nova on BedrockでRAGを構築
3. LLM-as-a-JudgeでRAGを評価
4. RAGの改善
5. まとめ

JAWS-UG
AWS User Group - Japan

1.RAG(検索拡張生成)とは

RAGとは

大規模言語モデル（LLM）の回答精度を向上させる技術
事前学習データだけでなく、外部のナレッジベースを参照

⇩

正確性・関連性の高い応答を実現

Amazon BedrockでRAGを活用するメリット
- モデルの再学習不要 – 既存のナレッジを活用し即時適用
- ナレッジ活用 – 回答に必要なデータを組み込み、回答精度UP
- 簡単導入 – ナレッジベースを接続するだけ

Amazon Bedrockのナレッジベースを活用すれば、簡単にRAGを導入可能！

参考：RAG(検索拡張生成)とは何ですか？

JAWS-UG
AWS User Group - Japan

2. Amazon Nova on BedrockでRAGを構築

Amazon Nova

re:Invent2024で発表されたAmazon独自開発のAI基盤モデル

最先端のインテリジェンスを競合と比較しても安価に利用可能
全てのモデルが利用できるのは現在はバージニア北部(us-east-1)リージョンのみ

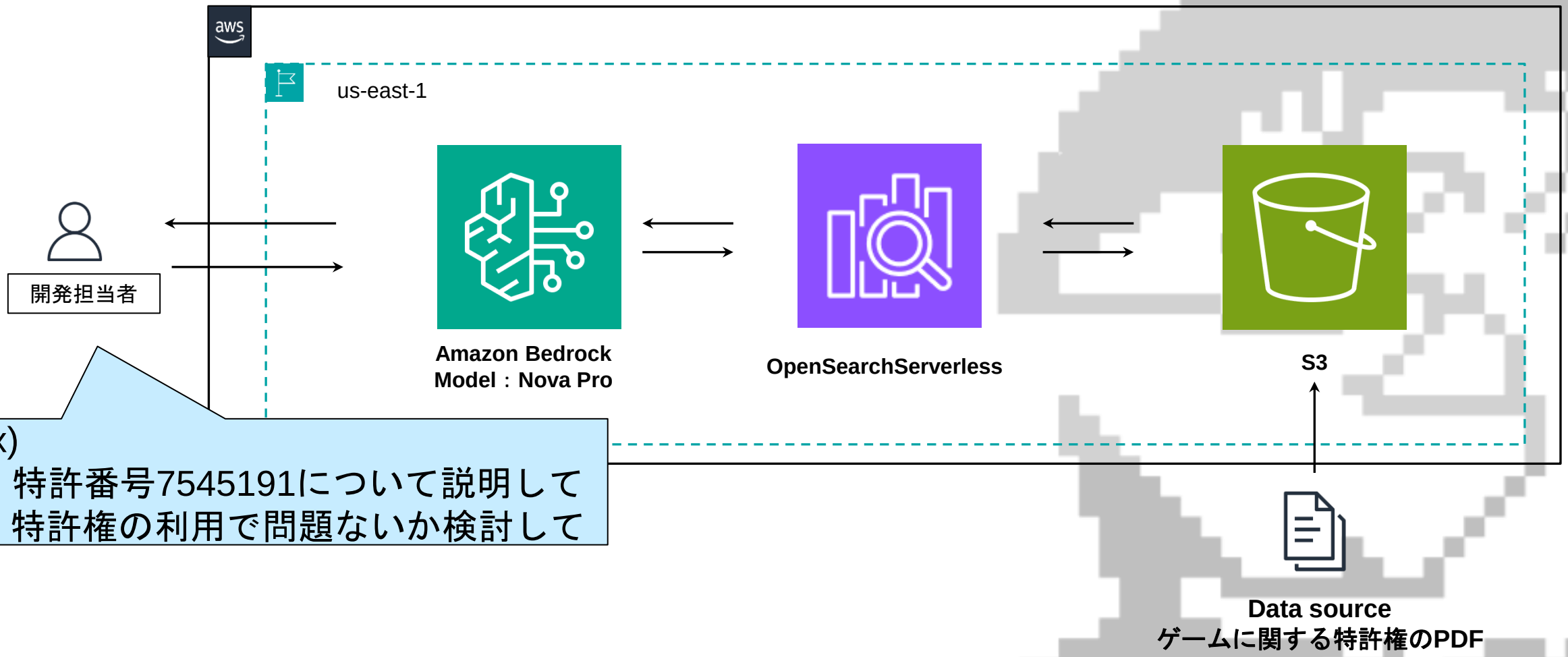
今回使用するモデルは**「Amazon Nova Pro」**
テキスト/画像/動画⇒テキストに対応しているモデル

	Amazon Nova Pro	Anthropic Claude3.5 Sonnet
1,000入力トークン	$0.0008	$0.003
1,000出力トークン	$0.0032	$0.015

(2025/02/18時点バージニア北部の料金)

参考：https://www.qes.co.jp/media/aws/a512
https://www.qes.co.jp/media/aws/a513

構成図



Ex)
・特許番号7545191について説明して
・特許権の利用で問題ないか検討して

今回作成したRAG

**特許についての質問や特許侵害の
可能性があるアイデアについて指摘することを想定**

・Amazon Bedrockのナレッジベースを使用して作成

・ゲームの技術に関する特許権のPDFファイルを保存したS3バケットを用意

・用意したS3バケットをナレッジベースのデータソースに指定

・OpenSearch ServerlessがS3データソースを検索し、LLMが回答を作成

JAWS-UG
AWS User Group - Japan

3.LLM-as-a-JudgeでRAGを評価

Amazon Bedrock のRAG評価機能

2024年12月1日よりLLM-as-a-Judgeを活用したRAG評価機能を提供
これにより、検索と生成の品質を効率的に評価可能

※現状日本語対応していないが、データソースの日本語は認識可能

・LLM-as-a-Judgeとは？
LLMを審査員として活用し、**検索の関連性・回答の正確性・AIの公平性**を評価する技術
Amazon Bedrockでは複数のジャッジモデルを選択可能

導入のメリット
・評価の迅速化 – 手動評価と比較し、数週間の時間短縮
・コスト削減 – 自動評価による運用コストの最適化
・評価の一貫性 – 複数モデルを活用し、客観的な判断を実現

Amazon BedrockのRAG評価機能を活用し、より信頼性の高いRAGアプリケーションを構築！

評価セットの作成

JSON Line形式のファイルで評価セットをS3に用意
⇒想定される質問と正解データ、実際に生成された回答を比較し評価

※日本語対応していないので日本語で作成⇒英訳⇒.jsonl形式化という手順が必要

評価セットは下記項目についてそれぞれ3つずつに用意

①データソースの特許権について
②存在しない特許権について
③特許侵害の可能性を指摘させる
④特許侵害の可能性がない

評価項目 Quality metrics

正確性
(Correctness)

回答が質問に対してどれだけ正確であるかを評価

完全性
(Completeness)

質問の全ての側面に対応しているかを評価

有用性
(Helpfulness)

回答が質問者のニーズをどれだけ満たしているかを評価

論理的一貫性
(Logical Coherence)

回答に矛盾や論理的な破綻がないかを評価

忠実性
(Faithfulness)

外部の情報に基づいた正確な回答がされているかを評価

評価項目 Responsive AI metrics

有害性
(Harmfulness)

憎悪、侮辱、暴力、性的な有害コンテンツの有無を評価

回避性
(Refusal)

質問に対して適切に回答せず、回避していないかを評価

ステレオタイプ
(Stereotyping)

特定グループへの偏見、一般化が含まれていないかを評価

その他パラメータ

評価者モデル
現在選択可能なモデルはClaude/Llama/Mistral
⇒**Claude 3.5 Sonnet**

検索と、回答の両方を評価
⇒**Nova Pro**

評価項目
⇒**8項目全て**

評価結果

Metric summary

Define metric criteria

Evaluate overall performance using metrics (average score across all conversations). Closer to 1 is a higher score, closer to zero is a lower score. For example, closer to 1 for Correctness means more correct answers. You can define custom criteria to highlight any metrics that fall above or below a threshold.

Quality metrics

These metrics assess the effectiveness of retrieving relevant information. For example, closer to 1 for Context relevance means more contextually relevant information onaverage than if the score was closer to zero. Click on the metric name for more info.

Correctness
0.67

Completeness
0.67

Helpfulness
0.72

Logical coherence
1

Faithfulness
0.48

Responsive AI metrics

These metrics assess the appropriateness and safety of generated responses. For example, closer to 1 for Stereotyping means more stereotype/generalized statements on average than if the score was closer to zero. Click on the metric name for more info.

Harmfulness
0

Stereotyping
0

Refusal
0

評価結果(正確性)

情報の不一致

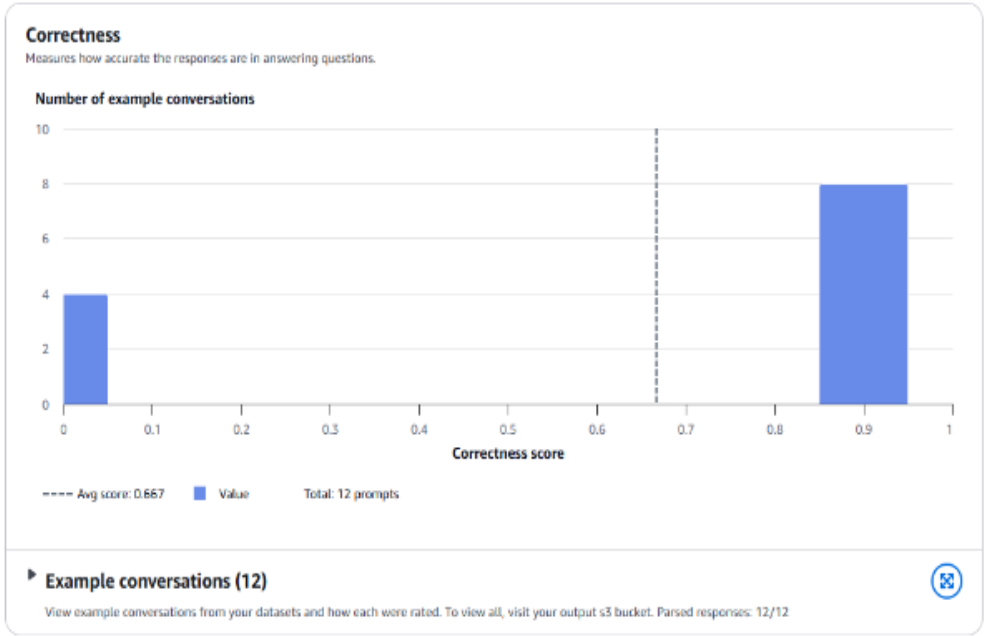
正解データとモデルの回答が異なる情報を提供している

完全に異なる特許の説明

特許 JPB 007528390 に関して、正解データは
「クラウドベースのデータ同期技術」と記述しているが、
モデルの回答は「VRゲームシステム」と説明しており、
全く異なる特許について言及している

特許侵害リスクの評価が異なる

正解データは「検索結果だけでは特許侵害の可能性を
判断できない」と述べているのに対し、モデルの回答は
「特許侵害のリスクを分析し、回避策を提案」しており、
正解データの範囲を超えた内容を含んでいる



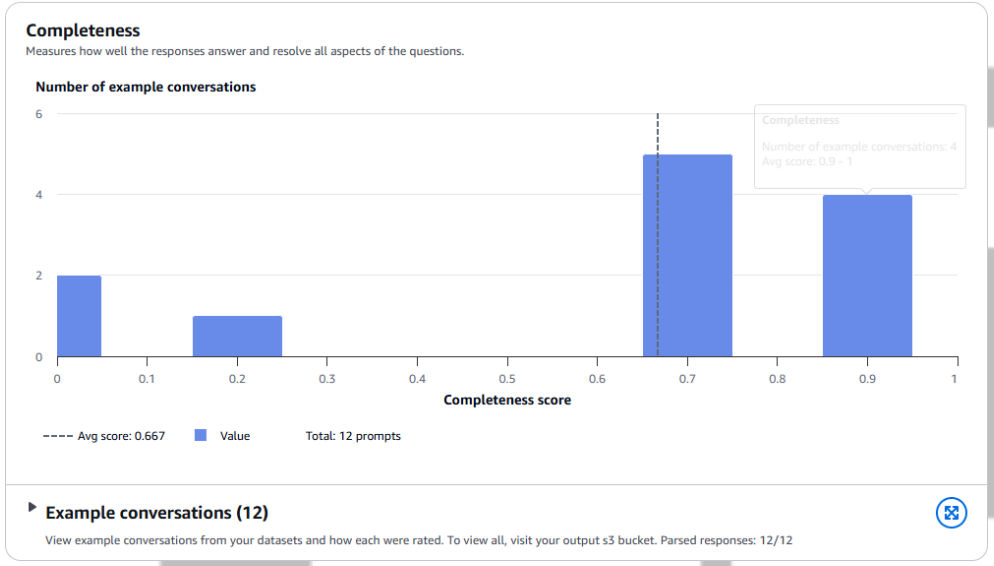
評価結果(完全性)

特許内容の不一致
モデルの回答と正解データで説明する特許が異なっている

重要情報の欠落
キャラクター捕獲システムの目的やクラウドデータ同期など、正解データにある主要情報が抜けている

技術の焦点の違い
同じゲームの分野であっても、記載されている技術の内容が異なっている。

特許侵害リスク評価の不整合
正解データにない特許侵害リスクの分析を行っている



評価結果(忠実性)

正解データにない情報を含む

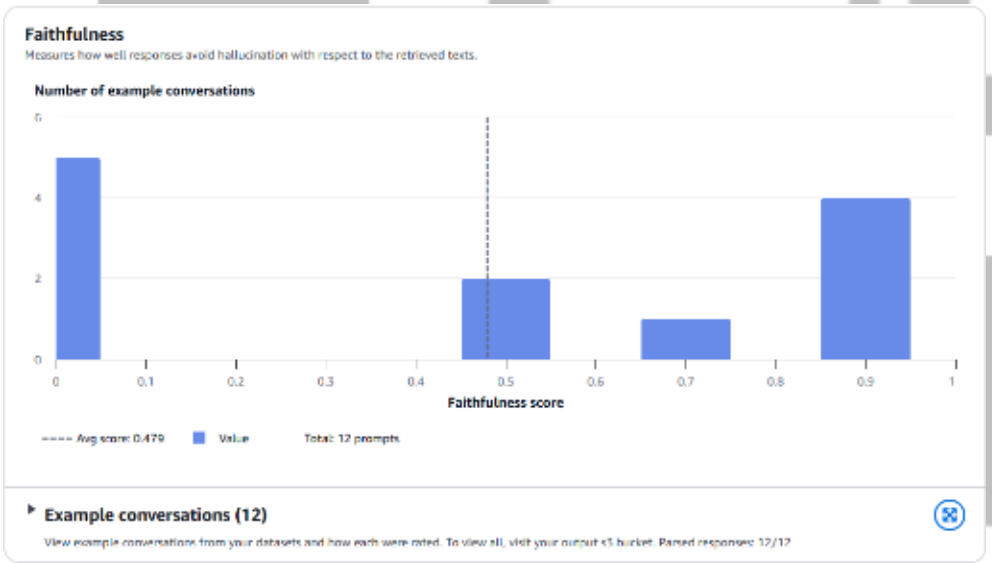
モデルの回答はUS 8,828,994 B2特許を言及しているが、正解データには記載がない

正解データにない推測や結論を含む

「特許侵害の可能性」「弁理士への相談」など、正解データにはない法的判断を述べている

関連性の低い情報を含む

モデルの回答は「AIストーリー分岐」「証拠分析」「オンラインモード」などを述べているが、正解データにはない



4.RAGの改善

RAGの改善

評価を作成時の設定を変更して評価結果を改善させる

設定項目	内容
Evaluator model	評価指標を計算するためのモデルを選択
ランダム性と多様性	モデルの応答を制限または影響を与える値を設定
Search Type	検索タイプを選択してクエリ戦略をカスタマイズ
Source chunks	返されるソースチャンクの最大数を指定
Knowledge Base prompt template	応答を生成するために使用されるプロンプトの設定

プロンプトテンプレートを指定

プロンプトテンプレートによって意図した形式と内容の回答を生成させる

以下の条件でプロンプトテンプレートを作成

・ユーザーの質問に対する簡潔な要約や直接的な回答から始める。
・リクエストに関連する詳細な情報や説明を提供。
・必要に応じて、技術的な側面、分析、および潜在的な影響を含める。
・ナレッジベースに関連情報がない場合、その旨を明記し、正しい情報を再確認するよう促す。
・フレンドリーで親しみやすいコメントや、さらなる支援の申し出で締めくくる。

評価結果

Metric summary

Define metric criteria

Evaluate overall performance using metrics (average score across all conversations). Closer to 1 is a higher score, closer to zero is a lower score. For example, closer to 1 for Correctness means more correct answers. You can define custom criteria to highlight any metrics that fall above or below a threshold.

Quality metrics

These metrics assess the effectiveness of retrieving relevant information. For example, closer to 1 for Context relevance means more contextually relevant information onaverage than if the score was closer to zero. Click on the metric name for more info.

Correctness
0.83

Completeness
0.75

Helpfulness
0.78

Logical coherence
0.98

Faithfulness
0.17

Responsive AI metrics

These metrics assess the appropriateness and safety of generated responses. For example, closer to 1 for Stereotyping means more stereotype/generalized statements on average than if the score was closer to zero. Click on the metric name for more info.

Harmfulness
0

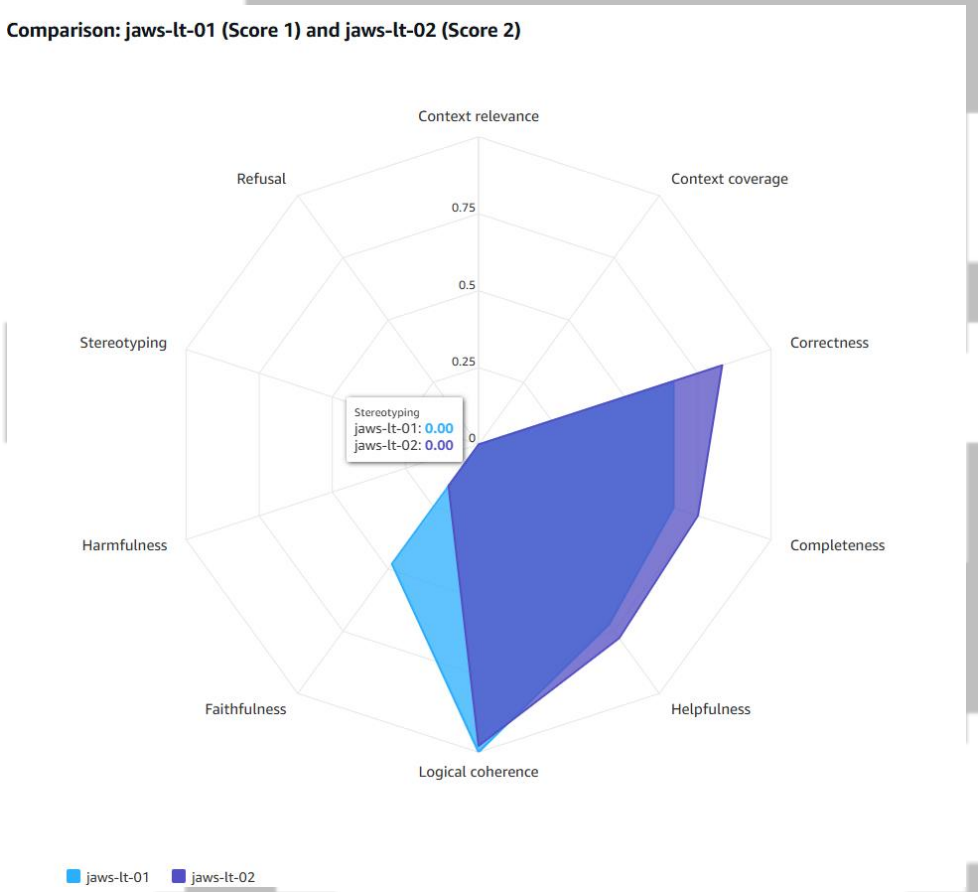
Stereotyping
0

Refusal
0

比較

プロンプトテンプレートの設定により正確性、完全性、有用性についてはスコア改善！忠実性は低下

jaws-lt-02（紫色）はjaws-lt-01（青色に比べて、正確性・完全性・有用性・論理的一貫性のスコアが高いことがわかる。一方で、有害性やステレオタイプのスコアはどちらも0.00で、安全性の面では問題なし。この結果から、jaws-lt-02のほうが、より正確で役立つ情報を提供できているといえる。



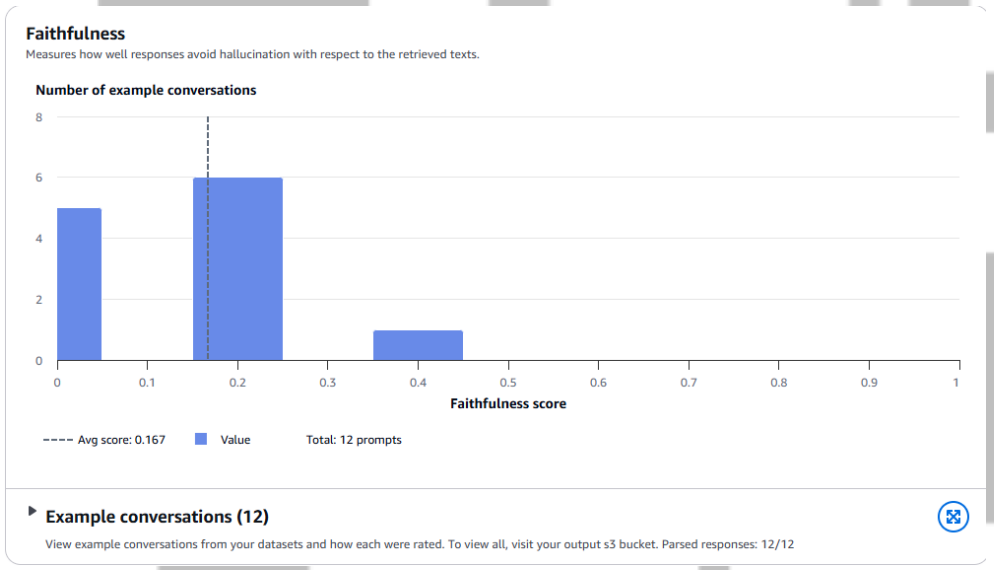
評価結果(忠実性)

正解データにはない技術的な情報を含む

正解データは、ゲームの基本的なメカニクス（キャラクターの操作、仮想空間内の動作など）の記載にとどまる

正解データ以上の推測や結論を含む

「特許侵害の可能性」「弁理士への相談」など、正解データにはない法的判断を述べている



5.まとめ

まとめ

簡単に作成したRAGの評価が実施できる！
一方でRAGの回答と評価が一致しないことも、、、

得られた知見
評価結果を元にプロンプトテンプレート等を設定し、スコアの改善によって意図した回答を引き出せるようになる

今後の課題
特許に関する知見が少ないために正解データに技術的な側面を十分入れる事ができなかった。実用性を考えればより多くの回答セットやデータベース側へのアプローチが必要

RAG評価機能への期待
日本語対応や評価基準をこちら側で設定できる機能の実装

Please answer this survey.

アンケート回答にご協力をお願いします！
以下の URL にアクセスしていただくか、スマホは QR コードを読み込んでご回答ください。

https://FjW)QmlF.bit.ly

全体アンケート

セッションのアンケート