大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

信頼性/正確性を見える化した RAG評価フレームワークの提案

伊左次 翔太 NECソリューションイノベータ株式会社　植松 凌太 株式会社日立製作所
川西 昂弥 三菱電機ソフトウエア株式会社　濱 憲仁 富士通株式会社

RAGの評価における問題点

- 近年 RAG（検索拡張生成）の需要が増加しており、その品質評価の重要性が一層高まっている。
- RAG 評価フレームワークである RAGAS を試行した結果、多角的な観点で性能を評価できるが、多数の指標が数値のみで出力されるため、専門知識を持たないステークホルダーへの説明には不向きであると考えた。

* RAG: Retrieval-Augmented Generation,
RAGAS: RAG ASsessment

問題解決のアプローチ

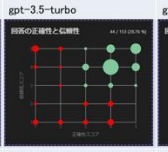
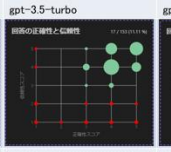
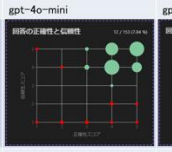
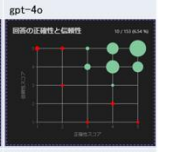
- 説明性を三つの要件に分解し、ステークホルダーが信頼・納得できる評価フレームワークを提案。
- RAGAS の指標を単純に可視化するだけでは不十分な点について、新たに三つの技術要素を開発。
  (A) 信頼性スコア: 回答と検索結果の関連度を評価
  (B) 原因分析: 回答不備の原因をレベル別に分類
  (C) 質問分析: 質問内容を難易度・テーマ別に分類

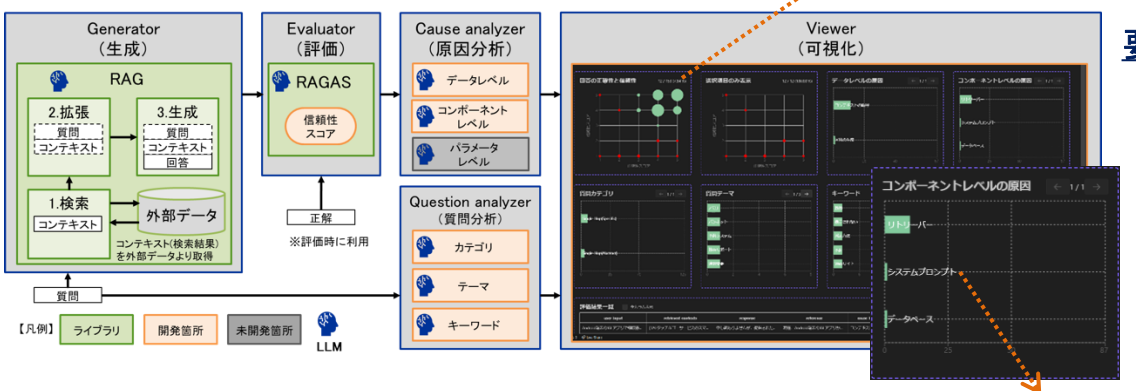
信頼性/正確性を見える化したRAG評価フレームワーク

要件①: RAGの性能が一目でわかる

正確性と信頼性の二つの観点で可視化

- 正確性スコア（5段階評価） RAGAS指標「Rubrics score」
- 信頼性スコア（5段階評価） 独自作成指標
  → 回答がコンテキストに基づいているかを評価可能に

	パターン①（開発初期1）	パターン②（開発初期2）	パターン③（開発中期1）	パターン④（開発中期2）	パターン⑤（開発中期3）
データ量	全データの1/3	全データの2/3	全データ	全データ	全データ
生成モデル	gpt-3.5-turbo	gpt-3.5-turbo	gpt-3.5-turbo	gpt-4o-mini	gpt-4o
結果					
低スコア回答数	111件（全体の72.55%）	44件（全体の28.76%）	17件（全体の11.11%）	12件（全体の7.84%）	10件（全体の6.54%）



要件②: できないこと、できない原因が明確にわかる

低スコア回答をカテゴリ別に可視化

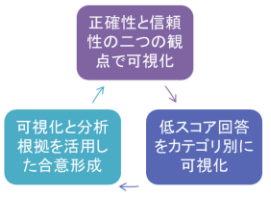
- 原因分析（データレベル）
  コンテキストの取得、回答の生成、正解の定義 など
- 原因分析（コンポーネントレベル）
  データベース、リトリーバー 、生成モデル、システムプロンプト など
- 質問分析（難易度別）
  単一検索（具体的）、単一検索（抽象的）、多段検索（具体的） など

→ 回答不備の特徴・傾向を分析可能に

要件③: 「どこをめざすべきか」をステークホルダーと議論できる

可視化と分析根拠を活用した合意形成

- 可視化を軸とした改善サイクル
- 分析根拠に基づいた詳細説明
  → 影響範囲と優先度を考慮した議論を実現



＜分析根拠＞
回答のfaithfulnessが0.67と低く、response_relevancyが0.0と極端に低いことから、回答がコンテキストに関連していない可能性が高いです。

コンテキストには、悪天候や自然災害の影響で搭乗予定便が欠航になった場合の払い戻し手続きが可能であることが明確に記載されていますが、生成された回答は払い戻しができないと述べています。このため、システムプロンプトがコンテキストの情報を正確に反映するように指示できていないことが原因と考えられます。

検証結果

技術要素の妥当性を確認

- 信頼性スコアが、期待通り、回答の裏付け有無の傾向をとらえた評価値を算出。
- 原因分析が、各ケースで、RAG の要素の内、不適切であった部分を原因として抽出。

信頼性スコアの評価結果（WikiEval[1] データセットを一部改変して利用） ＊ヒートマップとして可視化

項目	説明	信頼性スコアの評価値ごとの件数					評価値
データ	コンテキストの種類	Score 5	Score 4	Score 3	Score 2	Score 1	平均
Data 1	理想的なコンテキスト	30	20	0	0	0	4.60
Data 2	冗長なコンテキスト	30	20	0	0	0	4.60
Data 3	欠損したコンテキスト	2	4	10	9	25	1.98
Data 4	関連なしコンテキスト	0	0	0	0	50	1.00

すべてのデータで回答には正解を使用し、異なるコンテキストでの評価結果を比較
信頼性（高） ← → 信頼性（低）
＊青枠　は各行で正解として期待するスコア、太字は各行で最多となる件数

原因分析（コンポーネントレベル）の分析結果 ＊ヒートマップとして可視化

項目	説明	分類された原因（コンポーネントレベル）ごとの件数					
名称	RAGの特徴	不備なし	データベース	リトリーバー	生成モデル	システムプロンプト	その他
RAG 1	理想形（正解と回答が一致）	49	1	0	0	0	0
RAG 2	標準形（意図的な不備なし）	41	3	0	0	3	3
RAG 3	データベースに不備	2	40	8	0	0	0
RAG 4	リトリーバーに不備	1	0	48	0	0	1
RAG 5	生成モデルに不備	13	2	0	35	0	0
RAG 6	システムプロンプトに不備	3	0	0	10	37	0

＊青枠　は各行で正解として期待する原因（コンポーネントレベル）、太字は各行で最多となる件数

出典[1]：https://huggingface.co/datasets/explodinggradients/WikiEval

今後の展望

- **原因分析の精度向上・追加実装**
  （コンポーネントレベル、パラメータレベルへと分析を深度化）
- **多様なデータでの評価**
  （複数の情報を統合しないと回答できない質問では未検証）
- **RAG改良前後で評価結果の差分を確認可能に**
  （改良前は回答できていた質問が、改良後は回答できなくなるケースに気づける仕組みを導入する）