最先端ソフトウェア工学ゼミ 成果報告

2024年7月 11日

氏名 西山美恵子
所属 株式会社クレスコ

発表内容

1.設定テーマと選定理由
2.調査結果
3.考察
4.まとめ
5.参考文献

1. 設定テーマと選定理由

■選定テーマ

RAGの自動評価ツールにおける評価観点について

■選定理由

ハルシネーションを回避し,生成AIの回答精度を高める技術としてRAG注目されている.
実際にRAG構成のシステム開発をしている, 3つのプロジェクトに対してRAGシステムの品質評価に関してヒアリングをした.
その結果,いくつか課題あった.それらの課題を整理した.

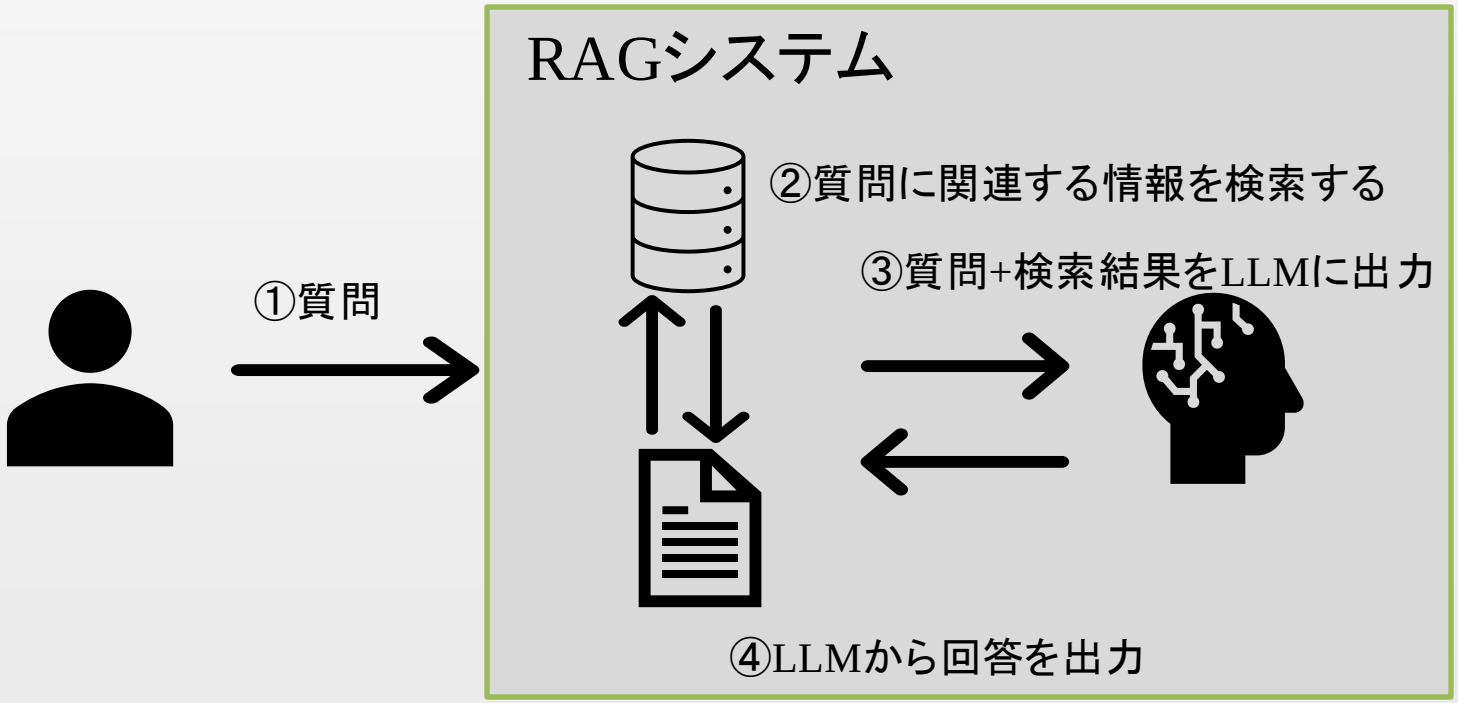
課題1:LLMの評価軸には何があるのか明確でない
課題2:RAGの評価にはLLMの評価軸のどれが重要なのか
課題3:RAGの評価ツールにはどんなものがあるのか
課題4:RAGの評価で重要とされているものが評価ツールで評価できるのか

上記課題に対して,課題と後述する調査のStepを対応付けて調査していく.

1. 設定テーマと選定理由

■RAGとは

LLMで学習済みの内容とは別に,外部の知識やデータを検索し,
その内容に基づいてLLMが回答する仕組みのこと.



2. 調査方針

調査を進めるにあたって,以下4つのStepで実施していく.

Step1：LLMの評価軸には何があるのかを調査する.
Step2：LLMの評価軸をもとに,RAGの開発者にどの評価軸を重要視するのかヒアリングをする.
Step3：RAGの自動評価ツールにはどのようなツールがあるのか調査する.
Step4：Step2で重要だとされる評価軸が,RAGの自動評価ツールで評価できるのか比較検討する.

2. 調査結果 Step1

LLM + 評価をキーワードとして論文やネット記事を検索し,被引用数が大方ものを採用した.

LLMの評価軸	説明
真実性	LLMが生成する情報の正確性と信頼性を指す。誤情報を避け、正しい知識を提供する能力が求められる。
安全性	LLMが害を与えないように設計されていることを示す。例えば、悪意のある利用や暴力的なコンテンツの生成を防ぐこと。
公平性	色々なバックグラウンドや意見を公平に扱い、バイアスを避けることが重要である。特定のグループを差別せず偏りのない対応を目指す。
堅牢性	さまざまな状況や入力に対して安定して正しい結果を出す能力を示す。予期せぬ入力や環境の変化にも適応できるかが評価される。
プライバシー	ユーザーの個人情報を適切に保護することを指す。ユーザーのデータが不適切に利用されないようにする措置が求められる。
機械倫理	LLMが倫理的な行動をするよう設計されているかを示す。倫理的なジレンマに適切に対応しうるかが評価される。
透明性	LLMの動作や決定の過程が明確であることを示す。ユーザーがモデルの仕組みを理解できるようにすることが重要である。
説明責任	LLMが行ったことやその結果に対する責任を負うことを示す。問題が発生した際に、どのように対応するか、誰が責任を負うかが明確にされているかが評価される。

1. https://arxiv.org/pdf/2401.05561 2. https://www.brainpad.co.jp/doors/contents/apply_generative_ai_to_business_tips/

LLMモデルそのものを評価する際の8つの評価軸があることが分かった.

2. 調査結果 Step2

■RAG構成のシステムの評価軸について
LLMの8つの評価軸をもとに,RAGシステムの開発者にRAGシステムで重要だと思う順に順位付けのアンケートをしてもらった.結果の順位を加算した.加算した数値が小さい項目が重要だと捉えることにする.

PJ種別	用途	真実性	安全性	公平性	堅牢性	プライバシー	機械倫理	透明性	説明責任
社内開発	社内FAQ	3位	8位	4位	5位	6位	7位	2位	1位
生損保	社内FAQ	1位	8位	6位	2位	5位	7位	3位	4位
専門商社	論文検索	5位	4位	6位	3位	1位	8位	7位	2位
合計		9	20	16	10	12	22	12	7

PJによってばらつきはあるが真実性,堅牢性,説明責任,が上位になった.ただし,専門商社ではプライバシーが最も重要とされていた.

2. 調査結果 Step3

■RAG構成のシステムの評価ツール
当研究では,3つあるRAGの自動評価ツールのうち,指標数および論文被引用数が最も多いRAGASを取り上げた.
上記以外にも, RAGASは実際に利用してみた記事も多く,実用性が高いと考えられるため当ツールを取り上げた.

ツール	指標数	論文被引用数
RAGAS	10	68
ARES	3	25
TruLens	3	–

ARES,TruLensはRAGASと指標が被っていたため今回は取り上げない.

2. 調査結果 Step4

■RAG構成のシステムの評価軸とメトリクスについて
LLMの評価軸とRAGの自動評価ツールであるRAGASのメトリクスを対応付けた.

赤字はStep2で重要視されているもの

LLMの評価軸	RAGASのメトリクス
真実性	忠実
	回答の関連性
	コンテキスト精度
	コンテキストの関連性
	回答の意味的類似度
	回答の正確さ
安全性	アスペクト評価
公平性	アスペクト評価
堅牢性	コンテキストリコール
	コンテキストエンティティ
プライバシー	-
機械倫理	-
透明性	-
説明責任	-

プライバシー,機械倫理,透明性,説明責任を評価するメトリクスが不足している.

3. 考察

■RAG構成システムの課題

RAGシステムの完全な自動評価は現状では困難である. しかし, 真実性や堅牢性に関しては評価するメトリクスが複数あるため,一部自動化は可能であると考察する.

またそれぞれの指標を,LLMのモデル,RAGのシステム,RAGの回答精度のどこで評価していくのか整理されていないことが分かった.

今後,不足している観点を何で補っていくのか,どの部分で評価していくのか現場関係者と検討して評価方針を定める必要がある.

4. まとめ

- LLMの評価軸として以下8つの指標に着眼した.
  真実性,安全性,公平性,堅牢性,プライバシー,機械倫理,透明性,説明責任
- ヒアリングをした結果,RAGでは真実性/堅牢性/説明責任が重要とされており,一部のプロジェクトではプライバシーが重要とされていることが分かった.
- RAGASでは真実性,堅牢性はメトリクスがあるが,説明責任やプライバシーを評価することは現時点では難しい.
- RAGシステムの評価の一部自動化は可能であると考察する.
- それぞれの指標を,LLMのモデル,RAGのシステム,RAGの回答精度のどこで評価していくのか整理されていない.
- 今後,不足している観点をどのように補うのか検討する必要がある.

5. 参考文献

1.TRUSTLLM: TRUSTWORTHINESS IN LARGE LANGUAGE MODELS
https://arxiv.org/pdf/2401.05561
2.生成AIをビジネス活用する上で押さえるべき8つの評価観点
https://www.brainpad.co.jp/doors/contents/apply_generative_ai_to_business_tips/
3.RAGAS
https://docs.ragas.io/en/stable/concepts/metrics/index.html#