

生成AIを用いた研究報告書検索システム

Enhancing Research Report Search with Retrieval-Augmented Generation

RAGの仕組みがもつ課題とその対策案の紹介

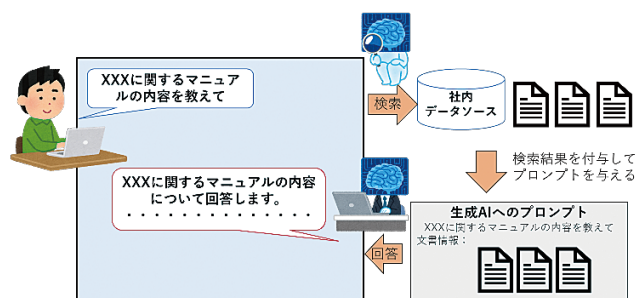
生成AI技術の進化に伴い、非公開の社内情報に関する回答生成を可能とするRAG技術が注目されている。当社でも自社構築によりRAGの検討を進める中で、様々な課題が明らかになってきている。本稿では、検討の中で明らかとなったRAGの仕組み上の課題や、その対策を検討し実装した研究報告書検索システムの内容について紹介する。



執筆者
先端技術応用研究所
情報技術グループ
追良瀬 利也

1 はじめに

近年、ChatGPTを代表とする生成AI技術の急速な発展により、チャット形式でまるで人間と話すかのようにAIと対話できる技術が広がっている。その中でも、RAG (Retrieval-Augmented Generation) と呼ばれる技術が広く使われている。これは、ユーザからの質問に応じて内部的に個別のデータソースを検索し、検索結果と質問内容を一緒に生成AIに送ることで、固有のデータをもとにした回答を可能とさせるものである。RAGのイメージを第1図に示す。



第1図 RAGの概略図

この仕組みにより、通常の生成AIが知識として持っていない、一般には非公開の社内情報等に関して回答させることが出来ることから、こうした取り組みは各社で社内用チャットボットとして活用が広く進んでいる。当社もその1つとして、自社構築で様々な部門のデータを対象に、新人教育や業務検索の高度化などを目的に活用を進めている。研究開発を主なミッションとする技術開発本部においても、過去の研究知見の活用や新たな知見・アイデア創出を効率的に行うため、研究情報がまとめられた研究報告書を対象としたRAGの仕組みを検討している。

検討を進める一方で、基本的なRAGの仕組みが持つ課題や苦手な領域についても明らかになってきている。本稿では、その課題・苦手な領域の一部を紹介しつつ、その対策を検討し試作した研究報告書検索システムについて紹介していく。

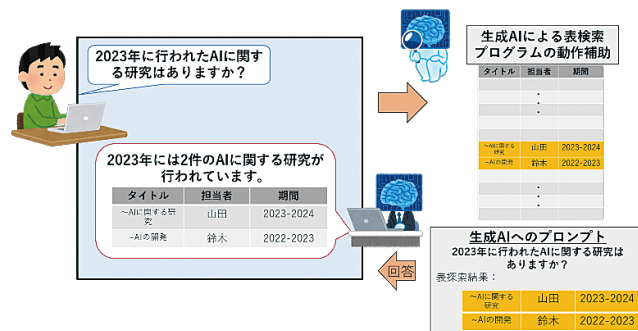
2 RAGの仕組みが持つ課題と対策

本章では、検討を進める中で実際に明らかとなった基本的なRAGの仕組みが持つ課題と、その対策として報告書検索システムに実装した仕組みの概要について紹介する。

(1) 大きな表データの検索と回答

最も大きな課題の1つは、大きな表形式データに関する質問回答である。生成AIはその仕組み上、大きな表形式のデータをそのまま与えて表に関する質問に回答させたり、データを選択させたりする回答生成が難しく、ハルシネーション※を起こす可能性も高まる。実際に、報告書検索システムにおいても、1000行を超える報告書のタイトルや担当者、実施期間などのデータをもつ表データを与え、ユーザの質問（例：2023年に行われたAIの研究はありますか？）に的確に答えさせることは難しく、ハルシネーションも見られた。そこで、ユーザから表データに関する質問が来た際、表から回答に必要なデータを事前に抽出する処理を組合せ、回答を生成させることにした。この処理においても生成AIを用いて、ユーザからの質問をもとに表探索用のプログラムの動作補助のための情報を抽出している。動作の概要は第2図の通り。報告書検索システムでは、表に関する内容や数を問う質問が来た際はこのような動作をシステム内部で行い、回答を行う。

※AIが誤った情報や事実とは異なる情報を生成する現象。

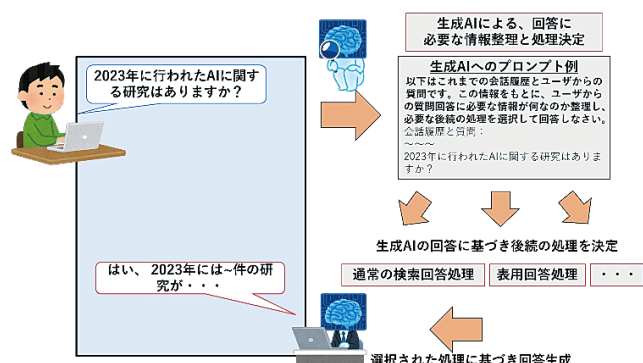


第2図 表探索の補助・生成を行う仕組みの概要

(2) 都度の情報探索による弊害と処理分岐

基本的なRAGは、第1図のような質問→検索→回答の流

れを毎回繰り返すことでユーザとの対話を行う。つまり、質問と一緒にプロンプトとして生成AIに与える検索結果の情報の品質が、回答の質と密接に関係することとなる。こうした仕組み上、ユーザとの会話の中で暗黙的に示される「その研究」や「この報告書」といった文脈依存の情報については、必要な内容を検索・回答させることが難しい場合がある。また、(1)で追加した表に関する回答用の処理が加わったことで、ユーザからの質問への回答に、どのような情報が必要なのかを一度判断して後続の処理を適切に選択する必要がある。そこで、この処理に生成AIを用いた。処理の概要を第3図に示す。質問から回答までの処理の流れは以下の通りである。



第3図 ユーザからの質問に基づき後続処理を選択する仕組みの概要

(ア) ユーザからの質問

ユーザからのチャット形式の質問を受け付ける。

(イ) 生成AI(エージェント)による情報処理

会話の流れと(ア)の質問をもとに、ユーザへの回答に必要な情報が何なのかを生成AIに回答させる。

(ウ) 生成AIの回答をもとにした処理分岐

(イ)の回答結果をもとに、どの後続の処理が必要か、処理を分岐する。例えば生成AIが表データによる回答が必要と回答したら表探索・回答を行う後続処理を、通常のデータソースからの検索・回答が必要と回答したら、その後続処理を行うように、処理プログラムを分岐する。

(エ) 選択された後続処理に基づく生成AI回答

(ウ)で選択された処理に基づき、情報を得た後生成AIによる回答生成を行う。

このような処理を内部的に行い、質問-回答を繰り返すことでユーザからは特に意識することなく、必要な情報をもとに回答してくれるRAGシステムが実現できる。

た回答を行っている。また、第5図では、第4図の会話に継続して、報告書の内容を問う質問を行い、データソースの検索を経て生成AIが回答している。いずれの処理も、会話履歴とユーザからの質問をもとに、生成AIが回答に必要な情報を整理し、内部的に処理を分岐した上で回答処理がなされていることが確認できる。

2020年以降の追良瀬さんの研究を教えてください

2020年以降の追良瀬さんの研究は以下の通りです。

No.	研究名	期間	書籍
1		2020/04-2021/03	本店先端技術応用研情報通信G 追良瀬 利也
2		2021/06-2022/03	本店先端技術応用研情報通信G 追良瀬 利也
3		2022/04-2023/03	本店先端技術応用研情報通信G 追良瀬 利也
4		2021/04-2022/03	本店先端技術応用研情報通信G 追良瀬 利也
5		2020/07-2021/03	本店先端技術応用研情報通信G 追良瀬 利也
6		2021/06-2022/03	本店先端技術応用研情報通信G 追良瀬 利也
7		2022/04-2023/03	本店先端技術応用研情報通信G 追良瀬 利也
8		2020/04-2021/03	本店先端技術応用研情報通信G 追良瀬 利也

引用： 1. 研究件名リスト.pdf

第4図 報告書検索システム画面その1
(社内情報につき部分的に黒塗り処理)

No.1の研究について概要を教えてください

追良瀬さんの「XXXXXXXXXX」についての概要は以下の通りです。

1. 研究目的

XXXXXXXXXX

2. 研究結果

- 検証手段の検討 XXXXXXXXXX
- フィールド検証 XXXXXXXXXX
- 現場適用に向けた基幹仕様決定 XXXXXXXXXX

引用： XXXXXXXXXX

第5図 報告書検索システム画面その2
(社内情報につき部分的に黒塗り処理)

4 おわりに

本稿では、近年活用が進むRAGがもつ課題や苦手な領域の一部について、その対策を検討し試作した研究報告書検索システムについて紹介した。今回紹介した以外にも、RAG技術はまだまだ発展途上であり、さらに回答精度を高めるため、検索対象のデータ保持・生成AIへの提供方法の改善[1]など様々な研究開発が盛んに行われている。今後も同技術を中心に最新技術をキャッチアップし検証を進め、社内情報探索の高度化や生成AIによる更なる業務効率化に貢献していきたい。

5 参考文献

1. Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., & Zhao, L. (2024). GRAG: Graph Retrieval-Augmented Generation. arXiv preprint arXiv:2405.16506.

3 研究報告書検索システム画面

本章では実際の報告書検索システムの画面を紹介する。第4図では、ユーザからの質問がリストからの回答を問うものであることから、2章で紹介した表探索処理を組合せ