最先端ソフトウェア工学ゼミ［個別ゼミ2］
成果報告
RAG技術の最新動向調査

2024年12月5日

家村 康佑（富士通）
西山 美恵子（クレスコ）

発表内容

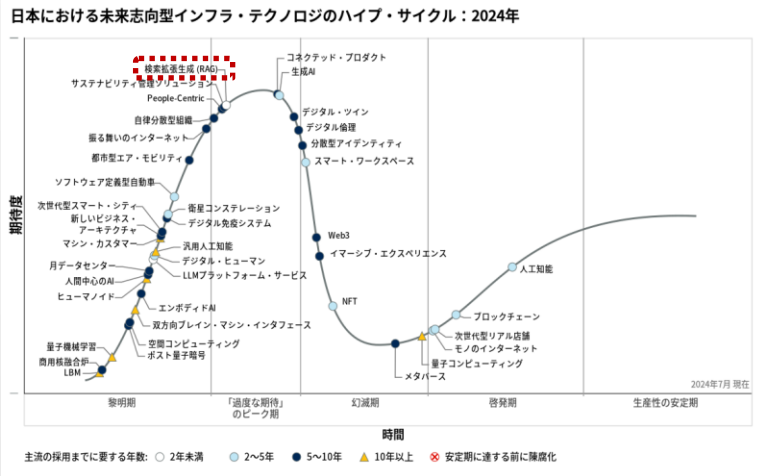
1. 設定テーマと選定理由
2. 初期調査
3. 追加調査内容
4. 調査結果
5. 考察
6. 所感
7. まとめ

1. 設定テーマと理由

- テーマ
  - RAG技術の最新動向調査
- 理由
  - 業務においてもRAG（Retrieval-Augmented Generation）技術を活用する機会が増え、RAGの可能性と課題が明確になりつつある。
  - RAGの技術自体は発展途上であり、最新動向を調査することでの議論の検討を実施したい。
  - 具体的なユースケースごとのRAGの適用技術マップを作成し、今後の業務に活用したい。

1-1. 調査背景

- 検索拡張生成（RAG）の技術動向
  - 大規模言語モデルの知識を外部データソースに拡張する技術として注目を集めている。
  - Gartnerが2024年8月に発行したハイプサイクルにおいても、RAGが新規テクノロジとして追加
  - データベースの検索技術やハルシネーションの抑制、多様なデータ形式の対応など技術開発や研究が活発



https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20240807-future-oriented-infra-tech-hc

1-2. 調査方法

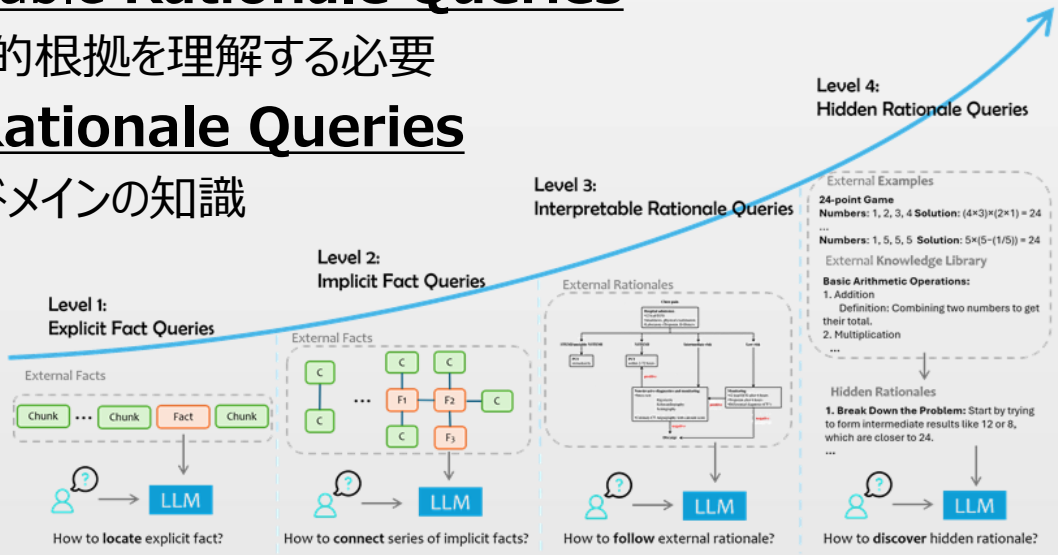
- RAG技術のSurvey論文を中心に調査
  - 論文①　：Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely 2024.09, Microsoft Research
  - 論文②　：Trustworthiness in Retrieval-Augmented Generation Systems: A Survey, 2024.09, Beijing Academy
  - 論文③　：Retrieval-Augmented Generation for Large Language Models: A Survey, 2023.12, Shanghai Research
  - 論文④　：Retrieval-Augmented Generation for AI-Generated Content: A Survey, 2024.02, Peking University
  - Web　：https://github.com/langchain-ai/rag-from-scratch
- RAG技術の最新論文を抽出
  - 論文⑤　：Golden-Retriever: High-Fidelity Agentic Retrieval Augmented Generation for Industrial Knowledge Base
  - 論文⑥　：Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach

2-1. 初期調査：調査結果（論文①）

- RAGを適用する際のQuery のタイプごとに、技術的な難易度で階層化、Levelごとの必要な技術について整理
  - Level 1 ：Explicit **Fact Queries**
    - 単一のドキュメント内にテキストとして存在
  - Level 2 ：Implicit **Fact Queries**
    - 複数のドキュメント内に分散して存在

RAGとしての評価データセットも様々存在

  - Level 3 ：Interpretable **Rationale Queries**
    - ドメイン固有の理論的根拠を理解する必要
  - Level 4 ：Hidden **Rationale Queries**
    - 暗黙的に存在するドメインの知識



2-1-1. Query Examples

- **Level1： Explicit Fact Queries**
  - 例：「2024 年夏季オリンピックはどこで開催されますか?」
- **Level2： Implicit Fact Queries**
  - 例：「キャンベラが位置する国の現在、多数党は何か」
- **Level3： Interpretable Rationale Queries**
  - 例：「胸痛と特定の症状がある患者はどのように診断および治療されるべきですか」
- **Level4： Hidden Rationale Queries**
  - 例： 5、5、5、1の数字を使って24ポイントを達成するには? 24ポイントゲーム

2-1-2. Level 1での要素技術

- マルチモーダル（表）
  - Table-to-Text
  - 画像のテキスト化、属性か
  - 埋め込み技術
- チャンク最適化
- データ検索
  - Sparse 検索：tf-idf, knn, keyword
  - Dense 検索：ベクトル
  - その他：ハイブリッド検索
- Query Doc Alignment：クエリ書き換え、HyDE
- Rerank：検索結果の修正（有用性、セマンティック）
- 再帰的検索：質問の段階的に分解して検索等

2-1-3. Level２での要素技術

元のクエリを複数の取得操作に分解し、結果を包括的な回答に集約する。
ドメイン固有の専門知識を必要とせずに常識的な推論を含む。
この種類のクエリには、統計クエリ、記述的分析クエリ、基本的な集計クエリ。たとえば、カウント、比較、傾向分析、選択的要約などの操作は、"いくつ" や "最も多い" タイプのクエリ

- Interactive RAG
  - プランニングベース：ReACT、RAT（CoT)、GenGround
  - 情報ギャップ埋めベース：ITRG、Self-RAG
- Graph/Tree RAG
- SQL RAG

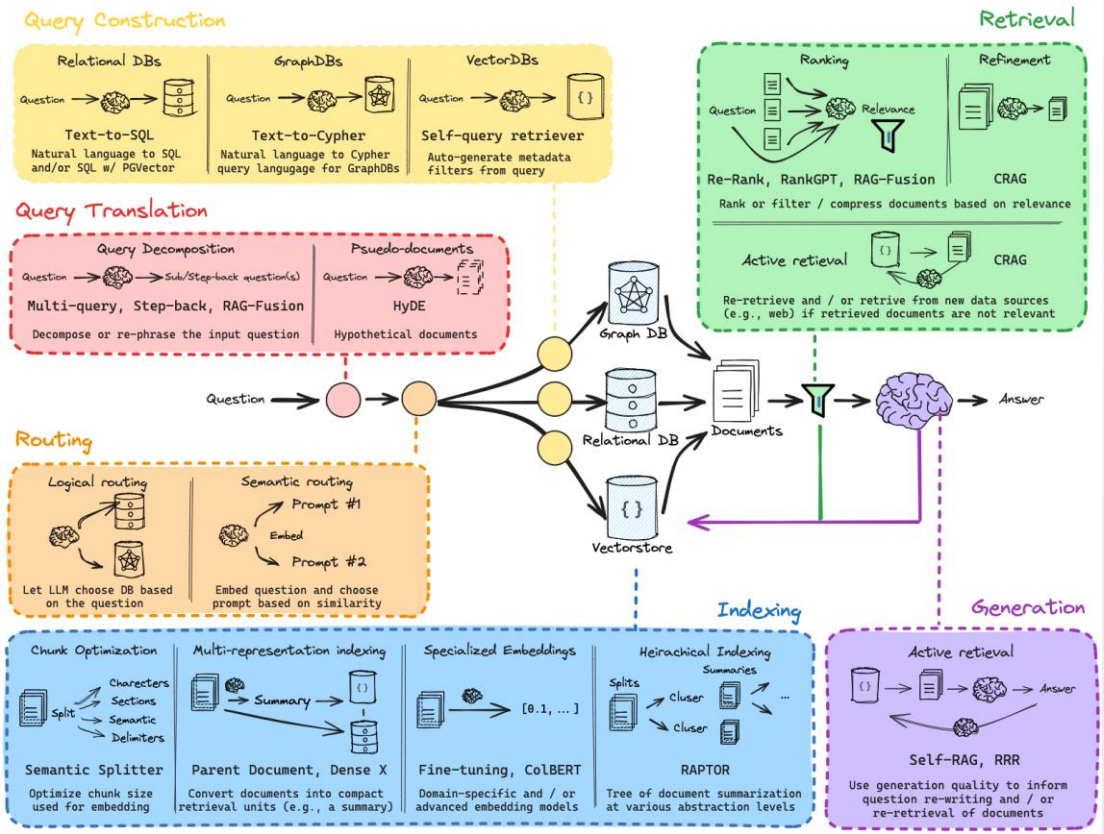
Query sample

- サンプルサイズが1000を超える実験はいくつありますか?
- 最も頻繁に言及される上位3つの症状は何ですか?
- X社とY社のAI戦略の違いは何ですか?

2-2. 初期調査：調査結果（Web）

- RAGを適用する際に活用できる要素技術を網羅的に紹介（2024年初期）

RAG-要素技術	要素技術の実現アプローチ
Query Translation（クエリ変換）	Input question translation
Routing（データベース選択）	Logical routing
	Semantic routing
Query Construction（クエリ構造）	Relational DBs
	Graph DBs
	Vector DBs
Indexing（データベース構成）	Chunk Optimization
	Multi-representation indexing
	Specialized Embeddings
	Heirachical Indexing
Retrieval（検索）	Ranking
	Refinement
	Active retrieval
Generation（生成）	Active retrieval



https://github.com/langchain-ai/rag-from-scratch

3. 初期調査を踏まえた追加調査

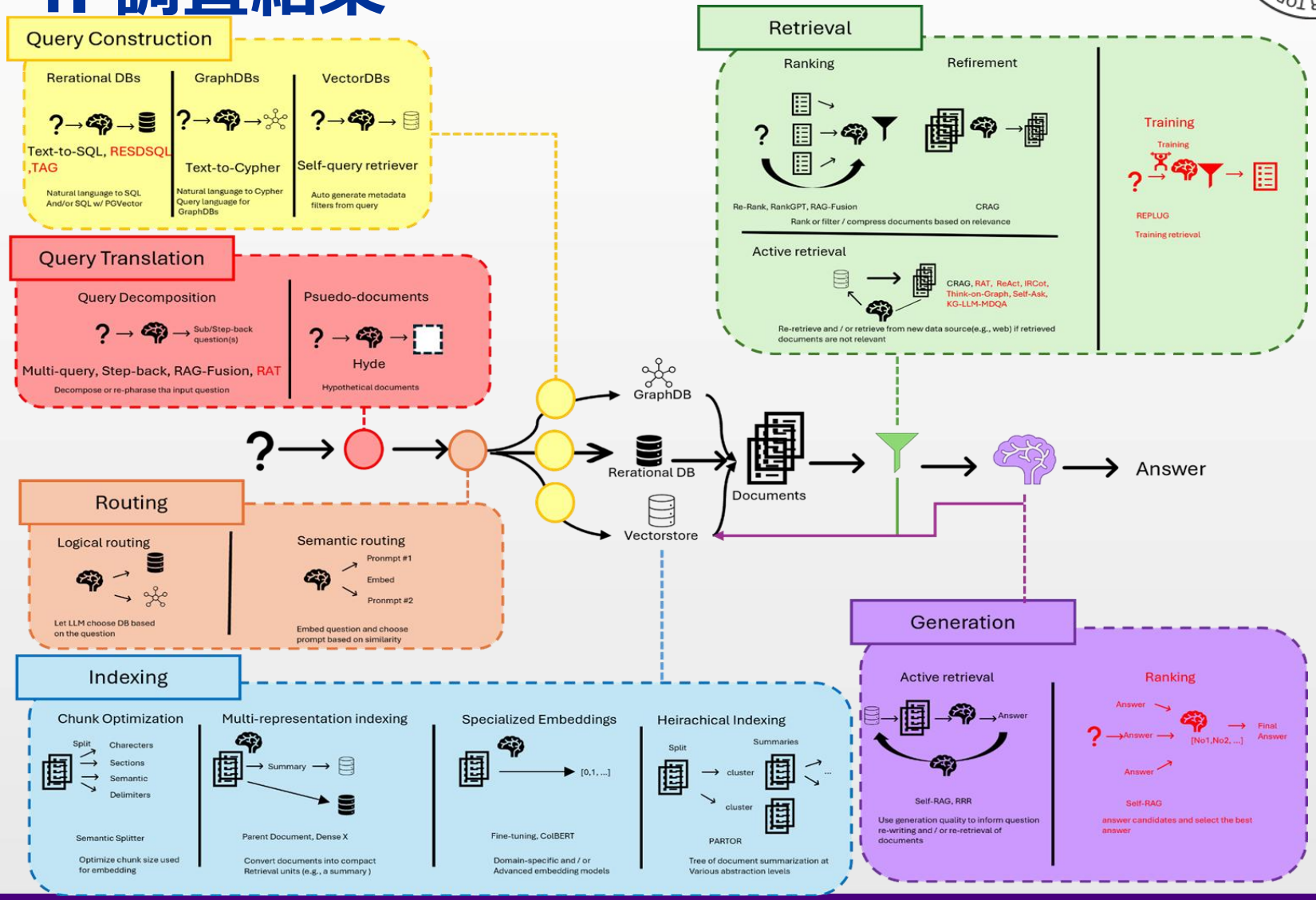
- RQ：2024年下期時点の技術動向はどのように変化しているか
  - 調査内容

    Survey 論文からLevel2に該当する論文を中心に追加調査しマッピング
  - 調査対象論文

    論文①　：Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely 2024.09, Microsoft Research

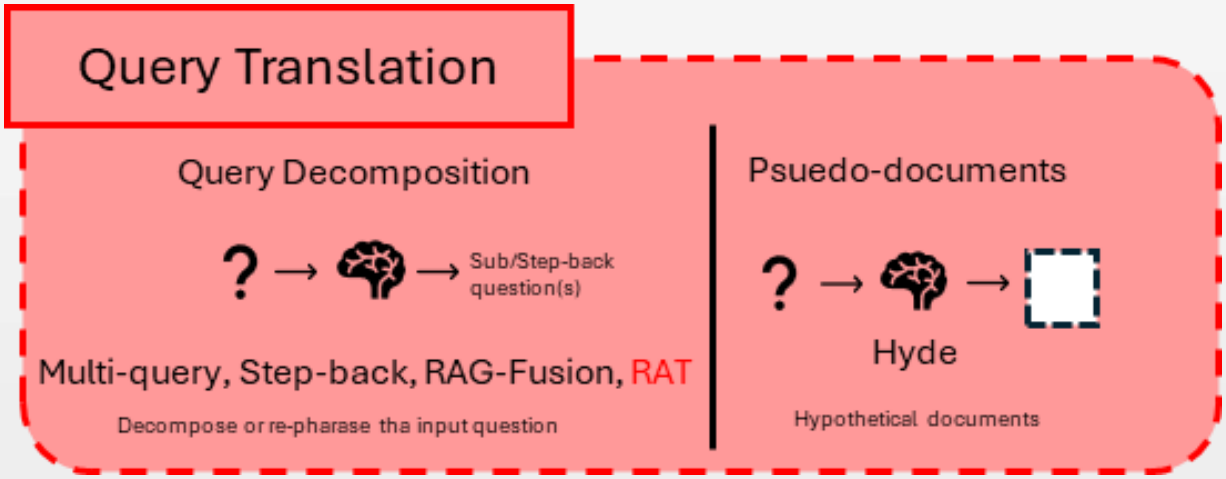
    論文②　：Trustworthiness in Retrieval-Augmented Generation Systems: A Survey, 2024.09, Beijing Academy

4. 調査結果



Query Transrationとは

入力された質問文をより適切に探索できるように、
質問文を調整する手法のこと。



4. 調査結果

RAT

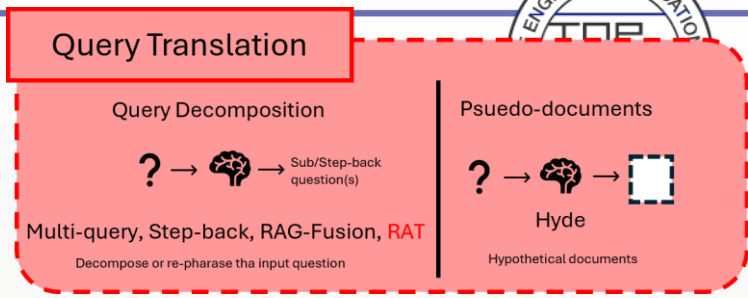
- Query Translation

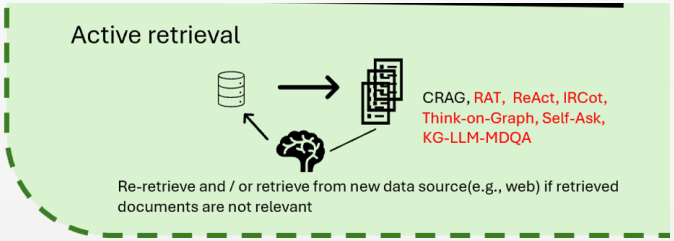
  Zero-shot CoT を使って、抽象的な質問を具体的な質問リストに変換

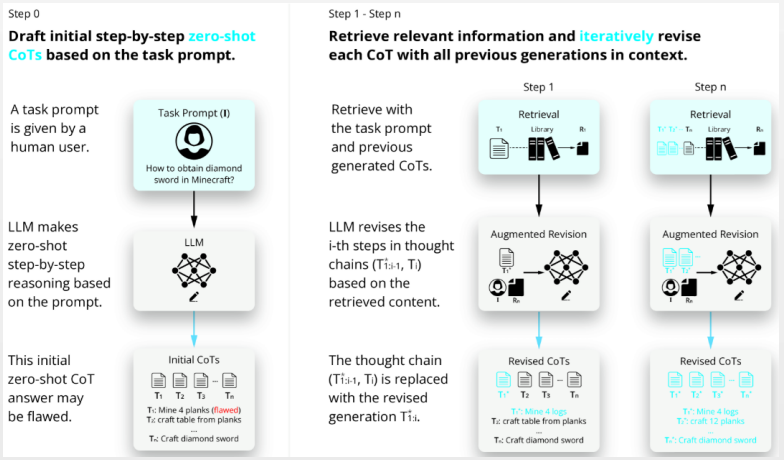
- Active Retrieval

  単なるクエリ分割にとどまらず、実行中も前の検索結果に応じてクエリを作成し、最終的な回答を出力

https://arxiv.org/html/2403.05313v1
(2024年)

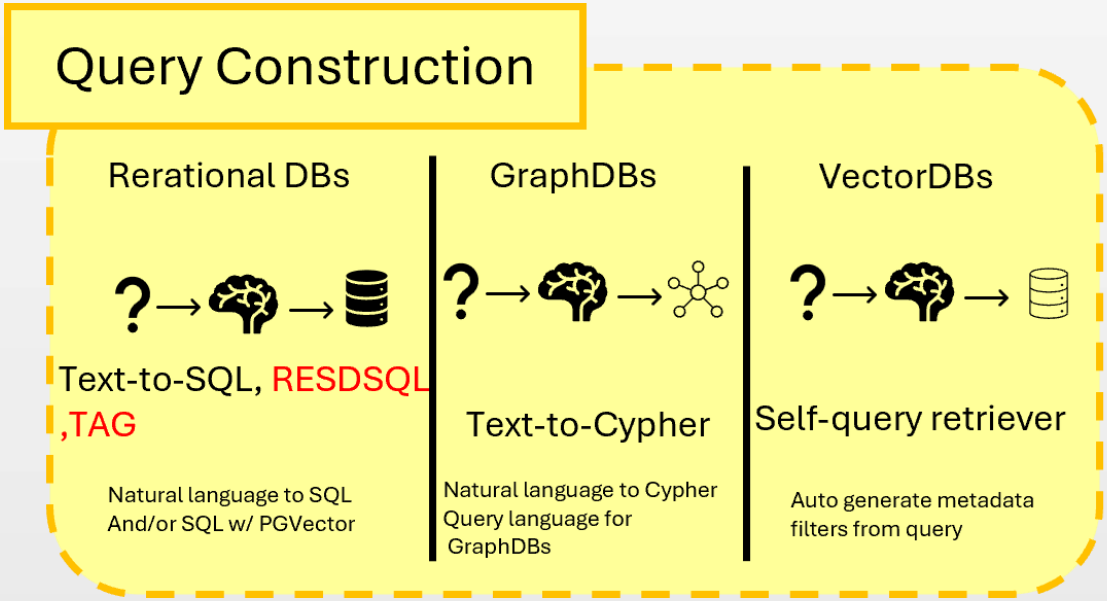




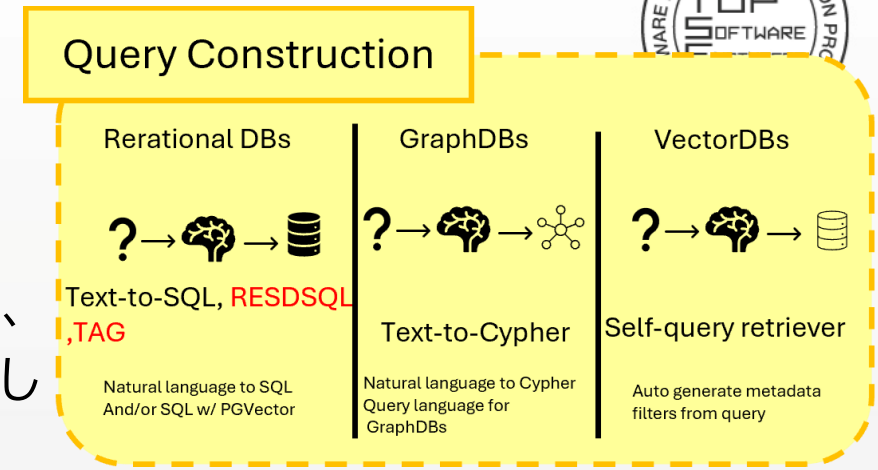


Query Constractionとは

曖昧な質問から適切に情報を探索するために、質問内容から構造化されたクエリを作成する手法



4. 調査結果



- **RESDSQL**

  DBの Schema 情報を単一文に変換し、質問と一緒に投入してSQLクエリを生成し探索

  https://arxiv.org/abs/2302.05965 (2023年)

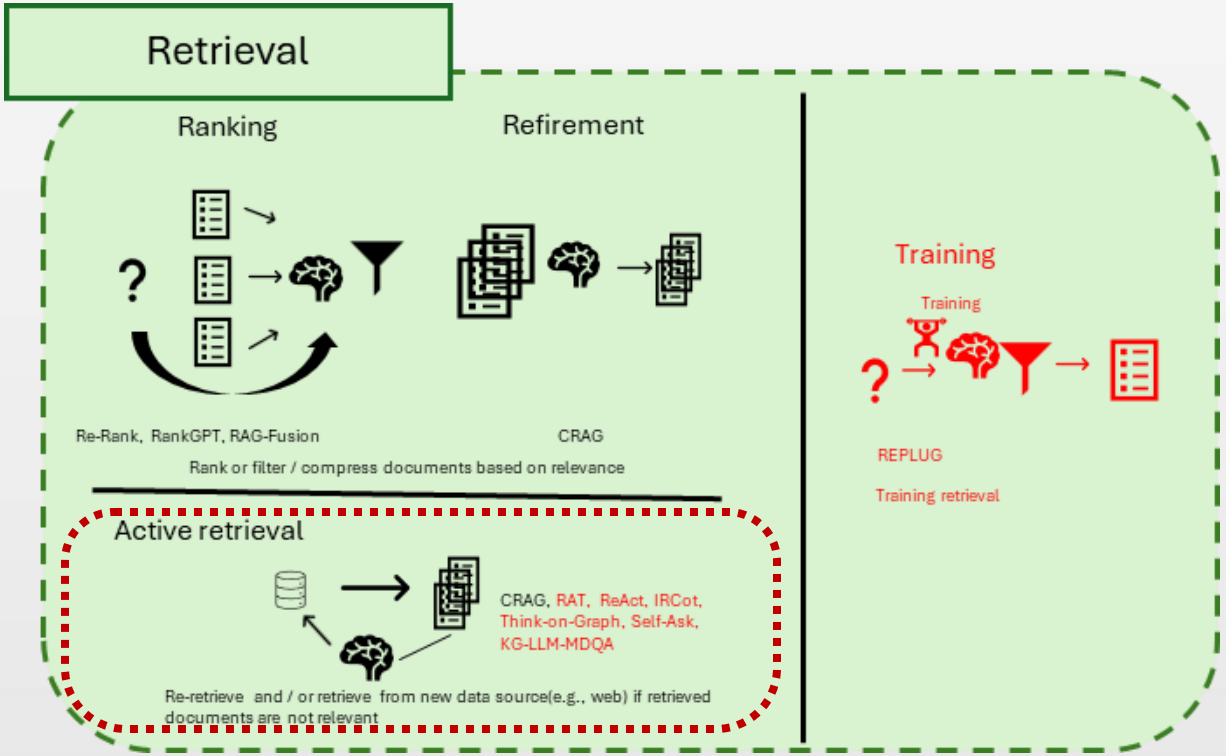
- **TAG**

  質問からSQLを生成し、SQL実行後のTable出力をLLMで判断
  検索結果をDocumentではなくTableとして処理

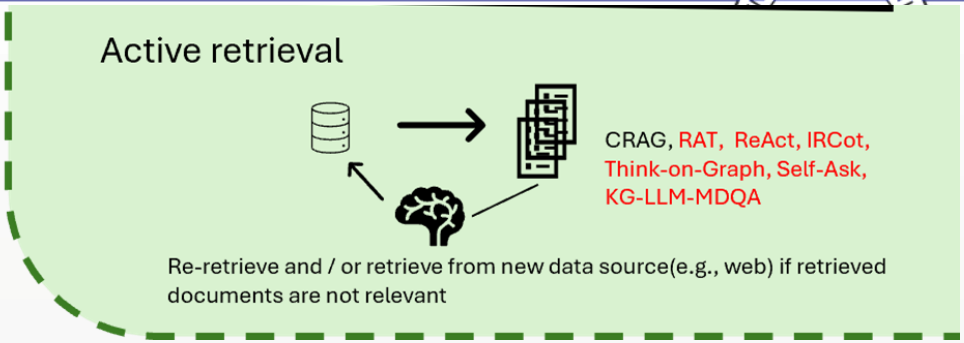
  https://arxiv.org/abs/2408.14717v1 (2024年)

Active Retrievalとは

Retrievalの技術の一つで、曖昧な質問から、データを探索し、取得した文書が適切でない場合、次の探索に必要なInputを得て、さらなる探索をする手法



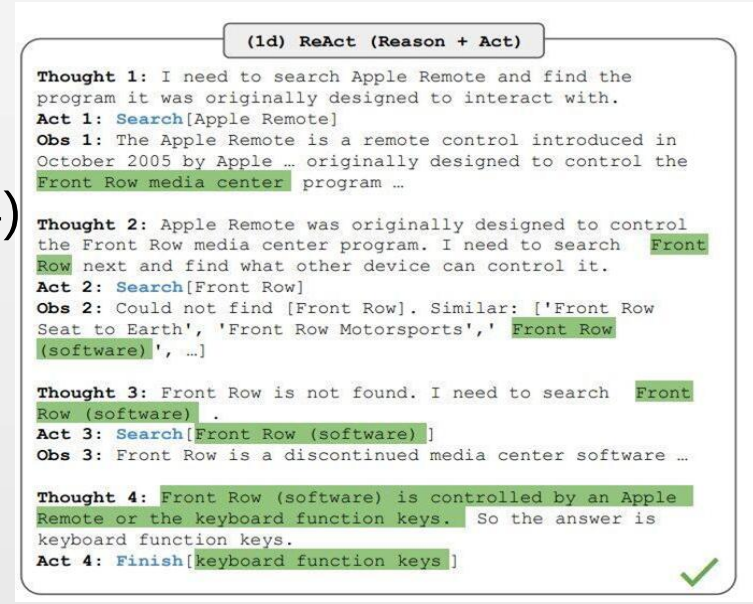
4. 調査結果



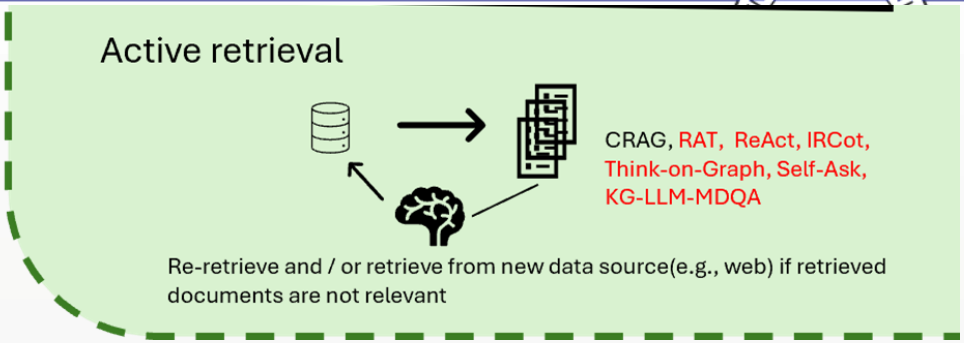
ReACT

- 次の質問の立て方
  前の検索結果から読み取れる内容(Observation)をLLMに抽出させ、
  次に必要な内容をLLMに考えさせ(Thought)、次の検索を実行(Act)
- 最終回答
  質問に対する答えをLLMが生成したと判断した場合に最終回答として出力

https://arxiv.org/abs/2210.03629 (2022年)



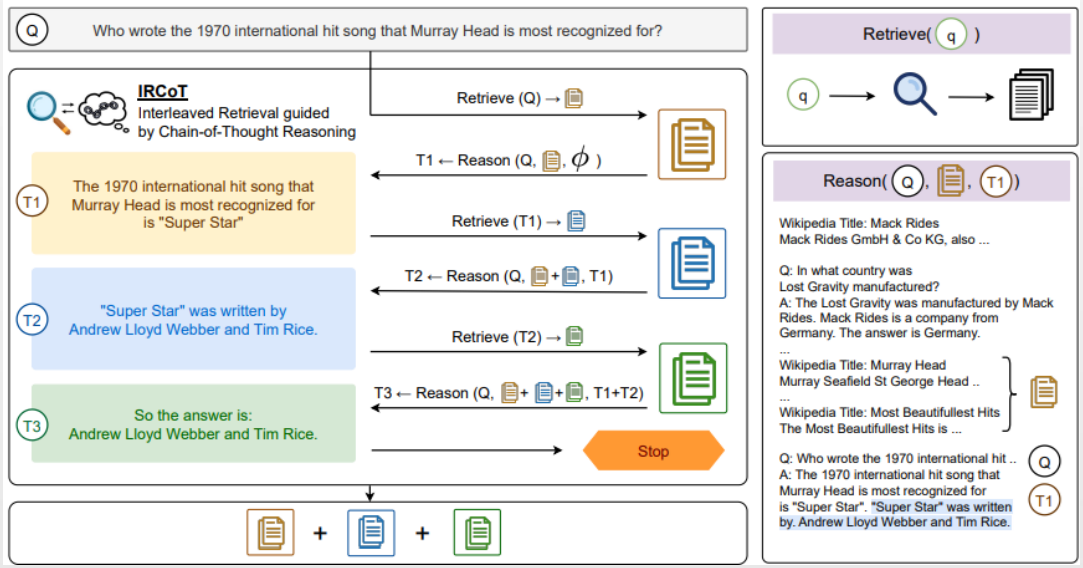
4. 調査結果



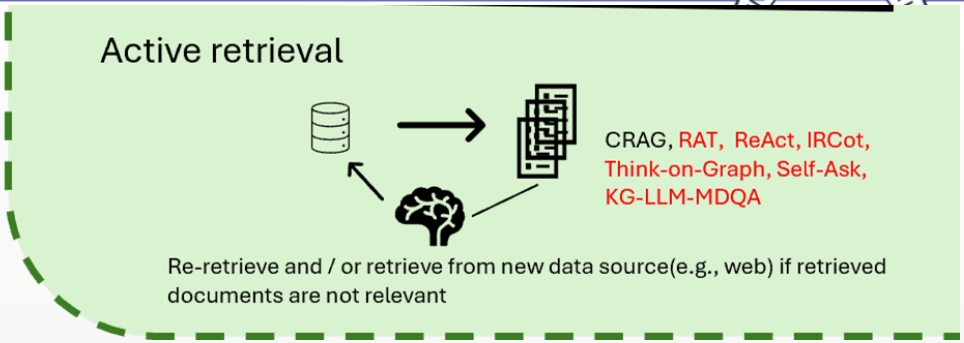
IRCoT

- 次の質問の立て方
  前の検索結果を用いた出力結果を用いて、次の検索を実施
- 最終回答
  質問に対する答えをLLMが生成したと判断した場合に最終回答として出力

https://arxiv.org/abs/2212.10509(2022年)



4. 調査結果



Think-on-Graph

- 次の質問の立て方

  Knowledge Graphの接続から段階的に深いNodeを探索

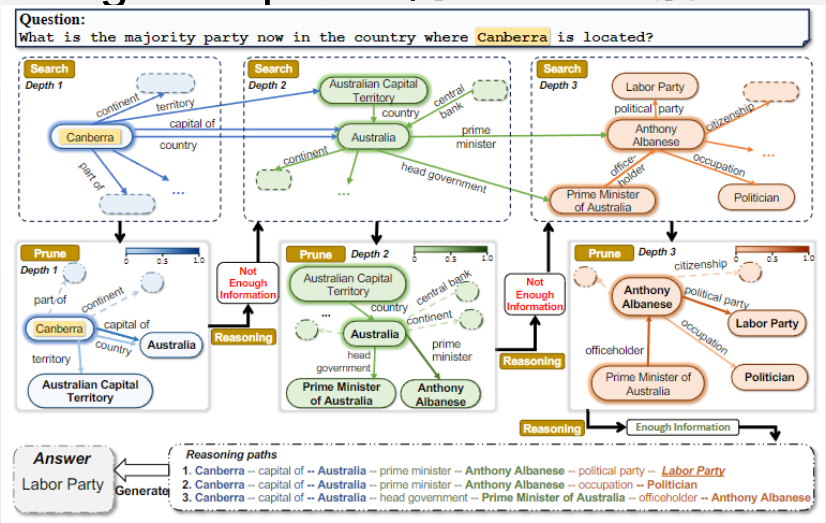
- 最終回答

  質問に対する答えをLLMが生成したと判断した場合に最終回答として出力

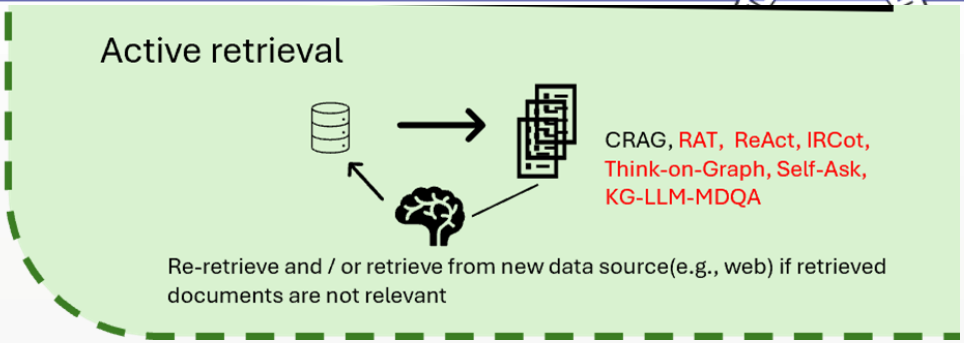
- LLMの知識の利用

  探索をした結果、最終的な回答をKnowledge Graphから得られない場合、LLMの知識で補完して回答

https://arxiv.org/abs/2307.07697
(2023年)



4. 調査結果



Self-Ask

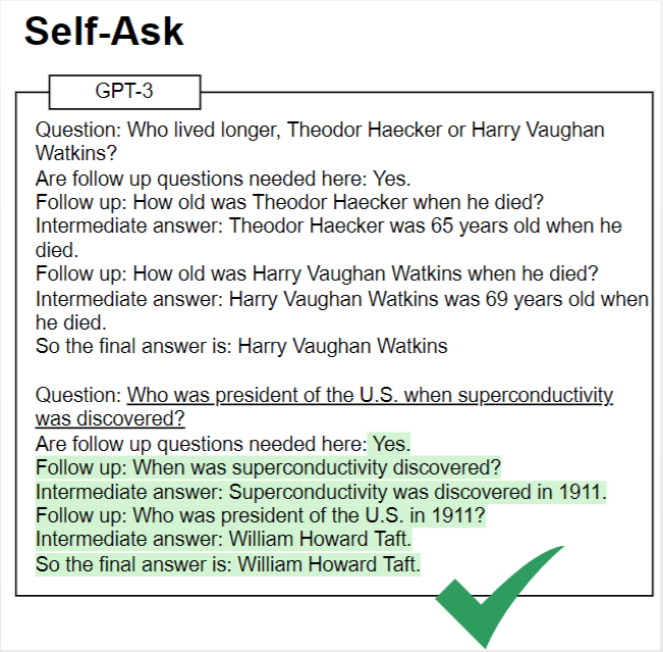
- 次の質問の立て方

  LLMでFollow Upの質問を生成、回答に必要となる追加情報を抽出

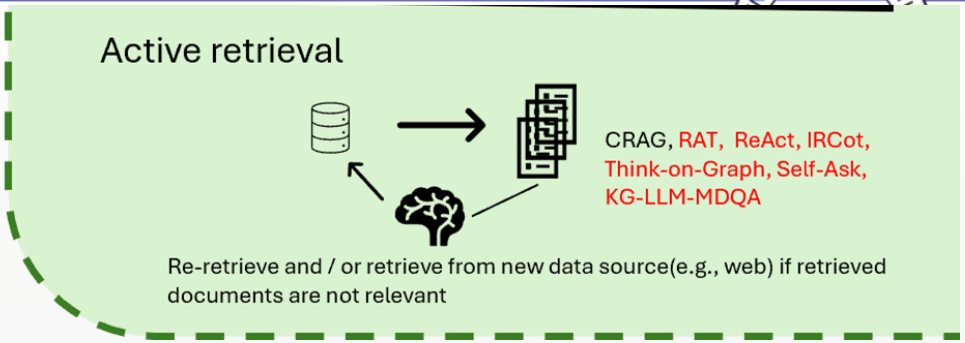
- 最終回答

  質問に対する答えをLLMが生成したと判断した場合に最終回答として出力

https://arxiv.org/abs/2210.03350 (2022年)



4. 調査結果



KG-LLM-MDQA

- 次の質問の立て方

  QuestionからLLMに推論させた結果と、Knowledge Graphの情報を比較、より類似性の高いノードから、次の検索条件を決定する。

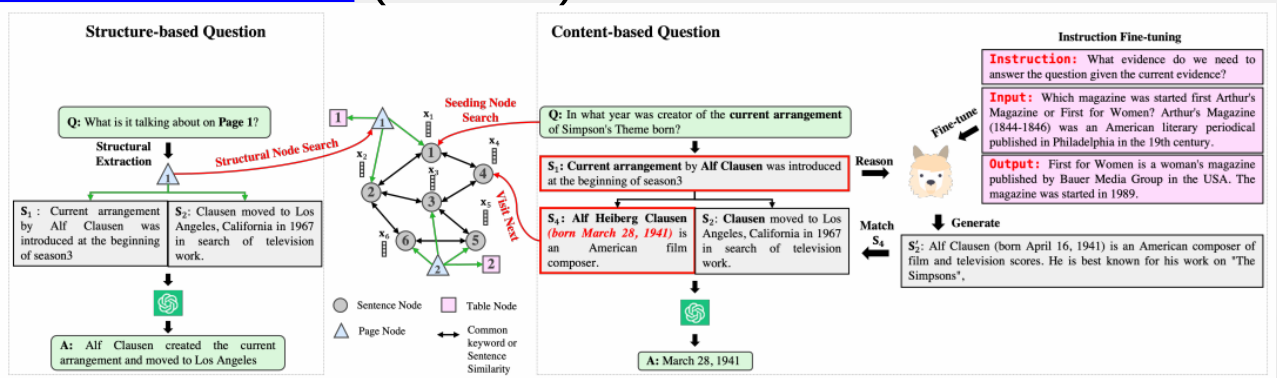
- 最終回答

  グラフの探索がない場合、最終的な回答を出力する。

- LLMの知識

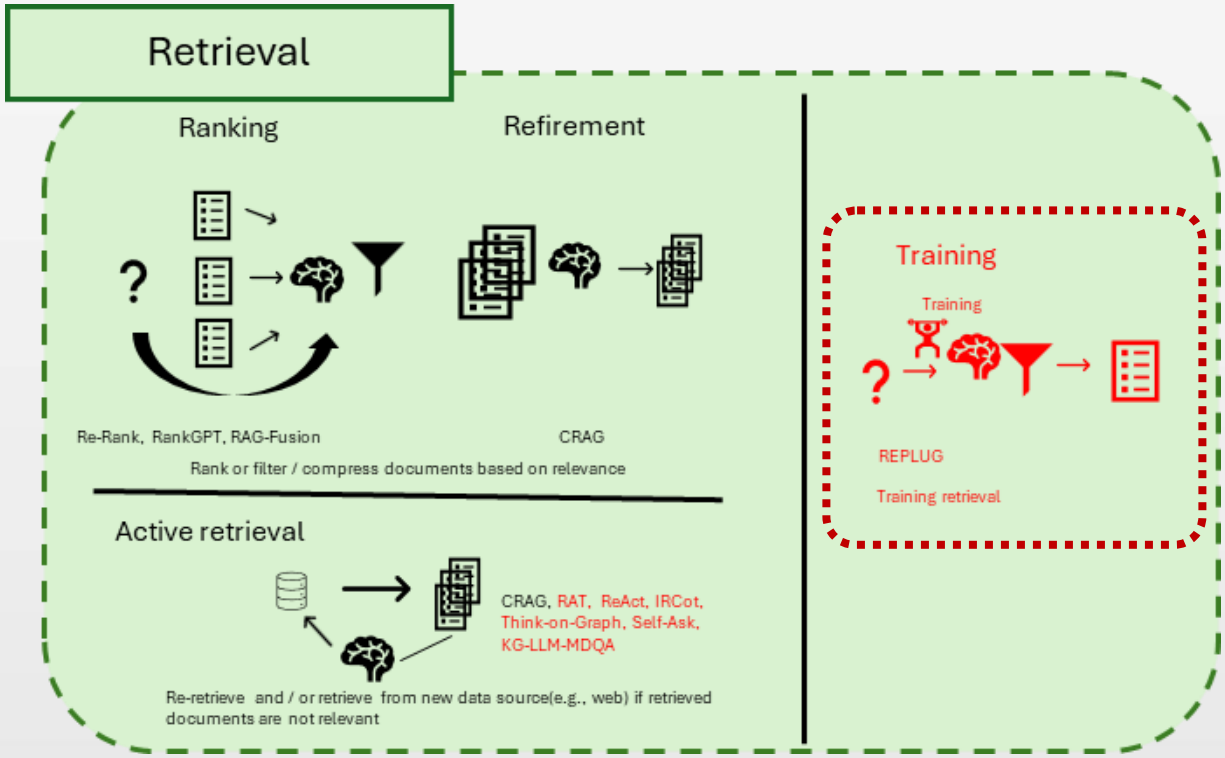
  探索の方向性を決めるために利用

  https://arxiv.org/abs/2308.11730 (2023年)

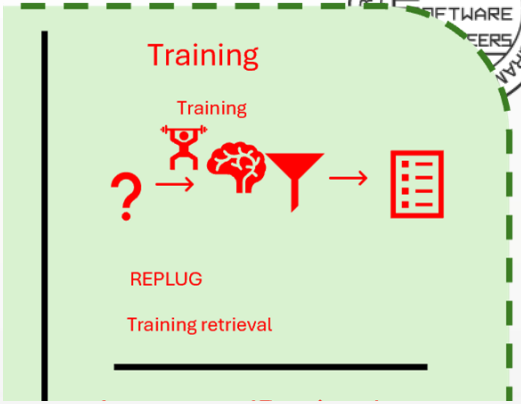


Trainingとは

LLMのパラメータ調整などをして調整するのではなく、探索する
Retriever自体をトレーニングする手法



4. 調査結果



REPLUG

- 学習方法（図1）
  LMの回答が正しくなるようにRetriever自体を学習
- 回答方法（図2）
  決まったルールで複数LMの出力結果から最もらしい、回答を答えとして出力
  https://arxiv.org/abs/2301.12652 (2023年)

図1：学習方法

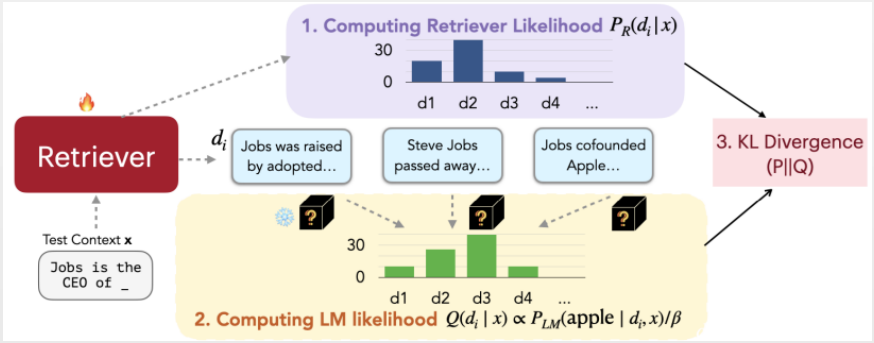
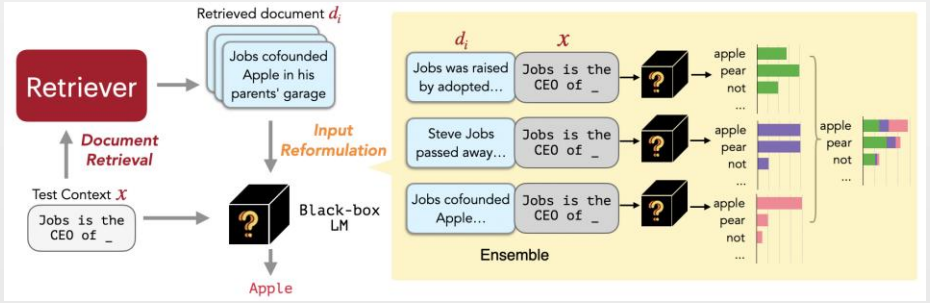
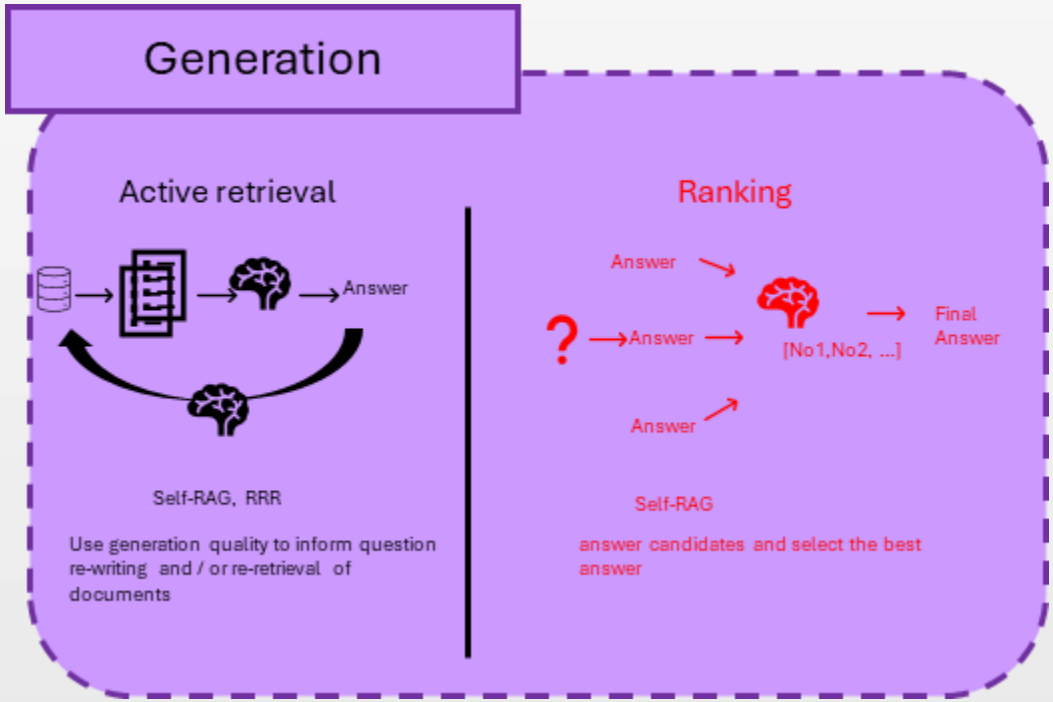


図2：回答方法



Rankingとは

LLMから複数生成された回答を、評価することで
より精度の高い回答を得る手法

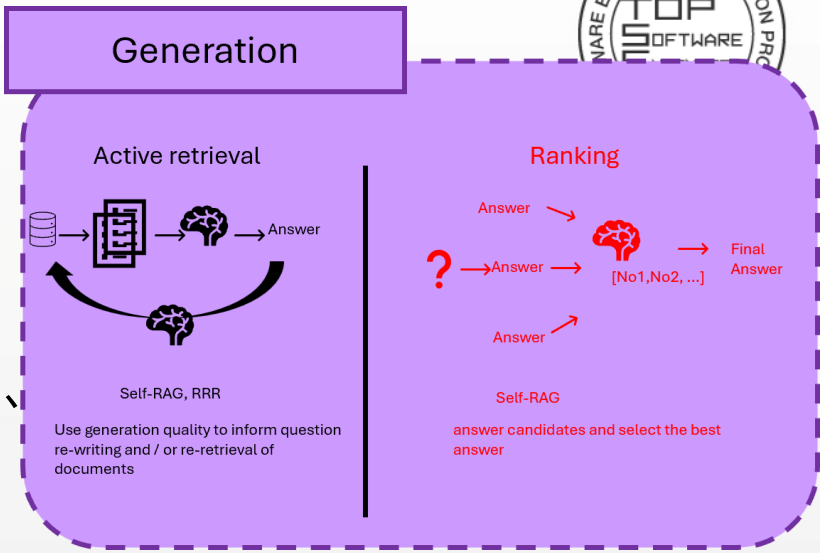


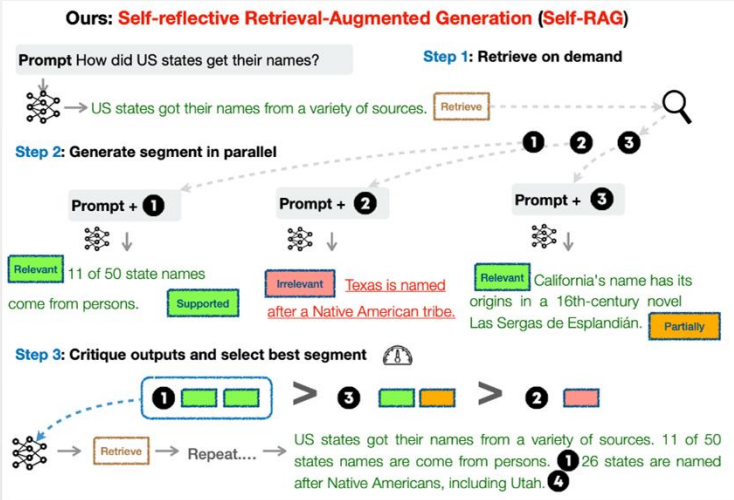
4. 調査結果

Self-RAG

- Documentの評価
  ドキュメント毎に、LLMに生成された回答を、決まったルールでランキング
- 最終回答
  最も順位が高いものを最終的な回答として出力する。

https://arxiv.org/abs/2310.11511 （2023年）





4. 調査結果（まとめ）

技術	要素技術	サンプルクエリ・適用例・概要
RAT	Query Decomposition Active Retrieval	「野口英世の生涯について説明して（Step by Stepで考えて）」 Stepに分解して検索。1:野口英世は・・・、2:若い時は・・・、
RESDSQL	Rerational DBs	SQLのDB Schema + Query ➡ 結果取得するSQL文生成
TAG	Rerational DBs	集計結果のテーブル＋Query ➡ 表の理解をLLMで実施
ReACT	Active Retrieval	「コロラド造山帯の東部地域の標高の範囲はどれくらいですか?」 思考・行動・観察 の繰り返し
IRCoT	Active Retrieval	「ロストグラビティはどこの国で作られたんですか?」 1:ロストグラビティはMack Ride、2:Mack Rideはドイツ
Think-on-Graph	Active Retrieval	「キャンベラがある国の多数党は？」 1:キャンベラはオーストラリア、2:オーストラリアの多数党はｘｘ
Self-Ask	Active Retrieval	「・・・の時のアメリカ大統領は？（深堀が必要なら何について知る必要があるか教えて）」
KG-LLM-MDQA	Active Retrieval	文書間のつながりをKnowledge Graphで効率的に表現、探索
REPLUG	Training	多くの文書を参照して、最も尤度の高い回答を選択 検索結果を学習
Self-RAG	Ranking	「アメリカ州の名前の由来は？」 似た文書を効率的に除外「Texasはアメリカ先住民族由来」等

5. 考察

今回調査したLevel2では、以下3点の技術活用が多い

■ Active Retrieval

今回の調査では**最も多い技術**、RAGでは検索結果に出力されることが前提のため、曖昧な質問や探索のさらなる深化を行い、検索結果の精度上げる技術として活用されていると考えている。

同様に Query Transration の技術も並行して用いられている。

■ Knowledge Graph

事前に知識の相互関係を定義しておくことで、通常の検索では検索できない文書間の知識を定義するGraphDBの技術の増加

■ 生成された回答の評価

生成された回答を自己評価して、精度の高い回答を得る技術として活用されている。

6. 所感

- 技術マップが整理されたことで、RAGシステムの中でも既存システムで精度向上のために取り組めそうな技術領域が明確になった。
- 社内にも当資料を展開して、RAGシステムの精度向上の参考にしていきたい。
- RAG自体は様々な技術的な展開がまだまだ可能な領域であり、実業務に合わせた工夫が必要

7. 今後の課題

- Level2以降の技術等についての整理
- RAG技術と業務課題との紐づけ
  1. RAG技術ごとの適用可能なアプリケーションの検証
  2. RAGに適用しやすいデータ構造の検証