RAGの検索ノイズ抑制のためのファインチューニング

SCSK株式会社
富田勇希、大塚真子、岡本怜奈、濱松大介、中本裕大

2026/3/13(金)

取り組み背景・目的

取り組みの背景

- 近年、RAGでは検索手法の発展により、回答に必要な情報を含むチャンクを高精度に取得できるようになっている
- 一方で、検索結果には無関係または紛らわしい文書が混在することもあり、誤答や、関連文書が得られない場合のハルシネーションが課題となっている
- RAGの高度化手法のうち、埋め込みモデルや検索アルゴリズムに依存しない、検索結果からの回答の生成精度向上に着目した

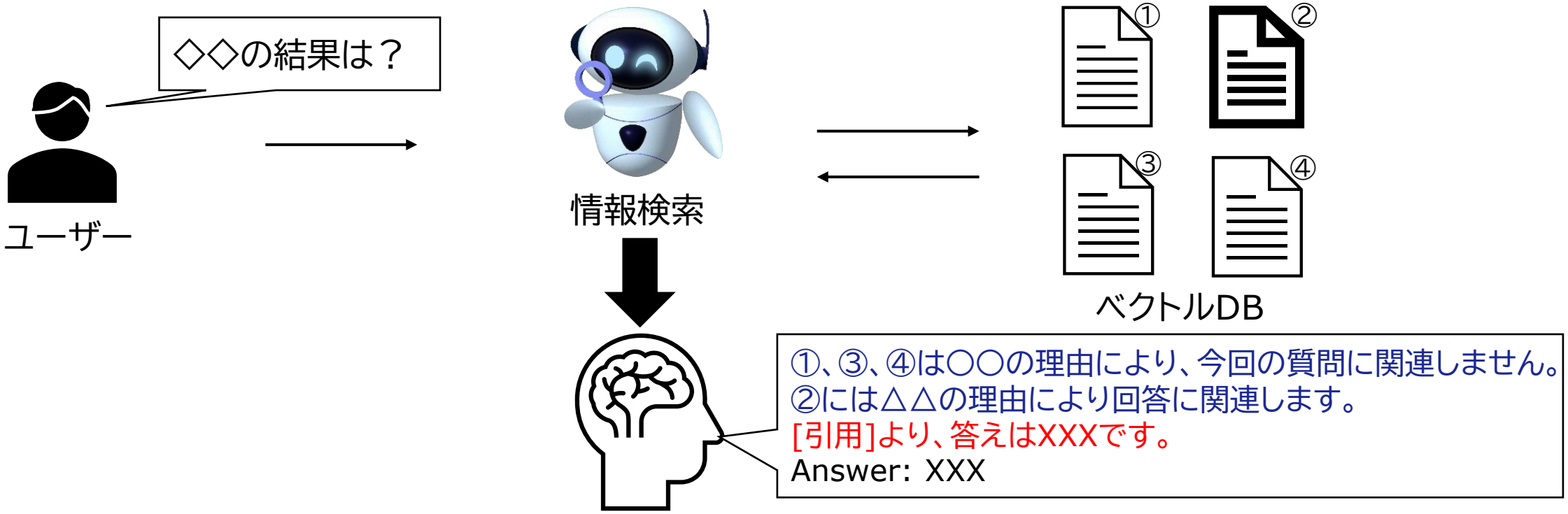
目的

- チャンク内の必要・不要な情報を適切に判断し、ノイズに影響されることなく、ユーザーからの質問に対し正しく回答できるモデルの開発を目指す

実施内容

4-8b-decay2m-ipt_v3.1-instruct4をファインチューニング
検索結果から必要な情報を引用する思考を2パターンで学習させる

- CoT①:関連箇所の引用と理由の説明を行う思考を学習
- CoT②:CoT①に加え、全チャンクの関連性判断、不要情報の排除を行う思考を学習

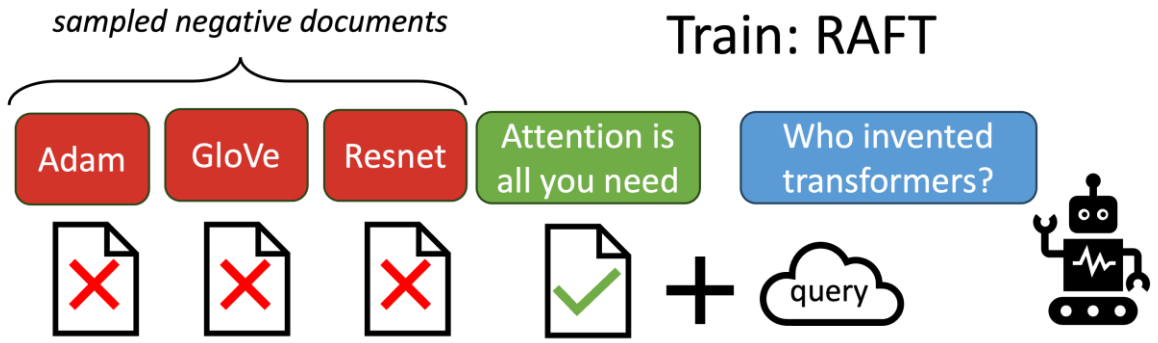


ファインチューニング済みモデル(CoT②)

既存手法

RAFT: Adapting Language Model to Domain Specific RAG[Zhang +24]

検索で得た複数のチャンクの中から関連する情報だけを選別し、無関係な「ノイズ文書」を無視しながら回答を生成できるよう、意図的にノイズを含むデータでファインチューニングする手法
※CoT①はRAFTを参考にして作成



Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models[Yu +24]

検索で得た複数のチャンクに対してLLMに逐次的な「読解メモ」を生成させることで、情報の関連性を段階的に評価し、ノイズや無関係な文書に惑わされず最終的な回答の信頼性を高める手法
※CoT②はChain-of-Noteを参考にして作成

学習データ

QAデータを生成し、検索結果から関連情報を引用しつつ段階的に推論する回答を生成する

- 日本語Wikipediaを基盤データとして使用
- 学習データの生成にはQwen-3-Next-80B-A3B-Instructを使用
- 合計1100件のデータを作成
  - 検索結果に回答に必要な情報が含まれる：880件
  - 検索結果に回答に必要な情報が含まれない：220件
- 上記のデータを統合して学習データと評価データに分割
  - 学習データ:880件
  - 評価データ:211件

CoT①の設計方針

- 関連部分のみを選択し、その理由を言語化
- 回答に関連する文をそのまま引用する

CoT②の設計方針

CoT①に加え、以下を追加

- 検索結果に含まれる全チャンクの関連性を評価
- 不要情報を排除し、回答に必要な部分のみ引用

評価結果

評価指標(評価モデル: Qwen/Qwen3.5-122B-A10B)

Answer Accuracy

- 特定の質問に対するモデルの回答と模範解答との一致度をLLMで評価
- 完全回答・部分回答・不正解に分けて判定させる
- <answer>タグ内の最終回答のみを抽出して評価に使用

評価結果

思考過程の学習の有無による性能への影響を評価するため、
未学習モデルでプロンプトだけ変えたものと、ファインチューニングしたモデルを対象に比較

モデル	Answer Accuracy		
	完全回答(件)	部分回答(件)	不正解(件)
未学習モデル	146	36	29
未学習モデル (CoT①プロンプト)	**136 (-10)**	43 (+7)	**32 (+3)**
未学習モデル (CoT②プロンプト)	**137 (-9)**	40 (+4)	**34 (+5)**
CoT①モデル	148 (+2)	**39 (+3)**	24 (-5)
CoT②モデル	**154 (+8)**	**39 (+3)**	**18 (-11)**

SCSK

夢ある未来を、共に創る。

参考文献

- Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, Joseph E. Gonzalez. (2024). RAFT: Adapting Language Model to Domain Specific RAG. arXiv preprint arXiv:2403.10131
- Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, Dong Yu. (2024). Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. arXiv preprint arXiv:2311.09210v2
- 基盤に使用した日本語Wikipedia データセット
  データセット(passages-c400-jawaki-20240401):singletongue/wikipedia-utils · Datasets at Hugging Face
  License: CC BY-SA 3.0 / GFDL
- Shahul Es, Jithin James, Luis Espinosa-Anke, Steven Schockaert. (2025). Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217v2

