

講演会の質疑応答を活用した生成 AI チャットボットの開発 ～大学における研究データマネジメント推進の支援～

和田 和浩¹⁾, 松原 茂樹^{1),2)}

1) 名古屋大学大学院情報学研究科

2) 名古屋大学 情報基盤センター

wada.kazuhiro.s8@s.mail.nagoya-u.ac.jp

Development of Generative AI Chatbot Using Q&A of Lectures for Enhancing Research Data Management in Universities

Kazuhiro Wada¹⁾, Shigeki Matsubara^{1),2)}

1) Graduate school of informatics, Nagoya Univ.

2) Information Technology Center, Nagoya Univ.

概要

オープンサイエンス推進に関する国際的な動きのもと、学術機関において研究データマネジメント（RDM）の組織的な対応強く求められている。しかし、RDM を推進するために知識や経験を有する支援者を十分に配置することは容易ではない。この課題の解消を目的に、本稿では、RDM 支援者向けチャットボットの設計と開発について述べる。生成 AI を活用した的確な回答生成のために、講演会などでの質疑応答記録を活用する。高い回答性能、構築・維持コストの軽減、操作のリアルタイム性を重視し、独自データベースと Web 検索を組み合わせたシステム開発に特徴がある。また本稿では、システムによる回答の生成例を示すとともに、他のドメインへの展開可能性についても考察する。

1 はじめに

研究データマネジメント（research data management: RDM）とは、研究活動で使用または生成された情報を適切に取り扱うことをいう [1]。

- 外付けハードディスクや NAS などに研究データを定期的にバックアップする。
- 研究グループの Web サイトや GitHub^{*1}などの公開プラットフォームで研究データを公開する。

などは RDM の典型的な行動であり、これまでも研究者や研究グループ等で実行されてきた。近年は、国際的なオープンサイエンス動向のもと [2]、より安全で効率的な研究データマネジメントに向け、

- 研究データの保管やグループ内での共有に、学術機関が提供するストレージ基盤を利用する。
- 研究データの取扱いにあたり、学術機関が定めたポリシーやルール、ガイドラインを遵守する。

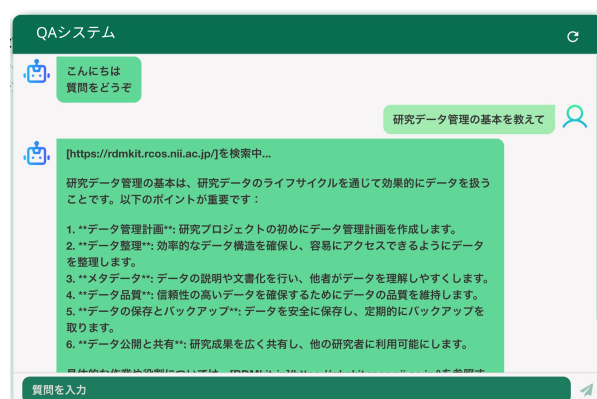


図1 RDM 支援者向け生成 AI チャットボットの動作

などの「学術機関による組織的な RDM」の推進が強く求められている [3, 4, 5]

学術機関において RDM を進めるためには、RDM の支援環境を大学として整備することはもちろん、支援サービスを運営する担当者を配備することも重要となる。しかし現実的には、

- 研究データマネジメントは、研究推進、産学連携、図書館、情報基盤、研究 IR など分野が多岐に

^{*1} <https://github.com/>

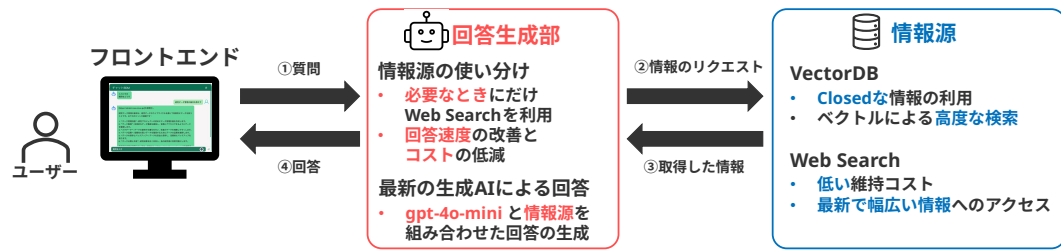


図 2 システムの概要

わたるため、統括するための専門性を備えた人員を新たに配置することは容易ではない。

- 支援サービスの運営を大学職員が担当することも多く、職員の異動が起りやすい環境で、支援者が十分な専門性を備えることは難しい。

という課題がある [6]。この課題の解消に向けて、RDM 支援担当者の経験や知識を補う情報環境の整備が望まれる。

そこで本稿では、学術機関における RDM 支援業務の知識を備えた生成 AI チャットボットの設計と開発について述べる。これを活用することにより、RDM 支援サービスの担当者は、学術機関の RDM に共通する概念や課題、解決のためのノウハウを容易に参照することが可能となる。

ユーザの問い合わせに対して的確な回答の提示を可能とするために、研究データマネジメントに関するセミナーやイベントで実施された講演や討論の記録を活用する。このような場では、講演者と聴講者との間で様々な質問や疑問について見解が交わされることが多く、これらの記録は、チャットボットによる回答生成のための有用な情報源となる。

生成 AI を利用した QA システムを開発した。システムの動作例を図 1 に示す。本システムの設計の特徴は以下の 3 点にまとめられる。

1. 回答生成の根拠となる情報源を効率的に構築するために、独自データを格納した VectorDB、及び、ネット上の情報を利用する Web Search を組み合わせている。
2. システム構築コストを軽減するために、回答生成部では、最新の生成 AI、及び、情報源から取得したコンテキストを利用する Retrieval Augmented Generation (RAG) [7] を使用している。
3. リアルタイム性の高いユーザインタフェースを実現するために、フロントエンドでは、ストリーミングレスポンス等の機能を導入している。

本論文の構成は以下の通りである。まず 2 章で、本システム構築の要点を述べる。次に 3 章では、本システムを構成する情報源、回答生成部、フロントエンドの 3 つの要素について論じる。4 章では、質問を受け取ってから回答を生成するまでの流れについて説明し、5 章では、システムの動作例を示す。最後に 6 章で、研究データマネジメントとは異なる領域への本システムの展開可能性について述べる。

2 QA システム開発の要件

実世界で運用する QA システムは、少なくとも以下の 3 つの要件を満たす必要がある。

1. 高い回答性能を有すること

ユーザは、自らの疑問を解消するために QA システムを参照する。システムが的確に回答できることは、信頼性を備えるための重要な要素である。

2. システムの構築や維持のコストが低いこと

サービス運用のために十分なリソース（費用や人員など）を割り当てることが困難な状況下では、システムの構築と維持のためのコストを軽減する必要がある。

3. ユーザ操作でリアルタイム性を感じさせること

チャット形式のシステムでは、ユーザはインタラクティブなやり取りを期待する。回答の精度が高さだけでなく、リアルタイム性を感じるユーザインタフェースを備えることも重要である。

3 システムの構成要素

生成 AI チャットボットシステムの概要を図 2 に示す。本システムは情報源、回答生成部、フロントエンドの 3 つの要素から構成される。このうちフロントエンドは、ユーザがブラウザを介して利用するアプリケーション部分であり、質問の送信、及び、回答生成部が生成した回答の表示を担う。回答生成部は必要に応じて、回答に必要な情報を情報源から取得する。

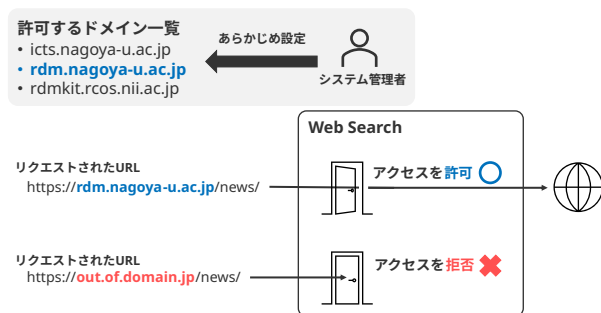


図3 Web Search のドメイン制限機能

以下では、2章で述べた3つの要件の充足を中心に、各構成要素の機能について詳細を述べる。

3.1 情報源

専門知識が必要な研究データ管理に関するQAでは、回答性能を向上させるために、情報源は重要な要素である(要件1)。本システムではVectorDBとWeb Searchの2つの情報源を組み合わせて使用した。

3.1.1 VectorDB

大学内で保管されている、研究データマネジメントに関するFAQ記録を利用するためにVectorDBを利用した。VectorDBはテキストとそれに付随するメタデータと共にベクトルを保存しておくためのデータベースである。通常、テキストを複数のチャンクに分割し、それぞれベクトル化して保存する。本システムでは、1つのQAペアを1つのチャンクとした。

研究データマネジメントに特化したデータを使用していることに加え、ユーザ質問とVectorDB内データとのベクトル間類似度に基づく高度な検索を実現することによって、回答性能の向上が期待できる(要件1)。

3.1.2 Web Search

VectorDBを補完するための情報源としてWebサイト上の情報を利用するWeb Searchを導入した。以下では、VectorDBの欠点とそれを補うためのWeb Searchの役割、ならびに、Web Searchの欠点への対策について述べる。

■ **VectorDBの補完** VectorDBは高品質な情報を提供できる一方で以下の欠点がある。

1. **DBの構築・維持コスト:** DBのみで質問に回答するためには、十分な量のデータを確保する必要があり、その構築コストが大きいことに加え、DBのサーバの維持管理コストも大きい。
2. **最新情報の維持の難しさ:** DB内のデータを最新に保つためには、内容を更新するたびに、該当デー

タを削除し、新しいデータを追加することが必要になる。このためにシステム管理者にかかる負担は大きく、常に最新の状態を維持することは難しい。加えて、データの更新が遅れれば、回答性能の低下を招く。

上記のVectorDBの欠点を補完するためにWeb Searchを導入した。Web Searchでは、検索したいURLを受け取り、そのURLにアクセスしたときのHTMLをMarkdown形式に変換した結果を返す。Webサイト上の情報を利用することで構築や維持のコストを軽減でき、最新の広範な情報にアクセスすることが可能になる(要件2)。

■ **情報の信頼性の確保** Web SearchはWebサイト上の情報を利用するため、不正確な情報が含まれやすいという欠点がある。そこで本システムでは、システム管理者があらかじめ許可するドメイン一覧を設定しておき、リクエストすることができるURLのドメインを制限する機能を追加している(図3)。この機能により生成AIが参照できる情報を信頼の置けるサイトのみに限定できるため、Web Searchの欠点に対処できる(要件1)。

3.2 回答生成部

本節では、回答を生成するための回答生成部について、回答生成時の工夫点と、情報源の使い分けによる効率化について述べる。回答生成部では与えられた情報とユーザーの質問を基に最新の生成AIを用いて回答の生成を行うRetriaval Augmented Generation (RAG) [7]を利用した。回答を生成するための情報には3.1節で述べたVectorDBとWeb Searchから取得したテキストが含まれる。ただし、コスト削減と回答生成にかかる時間の短縮のためにWeb Searchは必要な場合のみ使用する。回答の生成に使用したプロンプトを図4に示す。必要に応じてWeb Searchを利用する機能を実装するためにfunction calling^{*2}を使用した。

3.2.1 コンテキストを利用した回答の生成

回答の生成時にはユーザーからの質問だけではなく、過去の対話履歴、情報源からの情報をコンテキストとして加えるRetriaval Augmented Generation (RAG) [7]を利用した。また、信頼できる回答を生成させるために、情報源から得られた情報に従って回答を生成するための指示を加えている。

^{*2} <https://platform.openai.com/docs/guides/function-calling>

回答生成部のプロンプト

You are a helpful assistant at the Research Data Management department's help desk. The user asks you a question about research data management. You have to answer the question briefly by using the context information provided below.

You can also request other web content using a function calling if needed. If you use a function calling, the request URL must be included an enum parameter.

The answer must follow the information given by the context in DB and the web search results.

The fewer Web Searches, the better.

The links in the response are only for support and should be answered in writing as much as possible.

— Context retrieved from DB

{{dbContext}}

— Context retrieved from Web ({{url}})

{{webContext}}

—

図 4 回答生成部のプロンプト

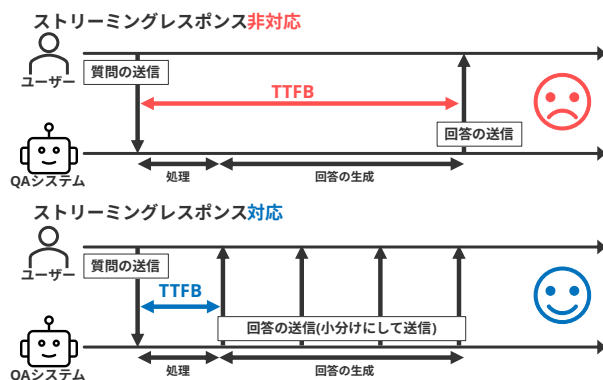


図 5 ストリーミングレスポンスの対応の有無による TTFB の差

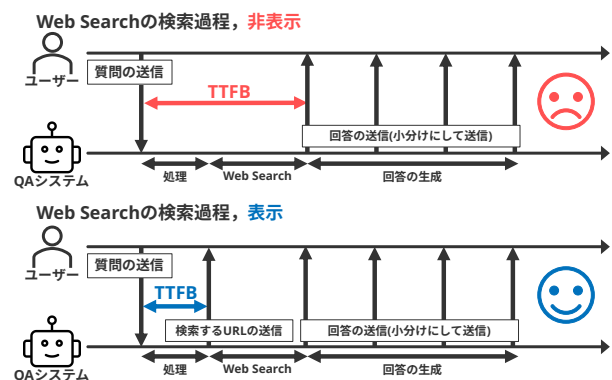


図 6 Web Search の検索過程の表示の有無による TTFB の差

3.2.2 情報源の使い分けによる性能と低コストの両立

VectorDB から十分な情報が得られている場合、必ずしも Web Search を行う必要はない。コスト面と回答生成にかかる時間を考慮すると、Web Search は必要な場合のみ使用されるべきである（要件 2, 3）。これを実現するために回答の生成時に生成 AI に Web Search による追加の情報が必要か判断させることで、 unnecessary Web Search を行わないようにした。

3.3 フロントエンド

フロントエンドはユーザーが QA システムに唯一触れる部分であり、実応用において、回答の性能と並んで重要な要素である。本節では、ユーザー体験の向上、特にユーザーの感じる回答のリアルタイム性の向上（要件 3）のために行った工夫点について述べる。回答のリアルタイム性を考える上で、Time to First Byte

(TTFB) レイテンシが重要な指標となる。TTFB レイテンシはユーザーが質問を送信してから QA システムから最初のレスポンスを受け取るまでの時間である。TTFB レイテンシが短い程、ユーザーの待ち時間が少なく、システムは高いリアルタイム性を有していると言える。

3.3.1 ストリーミングレスポンスへの対応

ストリーミングレスポンスとは準備が整った時点で部分的なレスポンスをクライアントに送り返す技術である^{*3}。これに対応することで全ての回答が生成された後に回答を表示するのではなく、一部の回答が生成された時点で、次々と回答を表示していくことができる。回答全体の生成に時間を要する本システムにおい

^{*3} <https://aws.amazon.com/jp/blogs/news/introducing-aws-lambda-response-streaming/>

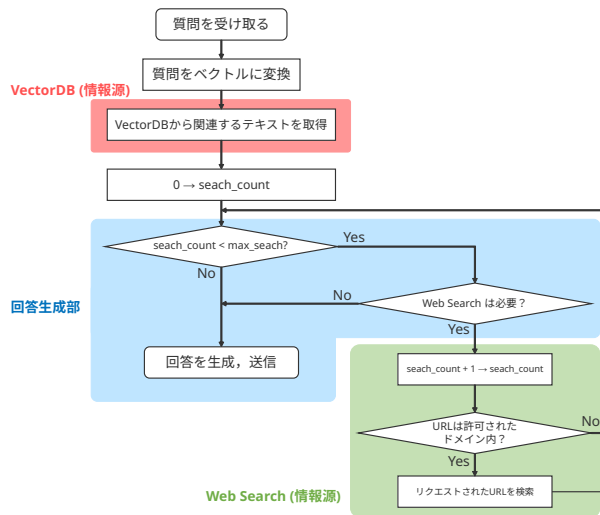


図 7 回答の生成フロー

て、この機能は TTFB レイテンシを大きく低下させ、リアルタイム性を向上させることができる（図 5）。

3.3.2 Web Search の検索過程の可視化

Web Search は回答の性能を向上させる一方で、追加の処理が必要になるため、回答速度が低下する。そうした状況でも TTFB レイテンシを最低限に抑えるために、Web Search を使用したときに、リクエストされた URL をリアルタイムに表示する（図 6）。これにより、Web Search を使用し、回答の生成が遅れた場合でもシステムのリアルタイム性を損なわない。

4 回答生成フロー

本章では、QA システムがユーザーからの質問を受け取ってから回答を生成するまでの詳細なフローについて述べる。図 7 にユーザーから質問を受け取ってからシステムが回答を生成するまでのフローを示す。まず、ユーザーの質問をベクトル化し VectorDB（3.1.1 節）から関連するテキストを取得する。次に、回答生成部（3.2 節）の生成 AI が追加で Web Search（3.1.2 節）が必要か判断する。Web Search が必要と判断された場合はリクエストされた URL の内容を追加し、再度、追加の Web Search が必要か判断する。Web Search が不必要と判断された、または、Web Search の回数が上限を超えた場合は回答の生成を行い、フロントエンドに送信する。

5 回答の生成例

実際に構築した QA システムで生成された 2 つの回答例を表 1 に示す。

表 1 の上部は VectorDB 内の情報のみで回答が可能

な質問に対する回答例である。Web Search は利用されておらず、回答生成部で Web Search が必要か適切に判断されていることが分かる。

一方、表 1 の下部は Web Search を使用しなければ回答することができない質問に対する回答例である。最近の情報に関する質問であるため、VectorDB には回答に必要な情報が含まれていないが、Web Search を使用して、該当する情報が記載されている正しいサイト（<https://rdm.nagoya-u.ac.jp/news/>）まで辿り、回答を生成することができている。

6 他ドメインへの展開可能性

研究データ管理以外の領域でも、専門的な知識や迅速な対応が求められる場面が多く見られる。チャットボットを用いたサポートシステムは、こうしたニーズを満たしつつ、効率的かつ継続的に支援を提供する手段となる。

本システムを他の領域に適用する場合、変更が必要なのは主に情報源である VectorDB と Web Search の 2 つである。VectorDB に登録するデータは用意する必要があるが、Web Search に必要なのはアクセスを許可したいドメインの一覧（e.g., rdm.nagoya-u.ac.jp）のみである。従って、本システムは他の領域でも容易に展開することが可能な汎用的なシステムである。

7 おわりに

本稿では、研究データマネージメント支援者のための生成 AI チャットボットシステムについて述べた。回答の精度に加えて、実応用において重要であるコスト面を考慮したシステムを構築した。また、RDM 以外の領域における同システムの展開についても考察した。

謝辞 本研究は、一部、文部科学省「AI 等の活用を推進する研究データエコシステム構築事業」の支援を受けたものです。本事業の「ルール・ガイドライン整備チーム」でご議論いただいた国立情報学研究所の南山泰之氏、名古屋大学附属図書館の田中幸恵氏に感謝いたします。また、大学における RDM 推進に関するイベントの利活用の在り方についてご議論いただく AXIES 研究データマネージメント部会の幹事団諸氏に感謝します。

参考文献

- [1] 青木学聡. オープンサイエンスと研究データ管理の動向. 情報処理. vol.62, no.5, pp.d1-d11

表 1 生成された回答例

VectorDB に保存されている内容だけで回答可能な質問	
QA システム	ユーザー
こんにちは 質問をどうぞ	
大学での研究データ管理のための取り組みの事例を教えてください	
エディンバラ大学の取り組みがあります。エディンバラ大学では、2011 年に全体に向けたデータ管理ポリシーを策定し、研究データがそのライフサイクルを通じて最高の水準で管理されることを明記しています。また、研修の実施、データ管理基盤の提供、データ管理計画の作成、個別の相談対応など、研究データのライフサイクルに合致した包括的なサービスが提供されています。これにより、研究者のサポートを強化し、研究力の向上につながっています。	
回答に Web Search が必要な質問	
QA システム	ユーザー
こんにちは 質問をどうぞ	
名古屋大学の RDM に関する最近の活動は？	
<p>[https://rdm.nagoya-u.ac.jp/] を検索中...</p> <p>[https://rdm.nagoya-u.ac.jp/news/] を検索中...</p> <p>名古屋大学では、最近以下のような RDM（研究データ管理）に関する活動が行われています：</p> <ol style="list-style-type: none"> 1. **研究データエコシステム東海コンソーシアムのセミナー**が開催され、研究データ管理に関する最新情報や課題共有が行われています。 2. **学術情報基盤オープンフォーラム**において、地域の力を活かした研究データ管理の未来について講演が行われました。 3. **大学における研究データポリシーの策定**に関する研修会も実施され、関係者への情報提供と意識向上が図られています。 <p>これらの活動は、大学および地域における研究データ管理の推進に寄与しています。詳しい情報は名古屋大学の RDM に関する公式サイトを確認することをお勧めします。</p>	
(2021).	kohyo/pdf/kohyo-24-t291-1.pdf.
[2] Department for Business, Innovation & Skills, UK Government, G8 science ministers statement: London, https://www.gov.uk/government/publications/g8-science-ministersstatement-london-12-june-2013 .	[6] 松原茂樹. 名古屋大学における研究データ管理の組織的推進. 学術情報基盤フォーラム 2022 大学での研究データ管理体制構築への道のり, https://www.nii.ac.jp/openforum/2022/day2_nii-rdc1.html .
[3] 内閣府. 国際的動向を踏まえたオープンサイエンスに関する検討会, https://www8.cao.go.jp/cstp/sonota/openscience/ .	[7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. <i>Proceedings Of The 34th International Conference On Neural Information Processing Systems</i> . (2020)
[4] 内閣府, 公的資金による研究データの管理・活用に関する基本的な考え方, https://www8.cao.go.jp/cstp/kenkyudx.html .	
[5] 日本学術会議, オープンサイエンスの深化と推進に向けて, https://www.scj.go.jp/ja/info/	