

# RAG で最適化した生成 AI による HPC ユーザ向けサービスの実現

三上和徳<sup>1,a)</sup>, 中村宜文<sup>1,b)</sup>, 庄司文由<sup>1,c)</sup>, 盛本マリア<sup>2,d)</sup>, 鈴木亮祐<sup>2,e)</sup>

1) 理化学研究所 計算科学研究センター 2) 株式会社 GFLOPS

a)kazunori.mikami@riken.jp, b)nakamura@riken.jp, c)shoji@riken.jp,

d)mariamorimoto@gflops-ai.com e)ryosuzuki@gflops-ai.com

## Realization of HPC User Services Using Generative AI Optimized with RAG

K. Mikami<sup>1,a)</sup>, Y. Nakamura<sup>1,b)</sup>, F. Shoji<sup>1,c)</sup>, M. Morimoto<sup>2,d)</sup>, R. Suzuki<sup>2,e)</sup>

1) RIKEN Center for Computational Science, 2) GFLOPS Co., Ltd.

### 概要

理化学研究所計算科学研究センター (R-CCS) ではスーパーコンピュータ「富岳」のユーザから寄せられる様々な技術的質問や要望へのサポートを行う「富岳サポートサイト」におけるサービスの一環として、2024 年度から生成 AI によるサービスを加え、ユーザ自身による迅速な自己解決を実現するための取り組みを推進している。本稿では R-CCS が生成 AI をサービスに採用した経緯、RAG を応用した最適な生成 AI サービスの構築、得られた効果などに関して報告を行う。また、同じ生成 AI 技術を応用してサービスを開始した HPCI 利用報告書の閲覧支援サービスについても紹介をする。

## 1 ユーザをサポートするサービス基盤

スーパーコンピュータ「富岳」[4] の利用にあたって、ユーザは利用手引書の内容を理解した上で各自の課題に取り組むことになるが、特に利用当初は利用方法についての疑問が生じたりエラーへの対処方法を調査することが必要な局面がしばしば発生することがある。そのような場合に質問や申請を受け付けて対処方法を示す、いわゆるユーザサポートは「富岳」を効果的に利用して成果の創出を後押しする上での重要な役割を担うことになる。以下に「富岳」ユーザに向けたサポートサービス基盤強化の推移を説明する。

### 1.1 「富岳サポートサイト」の開設

「富岳」の運用開始時点においては、ユーザからの質問や各種の申請は全てメールで受付けて、メールで回答を行っていた。2023 年度から、ウェブ上で質問や申請を受け付けてチケットを発行し、対応をアサインされた担当者がチケットの解決に至るまでウェブ上でユーザとチケットを更新し合うチケットサービスへと移行した。このチケットサービスは Zendesk を基盤とするクラウドサービスであり、「富岳サポートサイト」と称する。「富岳サポートサイト」トップページのユーザインタフェイスを図 1 に示す。

このサービスを導入する事により、ユーザサポート



図 1 富岳サポートサイトのユーザインタフェイス

の形態と質が大きく変わることとなった。ユーザが発行するチケットの内容は多岐にわたり、チケット毎にアサインされるサポート担当者は変わる。サポート担当者は複数の機関の所属メンバーから構成され、以下の様な体制となっている。

一次受付：高度情報科学技術研究機構（RIST）  
エスカレーション対応：理化学研究所（R-CCS）  
エスカレーション対応：富士通（保守対応企業）

ユーザ向けに整理されたチケット発行メニューの作成と、発行されたチケットをサポートスタッフが処理する各ステージで連動したツール類を利用する事で、ユーザの利便性とサポート側の運用効率化との両面において効果が得られることとなった。

チケットは基本的にプライベートな扱いであり、発行したユーザとサポートスタッフだけが当該チケットを参照・更新できる設定となっているが、同様な質問チケットが複数回寄せられたり、回答内容が他のユーザにとっても価値ある情報と判断できるチケットは、その質問回答内容を整理し直し、いわゆる FAQ として記事化して「富岳サポートサイト」へ掲示を行う方針としている。

「富岳サポートサイト」に対するユーザの満足度評価は年間平均で 97% 以上と大変高く、広くユーザに受け入れられたことを示している。

## 1.2 「富岳サポートサイト」への生成 AI 応用サービスの導入

「富岳サポートサイト」上で FAQ 記事を充実させてユーザへ利便性の高い情報を提供するという手法は妥当な手段であったと考えられるものの、FAQ 記事数が 300 を超える様になると、多数の記事の中から自分にとって有用な情報にたどり着くことが容易とは言えない状況となってきた。各 FAQ 記事をカテゴリごとに分類して検索が容易となる様なレイアウトを採用したり、記事の件名から本文内容を想起し易いように記述するなど、運用側での努力継続されているが、有用な FAQ 記事により直接的にたどり着くための手法の検討が必要となった。

さらに根本的な課題として、ユーザが「富岳」を利用して目的とする計算ジョブを実行して成果を得るために、「富岳」の利用手引書・各種マニュアル・講習会資料・性能データ等の 100 冊以上のドキュメント類に含まれる数万ページ相当の膨大な情報源から自分が必要とする情報を探し当てて確認する作業が相当の負担となる状況があった。この状況においては従来型のキーワード検索手法はユーザが意図する情報検索の手段としては不十分であることも指摘される。例えば、ある技術的な事項が複数のマニュアルに記載されることもしばしばあるが、それらの内容は同一の場合もあれば、用途に応じて焦点の当て方を変えた異なる説明

方法となっていることもある。さらには、調査したい事項そのものが概念として表現はできるが、具体的なキーワードとして想起できないという状況もしばしばある。

膨大な情報蓄積資源の中から、ユーザ自身にとって必要な情報を適切に得るための手段を提供すること、ひいてはユーザ自身による問題解決を促進することは「富岳」を運用するチームにとって重要な課題であった。

このような背景のもと、ChatGPT [1, ] を筆頭格として近年非常に進化が進んだ生成 AI を応用した質問への自動回答および高度検索サービスを導入する検討を 2023 年度から開始した。質問への回答がほぼリアルタイムで提示され、また通常の会話と同様に質問と回答の内容を踏まえた上で次の質問への回答を行うような動作継続性を持たせることができるため、ユーザが必要とする情報へ自然にたどり着きやすいという生成 AI の特性は、我々が期待するサービスに適していると考えた。サービスの実現手法として様々なアプローチがあり得たが、サービスを利用する対象者が「富岳」ユーザであり、彼らが必要とする技術的情報は全て上述したドキュメント群のいずれかに記載されていることがわかっているため、それらのドキュメントから容易にかつ正確に必要な情報を調査提示することが可能な技術と目される検索拡張生成（RAG: Retrieval Augmented Generation）[2] を採用することが相応しいとの判断を持った。RAG では専用の知識データベースを構築することになるが、データベース出典元文書の開示先が限定されている場合は、その開示対象者へ限定したデータ利用が可能であるため、「富岳」ユーザに限定される「富岳サポートサイト」のサービスとして追加することが可能である。また古くなったデータを明示的に削除・更新することができるという、進歩のペースが速い技術分野への適用するのに望ましい特徴をもっている。実際の検討作業は、RAG フレームワークを提供する複数の生成 AI サービス事業者と意見交換を行いながら進めた。実現の可能性についてある程度の見通しが得られた時点で事業者各社に概念検証を協力してもらい、具体的な動作・仕様の確認を行った後、入札による事業者の決定を経て、2024 年度に「富岳サポートサイト」へ「AI チャット」機能および「高度 AI 検索」機能としてサービスの追加を実施した。

## 2 生成 AI チャット AskDona

「富岳サポートサイト」の生成 AI サービスは株式会社 GFLOPS が開発した AskDona[3] を中心技術として採用している。AskDona はマルチエージェント型の RAG 機能を GFLOPS 社独自の技術で構成し、回答生成に大規模言語モデルを組み入れたサービスである。

ユーザは質問入力をテキストで行う。これは「富岳サポートサイト」の主要なサービスが基本的にはテキストベースで行われていること、および「富岳」のほとんどのユーザがプログラムのコーディングやスクリプト作成等を自身で行うと考えられ、質問の入力や解答の提示もなじみが深いテキスト形式が自然であろうとの判断による。

質問を受け付けた AskDona は、入力された質問文を意図やコンテキスト、RAG データベースの内容を考慮して、サブクエリ化する。各サブクエリに対して関連が強いデータチャンクを RAG の知識データベース（ベクトル検索用データベース、キーワード検索用データベース）から検索・取り出し（retrieval 処理）、取得した情報を重複削除や順序入れ替え、再加工などの処理を施す。RAG 知識データベースから取得し再処理を実行した情報をユーザーの元の質問文（クエリ）とシステムプロンプトとともに大規模言語モデル（LLM）へ送信し、LLM から得られた回答内容をユーザへの回答文として統合（synthesize 処理）した上で、チャットウィンドウ上で表示する。AskDona のデータ処理フローの概念図を図 2 に示す。

トップメニュー上段の「質問する」をクリックすると生成 AI AskDona のチャットセッションが開始される。AskDona のユーザインタフェースを図 3 に示す。

チャットセッション下部に質問入力スペースがあり、オプションとして選択可能な Fast モードボタンと Boost モードボタンが準備されている。通常は標準モードで適切な回答が得られることが多いが、回答速度を求める場合は、サブクエリ化処理や多段階検索、推論を一部省略する Fast モード、標準モードよりも多角的な視点でより多くの文書を分析した回答を求める場合は RAG データベースの調査深度がより深く、より多くの推論を実行する Boost モードの選択肢がある。

チャットセッションに表示されているタブの切り替えを選択すると RAG の知識データベースを調査する Deep Research 機能の選択肢も用意されている。

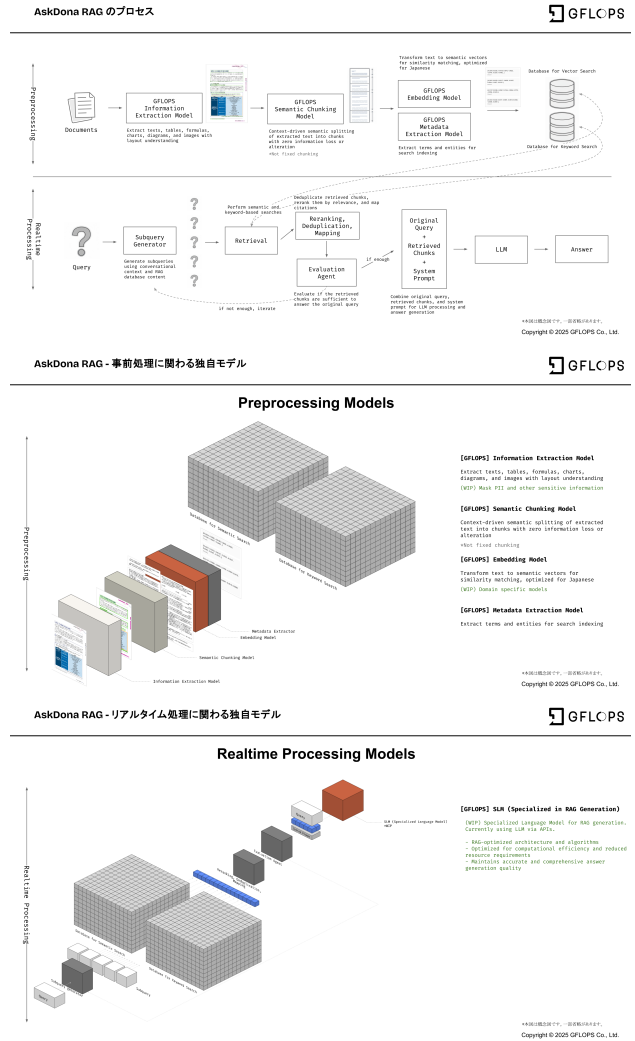


図 2 AskDona のデータ処理フロー



図 3 生成 AI 応用サービス AskDona の UI

Deep Research は対話的な回答生成ではなく、与えられた調査テーマに対して RAG の知識データベースの調査を AI エージェントが自律的に実行する機能である。入力された調査テーマに対して質問文を分析し、調査にあたって必要な補足情報をユーザーに聞き返すプロセスを経て、ユーザーが補足情報を送信すると調査設計書を生成し、調査が開始される。調査依頼内容および RAG 知識データベースの情報量に応じて、数十から 100 を超えるの参照元データから構成される 4000-6000 文字程度のレポートが生成される機能である。Deep Research 機能については、RAG の知識データベースを横断した体系的な調査を目的とする場合や断片的な情報を統合してレポート化したい場合に有用である。

以下に標準モードにおける実際のユーザからの質問とその解答例を図 4 に示す。右上に反転色で質問文をエコー表示した後、回答文の作成が始まる。この例では 30 秒程度の待ち時間の後、まず回答の要約が表形式で表示され、続いて具体的な指定パラメタの説明と関連する注意事項、そして回答文で引用した参照元文書名がそのリンクとともに示されている。改めて指摘するまでもなく、生成 AI の回答が 100% 正しい保証はない。「富岳サポートサイト」の生成 AI サービスが精度の高い RAG 技術で構成されているとはいえ、ユーザ自身が参照元文書の該当箇所で回答の正しさを確認することは、どのような生成 AI を利用する場合でも共通する留意事項である。

生成 AI AskDona による質問回答サービスは 2024 年度から始まり、当初はユーザが従来方式のチケット発行・有人対応サービスと、AskDona による自動回答サービスを選択できるメニューを提供していたが、半年間の経過観察を経て、全ての質問を AskDona が最初に受け付けるメニューへ変更をおこなった。ユーザの質問へ AskDona がまず質問対応にあたるが、もし期待する回答が AskDona から得られずに従来方式のチケット対応を希望する場合は、チャットセッションで「有人対応をお願いしたい」と入力することによりチケット発行メニューを呼び出すことができる。このフローで発行されるチケットには AskDona とのチャットセッション履歴を紐づける情報がサポートスタッフ向けに付加される。AskDona によるサービス加えたユーザサポート体制は図 5 のようになった。

R-CCS と株式会社 GFLOPS が継続的に協働してきたことにより、AskDona の回答精度は導入当初と比較して大きく向上している。この改善プロセスには、

AskDona

ノード内テンポラリ領域と共有テンポラリ領域のサイズのデフォルト値と最大値を教えてください

Dona です。ノード内テンポラリ領域と共有テンポラリ領域のサイズに関するご質問ですね。それぞれのデフォルト値と最大値について、資料から得られた情報をもとに詳しくご説明します。

**ノード内テンポラリ領域と共有テンポラリ領域のサイズ**

ノード内テンポラリ領域と共有テンポラリ領域のサイズについて、デフォルト値と最大値は以下の通りです。

領域の種類	パラメータ	デフォルト値	最大値 (理論値)	最大値 (実質ノード毎)
ノード内テンポラリ領域	localtmp-size	0MB	2,147,483,647MB	約87GB
共有テンポラリ領域	shoredtmp-size	0MB	2,147,483,647MB	約87GB

これらの値について、以下詳しく解説します。

**デフォルト値**

ノード内テンポラリ領域 (localtmp-size) と共有テンポラリ領域 (shoredtmp-size) のサイズをジョブ投入時に指定しない場合、デフォルト値は両方とも「0MB」になります。

- localtmp-size を 0 にした場合、ノード内テンポラリ領域にファイルを作成することはできません。
- shoredtmp-size を 0 にした場合、共有テンポラリ領域にファイルを作成することはできますが、データを格納することはできません。

**最大値**

最大値には、システム上の理論的な上限と、ハードウェアの制約による実質的な上限の2種類があります。

- 理論上の最大値: システムが受け付ける理論上の最大値は、ノード内テンポラリ領域と共有テンポラリ領域のどちらでも「2,147,483,647MB」です。
- 実質的な最大値: 実態に1計算ノードあたり利用できる容量は、ハードウェア (SSD) のサイズに依存します。最高では、1計算ノードあたり約「87GB」の容量を、ノード内テンポラリ領域、共有テンポラリ領域、そして第2階層ストレージキャッシュ領域の3つで分けて使用します。そのため、それぞれの領域で設定できるサイズは実質的に「8MB」から「87GB」の範囲となります。

**サイズ指定に関する重要な注意事項**

サイズを指定する際には、以下の制約にご注意ください。

- 合計サイズの制限: ノード内テンポラリ領域と共有テンポラリ領域は、同じストレージのノード上のSSDに確保されます。そのため、それぞれのサイズが上限以下であっても、両者の合計が利用可能な物理容量を超えることはできません。
- 第2階層ストレージキャッシュの確保: 第2階層ストレージのキャッシュとして、最低でも1計算ノードあたり「128MB」の容量を確保する必要があります。したがって、localtmp-size と shoredtmp-size の合計は、以下の式を満たす必要があります。

この制約を超えて指定しようとすると、ジョブ投入時にエラーが発生します。エラーメッセージには合計値の上限サイズが表示されます。

これらの情報を参考に、ジョブの特性に合わせて適切な領域サイズを指定してください。

**参照元**

- ジョブ運用ソフトウェア APIユーザーズガイド コマンドAPI編.pdf
- ジョブ運用ソフトウェア エンドユーザ向けガイド.pdf
- 利用手引き - システム利用 - 利用およびジョブ実行編.zip
- 利用手引き - プログラミングガイド - プログラミングガイド (IO編).zip

図 4 AskDona への質問と回答の例

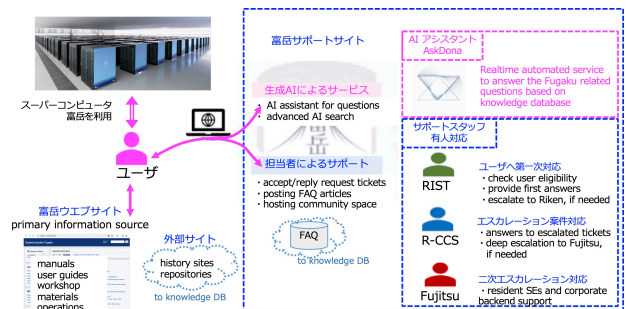


図 5 富岳ユーザへのサポート体制

2025 年 3 月から 5 月にかけて 20-25 人月のエンジニアリング工数を投入した。回答精度を向上するための具体的な施策として以下のようなことを行ってきた。

- 適切な元文書 (特に FAQ 記事) の充実化と知識データベースの更新
- 元文書からの情報抽出モデルの改善 (ノイズ除去や視覚情報の言語化含む)
- 抽出した情報を分割するモデルの改善 (文脈を考慮した Semantic Chunking)
- 分割した情報をベクトル化する Embedding モデルの改善
- 分割した情報からのメタデータ抽出モデルの改善
- Retrieval 手法の更新 (Agentic Retrieval)

- 適切な回答を得るためのシステムプロンプトの改良
- より高機能な LLM の採用
- 回答に不満足が示されたチャットセッション検知ツールの充実

### 3 生成 AI 導入の効果

生成 AI サービスを導入した効果を統計情報をもとに分析する [5]。図 6 は、富岳サポートサイトにおける 2022 年 5 月 11 日から 2025 年 8 月 31 日までの四半期ごとのユーザからの質問や申請、またサポートスタッフから個別ユーザへの通知や連絡についての新規チケット発行数である。2022 年第 2 四半期から 2023 年第 1 四半期までは、旧チケットシステムからの移行時期のため、新規チケット発行数は期を経るごとに増加している。2023 年第 2 四半期以降のいくつかの四半期において、第 2 四半期と第 4 四半期にチケット発行数が多く、第 1 四半期と第 3 四半期に少ないという傾向が見られる。これは、富岳において新規課題が毎年 4 月と 10 月の年 2 回開始されることに起因する。新しいユーザが利用を開始する際には、マニュアルに記載されている内容やスパコン・Linux の基本操作に関する初歩的な問い合わせ、自身のプログラムのコンパイル方法がわからない、性能が十分に出ない、特定のソフトウェアの使用方法がわからない、商用ソフトの利用を始めたいといった、さまざまな理由で問い合わせが増加するためである。

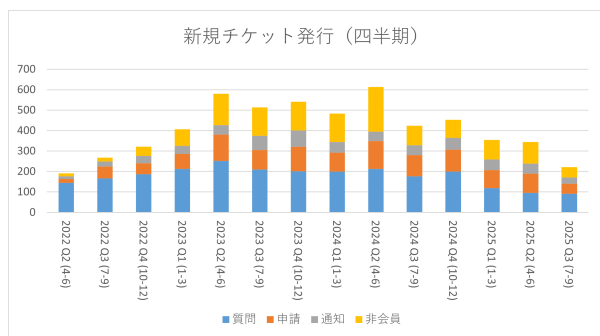


図 6 四半期ごとの新規チケット発行数

富岳サポートサイトには、富岳ユーザ向けサイトと非会員向けサイトが存在し、AI チャットおよび高度 AI 検索は富岳ユーザ向けサイトのみに導入されている。ユーザからの技術的な質問については生成 AI が有効に働くことが期待される一方、ユーザからの申請やスタッフからの通知に関するチケットには影響がほとんどない。そのため、チケットを質問、申請、通知、

非会員の 4 種類に大別し、生成 AI による効果を確認する。

生成 AI によるサポートは 2024 年 7 月 9 日に開始された。導入当初は、有人チケットフォームと AI チャットの双方が存在したが、AI チャットの利用促進と、新規質問者が生成 AI の実力を直接体験することを目的とし、2024 年 12 月 5 日には、有人チケットフォームへのアクセスを AI チャット経由にする予定が告知され、2025 年 2 月 1 日からは、有人チケットフォームを利用する場合は必ず AI チャットを経由することとなった。

2025 年第 1 四半期以降、質問チケットの発行は著しく減少しており、これは生成 AI の効果によるものと考えられる。なお、表 1 に示す通り、富岳の一日当たりのアクティブユーザ数および利用者数は年々増加しており、ユーザ数の増加にもかかわらず質問チケットが減少していることから、生成 AI によるサポートが実際に問い合わせ削減に寄与していることが示唆される。

表 1 富岳の年度ごとのアクティブユーザ数/日と利用者数

年度	アクティブユーザ数/日	利用者数
2020	約 170	1798
2021	約 330	2321
2022	約 355	2916
2023	約 395	3521

図 7 および図 8 は、それぞれ月ごとの新規質問チケット発行数と前年同月比を示している。これらの図から、2025 年 1 月以降、新規質問チケットが顕著に減少していることが確認できる。特に、前年同月比では最大で約 6 割の質問チケットの減少が認められた。

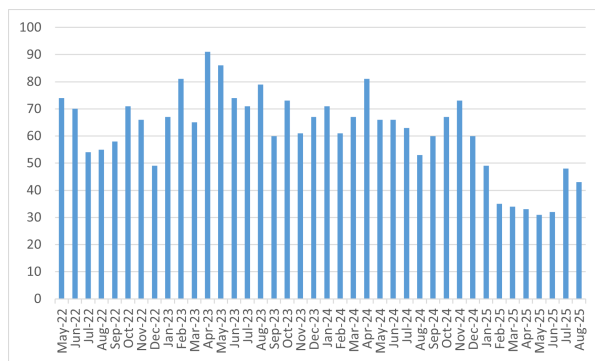


図 7 月ごとの新規質問チケット発行数

AI チャットはサービス開始以降、継続的に利用されている。図 9 および図 10 は、それぞれ月ごとの

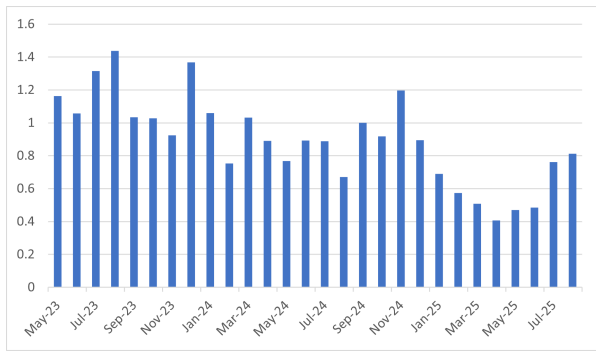


図8 月ごとの新規質問チケット発行数の前年同月比

AskDona のセッション数 (棒グラフ) とセッション当たりの質問数 (折れ線)、および月ごとの AskDona に対するユーザーからのフィードバック件数を示している。フィードバックはサービス開始月に多く、その後一時的に減少したが、最近では再び増加傾向にある。

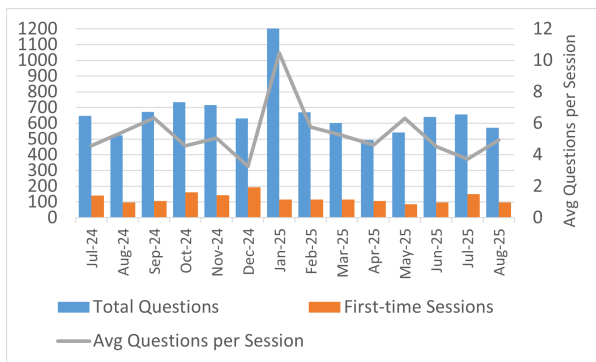


図9 月ごとの AskDona のセッション数 (棒線) およびセッション当たりの質問数 (折れ線)

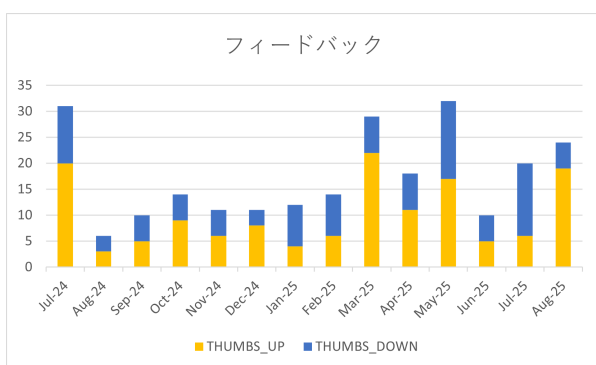


図10 月ごとの AskDona へのユーザーからのフィードバック

## 4 HPC ユーザに向けた生成 AI 技術の応用

前節までに説明した AskDona RAG による生成 AI 技術が相当に効果的であると判明したことをうけて、同技術を特定文書群の高度な調査支援ツールとして応

用してはどうかとの意見が上がり、その取り組みに着手した。

### 4.1 HPCI 利用報告書の閲覧支援 AI サービス

最初の応用先として、革新的ハイパフォーマンス・コンピューティング・インフラ (HPCI) が提供する国内大学・研究機関の HPC システムを課題利用した成果の報告書が HPCI 研究成果ページに登録・公開されているが [6]、これらを全て知識データベース化して、AskDona を用いることにより様々な角度からの検索・比較・調査が可能となるサービスを提供する試みである、「研究成果閲覧支援サービス」(略称: スパコン成果ナビゲータ) について報告を行う。

このサービスを実現するために、2012 年度から今年度までの利用報告書を全て知識データベース化する作業にとりかかり、1 ヶ月程度の作業期間でデータベースの構築と AskDona への組み込みが完了した。サービスの実装環境としては、「富岳サポートサイト」において既に AskDona のサービス環境を様々な改良を含めて実装済みであることから、「富岳サポートサイト」と同じ Zendesk 基盤の上にスパコン成果ナビゲータのページを設けた。興味ある課題の検索・報告書内容の論点整理・比較などの作業を自動的に行うことが可能なサービスであり、ユーザが課題に取り組む際に過去の研究事例を効率よく調べ上げるためのツールとしての利用が期待される。

スパコン成果ナビゲータのチャットページを開くと参考となりそうな質問例や必要に応じて選択できる絞り込みフィルターなどが準備されており、ユーザはこのサービスから得られる情報の期待感を持つことができる。チャットページの質問例から最初の一件を取り上げて回答を求めた場合の実際の出力回答を一部抜粋して図 11 に示す。

現在掲載されている利用報告書を人手によって網羅的に調査するには多大な労力がかかると考えられるが、この閲覧支援 AI サービスを用いることによって非常に効率よく情報収集・整理を行うことが可能であることが示される好例である。

スパコン成果ナビゲータは一般のユーザが利用可能なサービスであり、HPCI のホームページから「HPCI 研究成果」タグを辿ればアクセス可能である [7]。開始したばかりのサービスであり、これからユーザに向けた周知と利用を促す施策が必要な状態であるが、「富岳」ユーザに限らず全ての HPC ユーザに広く利用していただきたいと願っている。

新型コロナウイルスに関するテーマでの研究成果をリストアップして、課題の概要と、使用した計算資源をまとめでくわしい。

検索しました。ご提供いただいた資料に基づき、新型コロナウイルスに関連する研究課題をリストアップし、それぞれの概要と使用した計算資源をまとめました。

これらの研究は、ウイルスの基本的な性質の解明から、治療薬の開発、感染リスクの評価まで、多岐にわたるテーマを扱っています。

**新型コロナウイルス関連研究課題一覧**

以下に、各研究課題の概要と、その研究で利用されたスーパーコンピュータなどの計算資源をまとめました。

課題番号	課題名	課題の概要	使用した計算資源
hp20007 8	Cryo-EM プレートマッピング法による新型コロナウイルス-スプレッドによる変異型スパイクタンパク質 (S140) を対象に、変異がもたらす影響を調査しました。その結果、DS14G 変異がウイルスの感染に重要な役割 (感染抑制メカニズム) の「フックコンフォメーション」を安定化させる傾向があることを明らかにしました。	クライオ電子顕微鏡 (Cryo-EM) の実験画像からウイルスの構造を推定する統合ソフトウェアを「富岳」上に構築。野生型および変異型スパイクタンパク質 (S140) を対象に、変異がもたらす影響を調査しました。その結果、DS14G 変異がウイルスの感染に重要な役割 (感染抑制メカニズム) の「フックコンフォメーション」を安定化させる傾向があることを明らかにしました。	理化学研究所 スーパーコンピュータ「富岳」 (R160,599 ノード稼働) 金沢大学 Psaline (63,585 ノード稼働) 京都大学 XC40 (System A) (156,934 ノード稼働)
hp21008 6	新型コロナウイルスを宿主とした統合的免疫感受性リスク評価システムの開発と社会実装	計算高度科学を用いて、新型コロナウイルスの感染・スプロールによる感染リスクを評価するシステムを開発し、社会実装を目指す研究です。	理化学研究所 スーパーコンピュータ「富岳」 (5,169,989 ノード稼働)
hp22002 67	新型コロナウイルス主要タンパク質の相互作用を高精度に解析する大規模電子分子動力学シミュレーション (MD)	ウイルスの複製に不可欠な「主要タンパク質」の働きを調べる化合物を設計・シミュレーション (コンピュータでの計算) 研究です。電子分子動力学法を用いて、複製や宿主のメカニズムを詳細に解析します。	理化学研究所 スーパーコンピュータ「富岳」 (13,828,690 ノード稼働)
hp20013 0	SARS-CoV-2 ウイルスエンベロープ膜タンパク質の構造解析	ウイルスを覆うエンベロープ膜とそこに存在するタンパク質の構造をシミュレーションし、ウイルスの構造や感染機序の解明を目指す研究です。	理化学研究所 スーパーコンピュータ「富岳」 (3,676,514 ノード稼働)
研究*	新型コロナウイルスなどのパンデミック発症時に、関連タンパク質の相互作用を迅速に解析できる計算科学環境を整備する継続的研究です。FMO (フラグメント分子動力学) 法 (MD) (分子動力学) シミュレーションを用いて、ウイルスの複製機序 RBD と宿主細胞の ACE2 受容体の結合などを解析しています。	新型コロナウイルスなどのパンデミック発症時に、関連タンパク質の相互作用を迅速に解析できる計算科学環境を整備する継続的研究です。FMO (フラグメント分子動力学) 法 (MD) (分子動力学) シミュレーションを用いて、ウイルスの複製機序 RBD と宿主細胞の ACE2 受容体の結合などを解析しています。	理化学研究所 スーパーコンピュータ「富岳」・hbr-hp210000: 2342497 ノード稼働 (4)・hp2200205: 3,081,160 ノード稼働 (4)・hbr-hp230017: hp2300300 算的資料の必要使用量の記載なし (4) (4)
hp20014 2	COVID-19 ウイルスの RNA ポリメラーゼと宿主細胞の分子動力学シミュレーション	ウイルスの複製情報 (RNA) を複製する「RNAポリメラーゼ」を模倣とし、シミュレーションやアビシブルといった計算機環境がどのように作用するかを分子動力学シミュレーションで解析する研究です。	東京工業大学 TSUBAME3.0 (17,460 ノード稼働)
hp23001 3	新型コロナウイルスのスパイクタンパク質と宿主細胞の相互作用を高精度に解析する大規模電子分子動力学シミュレーション (MD)	ウイルスの複製に関わるもう一つの酵素「パリンプロテアーゼ」を模倣する化合物 (リガンド) との結合のメカニズム (結合自由エネルギー) を計算し、より効果的な阻害剤の設計を目指す研究です。	東京工業大学 TSUBAME3.0 (66,493 ノード稼働)

図 11 研究成果閲覧支援サービスへの質問と回答の例

## 4.2 JHPCN 成果報告書への拡大適用

スパコン成果ナビゲータは当初 HPCI の利用報告書を対象としていたが、同様な文書に対して容易に拡張適用が可能なサービスである。大学 ICT 推進協議会が登録管理する JHPCN 成果報告書に対してもそのような拡張適用を行った。

JHPCN 成果報告書 [8] を前節で述べた HPCI 利用報告書と同様に知識データベース化と AskDona への組み込みを行い、ユーザがスパコン成果ナビゲータを利用する際に絞り込みフィルターで「利用枠」を「JHPCN」と選択すると、JHPCN 成果報告書に対して様々な角度からの検索・比較・調査が可能となっている。

## 5 まとめと今後の展望

RAG で最適化した生成 AI による HPC ユーザ向けサービスの実現事例を紹介した。AI 技術分野の著しい進展に伴い、実務レベルでの活用を目指す同種の取り組みによって、高品質かつ自動化されたサービスが多様な分野で実現可能になると考えられる。

「富岳サポートサイト」への生成 AI サービスの追加に際しては、専門性の高い知識に基づいた正確な回答が生成され、ユーザーの期待水準を満たすことが必須条件であった。実現にあたって富岳の運用とユーザサービスの提供する側と、生成 AI の技術開発を担う側の連携と協力が不可欠であり、サービスが定着したことは両者による協業成果と言える。

より広範な HPC ユーザに向けた生成 AI サービス

として開始した「研究成果閲覧支援サービス」(スパコン成果ナビゲータ) については、萌芽的な取り組みではあるが、実際の HPC ユーザである研究技術部門の担当者のみならず、企画事務部門の担当者にとっても利用価値が高いという声が届いている。さらなる利用の拡大を期したい。

## 参考文献

- [1] OpenAI: Introducing ChatGPT, 2022. URL <https://openai.com/blog/chatgpt>
- [2] Patrick Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks arXiv:2005.11401 [cs.CL] <https://doi.org/10.48550/arXiv.2005.11401>
- [3] AskDona ホームページ <https://askdona.com/>
- [4] R-CCS ウェブページ 「富岳」について <https://www.r-ccs.riken.jp/fugaku/>
- [5] 中村宜文, PC クラスタワークショップ in 柏, 2025 [https://www.pccluster.org/ja/event/data/250627\\_PCC-WS-Kashiwa\\_04\\_nakamura\\_v2.pdf](https://www.pccluster.org/ja/event/data/250627_PCC-WS-Kashiwa_04_nakamura_v2.pdf)
- [6] HPCI 研究成果 利用報告書 URL [https://www.hpci-office.jp/achievements/user\\_report](https://www.hpci-office.jp/achievements/user_report)
- [7] 研究成果報告書 閲覧支援サービス URL [https://fugaku2nd.zendesk.com/hc/ja/p/report\\_viewer\\_top/](https://fugaku2nd.zendesk.com/hc/ja/p/report_viewer_top/)
- [8] JHPCN 採択課題【年別一覧】 URL <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/adoption>