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 [1,a)], 中村宜文 [1,b)], 庄司文由 [1,c)]

1) 理化学研究所 計算科学研究センター

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

Realization of HPC User Services Using Generative AI Optimized with RAG

Kazunori Mikami[1,a)], Yoshifumi Nakamura[1,b)], Fumiyoshi Shoji[1,c)]

1) RIKEN Center for Computational Science

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。

論文の構成案

1. サービスの形態が段階的に高度化してきた経緯説明
   （a）富岳サポートサイトの開設）
   （キーワード：Zendesk を基盤とするチケットサービスへの移行、ウエブブラウザ UI、チケットサービス）
   （b）生成 AI AskDona による自動回答サービスの追加
   （氾濫する情報量、効率の良い情報検索手法の模索、質問への回答を自動的に提示するサービスの提供）
   （＝＞生成 AI ＋ RAG、リアルタイム性、自動性、）
2. 高度な生成 AI サービスに向けた RAG フレームワーク
   （a）知識データベースの構築と改良
   （GFLOPS と R-CCS とで継続的に協働していることを示す）
   適切な文書の選択とベクトルデータベースの構築・更新
   適切なデータの取り出し（特に PDF 文書のクリーニング）
   適切なアノーテーション機能
   （b）回答精度を向上するための取り組み
   回答内容に満足・不満足が示されたチャットセッションを検知するツールの組み込み
   AskDona とのチャットセッションから有人対応を希望する場合の対応ワークフローの実現
   - 質問チケット発行メニューの自動呼出し
   - 発行されるチケットへの AskDona セッション紐付け情報の付加
3. 生成 AI サービスの性能と導入効果、利用者の評価
   （a）実例の紹介
   （b）質問チケット数の推移
   （c）Good/Bad 評価の統計
   （d）ユーザの質問内容の高度化
4. 同じ生成 AI 技術を応用したサービスの水平展開
   （a）HPCI 利用報告書の閲覧支援サービス
   （b）JHPCN 成果報告書の閲覧支援サービス（予定）
5. 今後の計画

用語確認：生成 AI か、単に AI とするか

1　ユーザをサポートするサービス基盤

「富岳」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となる局面がしばしば発生することがある。そのような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、ウエ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、チケットの内容が解決に至るまでウエ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を提供する相互サイト「富岳サポートサイト」を開設した。「富岳サポートサイト」は Zendesk を基盤とするクラウドサービスである。このサービスを導入する事により、ユーザサポート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担当者は変わる。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。

十分に整理されたチケット発行メニューとチケット処理の各ステージで連動する各種自動化ツールを利用する事で、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。

サポートチケットの受付と対応を行う期間

- 一次受付機関：高度情報科学技術研究機構（RIST）
- 二次受付機関：理化学研究所 計算科学研究センター（R-CCS）
- 保守対応企業：富士通

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 を応用したサービスの導入

「富岳サポートサイト」での FAQ 記事を充実させてユーザへの情報提供を強化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。

また「富岳」の利用にあたって、「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 冊以上の膨大なドキュメント類から、自分が必要とする情報を探し当てて問題の解決をはかる作業が相当の負担となり、

2　富岳サポートサイトへの生成 AI の導入

論点：導入背景：ユーザ側でのハードル、運用側でのハードル
論点：ユーザ自身による問題解決の促進
論点：生成 AI 提供業者数社による概念検証の実施
論点：入札による業者決定。機能強化・改善要望に機動的に対応する業者が選定されたことは重要なポイントであった

3　生成 AI AskDona

論点：AskDona の構成概要
論点：RAG のメリット（回答正解率の高さ、ハルシネーションが起きにくい構成）
論点：富岳サポートサイト専用の知識データベースの構築
論点：チケットサービスと生成 AI チャット機能の統合
論点：個人情報の不所持方針

4　生成 AI 導入の効果

論点：発行チケット数の変化
論点：ユーザからのフィードバック
論点：ユーザの利用方法の変化

論点：課題：ユーザフィードバック率の低さ
論点：課題：人手で対応するチケットサービスの満足度 V.S. 生成 AI サービスの満足度

5　HPCI 利用報告書の閲覧支援 AI サービス

論点：同じ AI 技術の水平展開
論点：HPCI 利用課題の検索、論点整理、調査・比較
論点：JHPCN 成果報告書の閲覧機能を追加

%参考文献

参考文献

[1] 雑誌の場合：著者名、タイトル、雑誌名 巻、号、ページ、発行年.
[2] 書籍の場合：著者名、書名、参照ページ、発行所、発行年.