技術紹介

生成 AI を活用したユーザサポート支援システム

Customer Support System Powered by Generative AI

西島　隆人 *1
NISHIJIMA Takato

四月朔日　勉 *2
WATANUKI Tsutomu

窪田　圭志 *3
KUBOTA Keiji

1. はじめに

当社では，大規模言語モデル（Large Language Model，以下，LLM）を用いて，自社製品に関する質問に回答するシステムを構築し，ユーザサポート業務への適用可能性を検証しています。ChatGPT など一般に用いられる LLM はウェブ上の公開情報しか知らないため，自社製品に関する質問には正確に回答できませんが，本システムでは，検索拡張生成（Retrieval-Augmented Generation，以下，RAG [1]）という手法を用いることで，自社製品独自の情報に基づいた正確な回答を得ることができます。

本稿では，まずシステムの機能と構成を紹介し，次いで，事前検証の結果について述べます。最後に，実業務での試験運用の状況を述べ，そこで得られた課題と改善策について考察します。結論として，生成 AI のユーザサポート業務への適用が有効であることを示します。

2. システムの機能と構成

本システムは，以下の機能を持ちます。

（1） 関連情報に基づく回答生成

自社製品に関する質問をシステムに投じると，ウェブ上に存在する一般的な知識だけでなく，システムに事前に登録された関連情報に基づいた回答が生成されます。図 1 に実際のシステムの画面を示します。赤枠で囲まれた箇所が，質問に対応する関連情報です。また，回答中の[1], [2]といった記号は，関連する情報のうち回答の根拠となったものを示しており，生成された回答の確からしさをその場で確認できます。

（2） 返信用メール文書の自動生成

宛名込みの返信用メール文書をワンクリックで生成できます（図 1 青枠）。問い合わせメールに含まれるユーザ名や所属といった情報を利用して，あたかも，サポートセンターの一員が作成したかのような，適切な言葉使いでメール文書が作成されます。

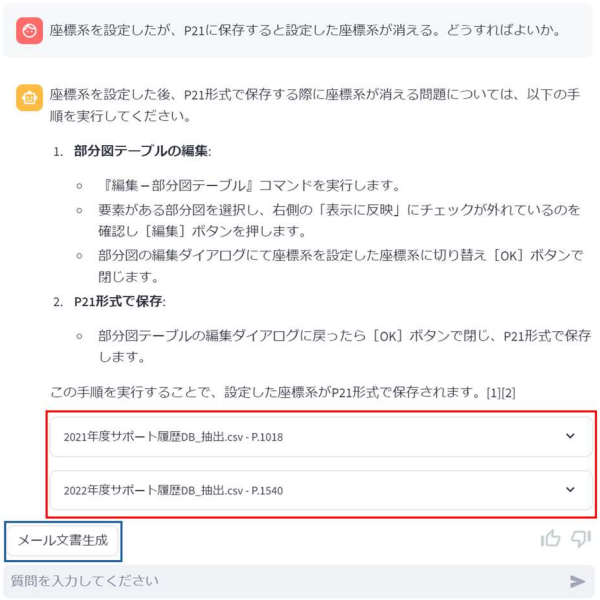

図 1　システム画面

続いて，本システムの構成と処理フローを図 2 に示します。一般に RAG は，質問に対する関連情報を検索し，それを文脈に含めて質問を拡張した指示書（プロンプト）を LLM に入力することで回答を生成する仕組みです。本システムで実装している RAG の一種である仮説的文書埋め込み（Hypothetical Document Embeddings，以下，HyDE [2]）では，質問に対する仮回答を LLM で生成し，質問と組み合わせて検索を行います。関連情報の多くは，質問形式ではなく，説明形式（回答に相当）であるため，仮回答と質問を組み合わせて検索することで，より回答に適した関連情報を取得できることになります。

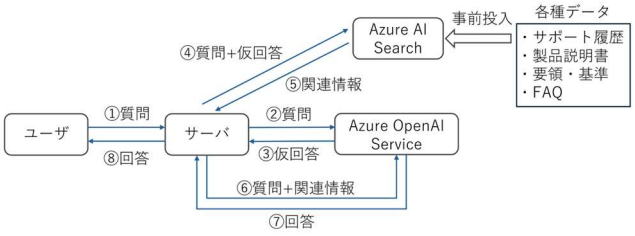

図 2　システム構成と処理フロー

*1 川田テクノシステム㈱開発本部クラウド開発部 主任
*2 川田テクノシステム㈱開発本部クラウド開発部 部長
*3 川田テクノシステム㈱開発本部エンジニアリング開発部 主任

回答生成には，Azure OpenAI Service を通して GPT-4o を利用しています。なお，Azure OpenAI Service は，データ提供・監視のオプトアウトが可能なため，顧客データなどの機密情報を守りながら利用することができます。

検索処理では，Azure AI Search を使って全文検索を実行しています。Azure AI Search には，事前に，サポート履歴や製品説明書，土木分野の要領・基準等，合計 900MB 強のデータを投入しています。より詳細には，上記データよりテキストのみ抽出し，チャンクと呼ばれる文の単位に分割した上で登録しています。

3. 事前検証

実業務に適用する前に，システムが実用できる水準にあるかを判断することや性能改善のために，定量的に回答性能を評価する方法を確立し，事前検証を行いました。

評価方法としては，評価データとして質問とその理想の回答，関連情報の 3 つ組を 100 件作成し，各質問に対してシステムが出力する結果を人手で点数化する手法を採用しました。なお，評価データは，サポート履歴を参考に，関連情報が被らないように作成しました。

事前検証の結果を**表 1** に示します。通常の RAG に基づいて構築した初期バージョン v1 は実用困難な回答精度でした。この結果を分析し，回答精度に大きく寄与するのは検索精度であると仮説を立て，データ追加，HyDE 実装，チャンク分割の工夫を行って構築した v4 では，検索精度，回答精度ともに改善し，実用可能な水準となりました。なお，v4 では Few-Shot Learning [3]などのプロンプトチューニングを取り入れることで，根拠の表示ができるようになるなど回答品質の向上も見られました。

表 1　事前検証の結果

評価観点	システムバージョン	評価精度（得点÷満点）
質問に対する適切さ	v1	52%
	v4	68%
検索結果の充足性	v1	29%
	v4	64%

4. 実業務での試験運用

2024 年 7 月より構築したシステムをオペレーションパートナーズ㈱（以下，OP）で試験運用しており，回答の正誤の二値評価とその評価理由を集計しています。**表 2** に 2024 年 9 月までの評価結果を示します。事前検証と比較して回答精度が著しく下がっていますが，この理由は，評価データにおいて，直接的な答えが関連情報として既に含まれている質問が多かったため，事前検証での精度が高く出やすかったからであると考えられます。

表 2　2024 年 9 月までの回答の評価結果

誤り	正しい	合計
83	46	129
64.34%	35.66%	100.00%

誤りの代表的な要因を集計した結果を**表 3** に示します。

表 3　誤りの代表的な要因

データの問題	専門用語の無理解	ハルシネーション
50	26	23

※1つの誤りにつき要因は複数あるため合計しても83件にはならない。

要因の大半を占めるデータの問題は，質問に対応する関連情報がない問題ですが，これは，適切な回答をシステムに追加登録することで解決できます。

次に，専門用語の無理解は，LLM が専門用語の意味を理解できずに無関係な回答をしたり，適切に言い換えできずに検索が失敗したりする問題です。この改善策のひとつは，意味的類似度での検索を実現するベクトル検索の導入です。実際，ベクトル検索は言い換えに頑健です。例えば，「プロット」で検索すると「作図」を含むデータを抽出できます。また，調整により専門用語の意味を正しく捉えて検索できるため，LLM に専門用語の意味を伝える関連情報を適切に取得できるようになります。

最後に，ハルシネーションは，LLM がもっともらしい誤情報を生成する問題です。特に，回答に役立つ関連情報がない場合に，造語を使って無理やり回答するケースが多くありました。故に，検索結果の質問との関連度での絞り込みと，回答不能なときにその旨を答えさせるプロンプトチューニングで改善可能と考えられます。

一方，明確に正しいと評価された回答については 46 件中 34 件が Q&A 形式のデータに基づくものだったため，Q&A の拡充が性能向上に繋がる可能性が示唆されます。

5. おわりに

実業務での試験運用の結果（**表 2**）は，満足の行くものではありませんでしたが，前章で考察した通りに，精度を上げる目処は立っており，適切な改修と適切な関連情報を追加することで実用的なシステムになることは確実です。また，OP へのヒアリングから，文章作成の省力化，原因調査の高速化など，業務負荷の軽減に繋がることもわかっています。さらに，根拠付きの回答が得られることは経験の浅い担当者の教育にも有効と思われます。

以上から，RAG に基づくシステムは，ユーザサポート業務のような，ドメイン固有の知識が必要であり，比較的多数の要員が関与し，運用中に関連情報を蓄積可能である業務への適用効果が高いと結論付けられます。

末筆ながら，OP の皆様のご協力に感謝申し上げます。

参考文献

1) P. Lewis, et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks, NIPS'20, article 793, pp. 9459–9474.
2) Luyu Gao, et al.: Precise zero-shot dense retrieval without relevance labels, ACL'23, Vol.1, pp. 1762–1777.
3) Brown, T. B., et al.: Language Models are Few-Shot Learners, NeurIPS 2020, Vol.33, pp. 1877–1901.