検索拡張生成（RAG）で実現する生成 AI 型チャットボット導入に向けた取組

本間 隼人・日本女子大学 管理部 システム課
hommah@atlas.jwu.ac.jp
北 真一,松下 有稀,高島 咲帆・日本女子大学 管理部 システム課
長谷川 治久・日本女子大学 理学部

１．はじめに

近年、生成 AI はその応用範囲の広さより注目を集め、教職員の研究・教育・業務への生成 AI の適用は、競争力強化に資すると考える。本学においても、生成 AI の利活用環境の整備が、重要な課題であった。

一方、生成 AI サービスは発展の過渡期であり、本学の組織での利用の際には、コストやセキュリティの面に解決すべき課題が複数存在した。

各種課題に対処した、本学専用の生成 AI サービスを内製、全教職員へ提供し、一定の成果を上げた。

また、現在、生成 AI の応答範囲を組織内の情報に拡大する技術、検索拡張生成（RAG：Retrieval Augmented Generation）が注目されている。

本学では、本技術を適用した生成 AI 型チャットボットを内製開発、従来のシナリオ分岐型チャットボットと比較し、質問に柔軟な応答を実現した。

２．本学専用の生成 AI サービス

本学の「全教職員の生成 AI 利活用環境を整備する」という課題に対し、本学専用の生成 AI サービス（以下、JWU-GPT：Japan Women's University－Generative Pre-trained Transformer）を、2023 年 1 月よりトライアル運用、2024 年 5 月より本運用を開始した。

２．１．JWU-GPT 導入の背景

目的の達成には、「①利用コスト」「②ユーザー管理」「③意図しない情報流出」の課題を解決する必要があった。

①利用コスト：約 5,040 万円/年
（=240 ドル/人・年※1×約 1,400 人×約 150 円/ドル）

※1:OpenAI 社 ChatGPT Plus の場合（2024 年度 8 月時点）

②ユーザー管理：検討当初、生成 AI サービスは個人契約のみであったため、システム管理者側でのユーザーの権限や機能制限の制御が不可能であり、ユーザーが誤った利用をするリスクが懸念された。

③意図しない情報流出：生成 AI サービス利用時に入力した情報は、サービス提供事業者側に保存され、LLM（Large Language Models、大規模言語モデル）の学習に利用されるリスクが存在した。また、学習利用のオプトアウトは可能だが、②の理由より、ユーザー側に依存する点も課題であった。

２．２．JWU-GPT のシステム構成

JWU-GPT のユーザーインターフェースは、本学で導入済みの Microsoft 社の Teams を採用した。Teams 上に同社 Copilot Studio（旧 Power Virtual Agents for Teams）で構築したエージェントを展開、本エージェントは、API 経由で OpenAI 社の GPT-4 を LLM とした応答を生成する（①）。

コスト管理を目的に、入出力トークン数（生成 AI が処理する基本単位文字数）のみ、データベースに格納する。Microsoft 社 Power BI により、利用状況をダッシュボードで確認が可能である（②）。

また、本学の教職員・学生アカウントを管理する AD（Active Directory）は、AADC（Azure AD Connect）により、Entra ID（旧 Azure AD）に同期している。Entra ID のセキュリティグループ（例.教員グループ等）に対して JWU-GPT の利用権限を付与する（③）。

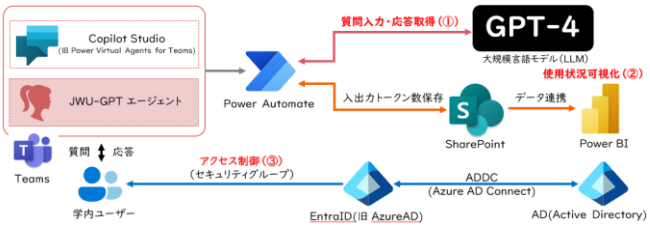

図 1. JWU-GPT のシステム構成概要

ユーザーは、Teams 上の JWU-GPT とのチャット画面から、質問を入力し、GPT-4 の応答を取得できる。

さらに、同一トピックに関する質問は 5 回まで継続が可能である。本機能は、API で GPT-4 をコールする際に、会話履歴を入力プロンプトに投入することで実現している。

図 2. JWU-GPT のユーザー画面

２．３．JWU-GPT の利用実績および結果

2024 年 1 月～7 月末の JWU-GPT の月別利用実績を図 3 に示す。全教職員への導入の本運用を開始した 2024 年 5 月以降、大幅に利用が増加している。合計で約 548 万トークン（≒文字）の入出力を確認し、

「全教職員の生成 AI 利活用環境を整備する」目的の達成に寄与したと考える。

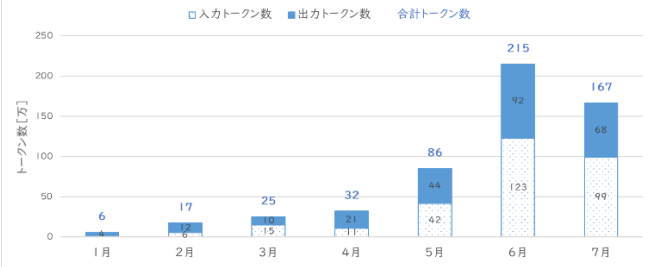

図 3. JWU-GPT の月別利用実績（2024 年 1 月～7 月）

また、「①利用コスト」「②ユーザー管理」「③意図しない情報流出」の課題は、次の結果となった。

①利用コスト※2：約 2 万円/年

（キャッシュアウト抑制：5,038 万円/年）

※2: 2024 年度 8 月時点の gpt-4-o の料金、8 月～12 月は 5 月～7 月の平均値を利用

API は入出力トークン数による従量課金のため、利用実態に応じた利用コストとなった。

②ユーザー管理：システム管理者側で本学の教職員アカウントが所属するセキュリティグループに対して権限を付与、ユーザー側の設定変更はできない。

③意図しない情報流出：API 経由の入出力はサービス事業者の LLM の学習に利用されず、情報流出リスクがない。

3.生成 AI 型チャットボット

生成 AI の応答を組織内の情報に拡大する技術である、検索拡張生成（以下、RAG：Retrieval Augmented Generation）を活用し、本学運用に特化した JWU-GPT の機能追加に着手した。本機能は、学生・教職員からの問合せの一次窓口を想定する。従来の問合せ対応の工数を削減、および、時間・場所を問わない迅速な回答によるサービスレベルの向上が期待される。

3.1.RAG による応答

RAG では、学生・教職員からの問合せの回答元となるドキュメントを加工し、ベクトルデータベースに事前に格納する。生成 AI へ問合せに対し、ベクトルデータベースから類似情報を検索・選択、入力プロンプトに情報を追加する。この追加した類似情報を基に回答を生成することで、組織内の情報に関して回答が可能となる。

しかしながら、RAG では、入力プロンプトに情報を追加する工程による、入力トークン数の肥大に留意する必要が生じる。

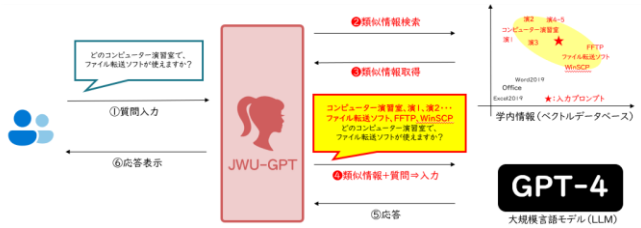

図 3. RAG の概念図

3.2.生成 AI 型チャットボットのシステム構成

生成 AI 型チャットボットの基本構成は、JWU-GPT を活用し、RAG による応答モードをユーザー側で選択する方法とした。前述の仕組みにより、本学の情報に関する応答を取得できる。

検索対象となるベクトルデータベースは、学内ファイルサーバに保存した回答元となるドキュメントを保存・加工し、投入する。また、高効率なサービス運用も視野に、学内サーバ上のドキュメントが更新されるとベクトルデータベース上の情報も更新される仕組みを構築している。

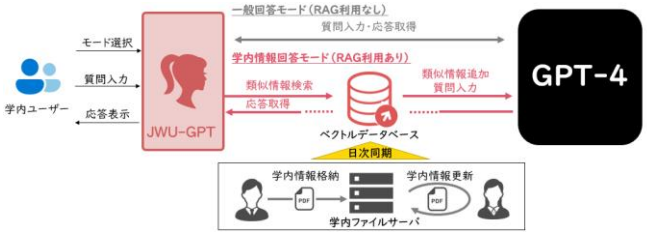

図 4. 生成 AI 型チャットボットのシステム構成概要

3.3. 生成 AI 型チャットボットの検証および結果

検証対象として、本学の ICT 関連相談担当のメディアセンターの QA 履歴（2023 年 4 月～2024 年 7 月）及び、メディアセンターHP の情報をベクトルデータベースに格納し、応答の検証を行った。

図 5. 生成 AI 型チャットボットの応答一例

結果、良好な応答精度であった。さらに、RAG での入力トークン数の肥大は、入力文字数に対し、約数千～1 万倍と推定されたが、実運用に問題がないコストと判明した。上記の結果を踏まえ、メディアセンターのスタッフ向けに先行運用を開始した。

3.4. 生成 AI 型チャットボット展開に向けた動き

実運用に向け、2024 年 9 月より各課職員よりサービス構築プロジェクト体制を確立する。回答元となるドキュメント準備及び、想定質問と回答の評価を実施する。さらに、評価結果に基づき、各種設定パラメーター調整やドキュメント更新を実施し、2025 年度 1 月にトライアル運用を開始する予定である。

さらに、学生サービスとしての展開を見据え、学生目線での評価および開発にも注力する。

謝辞

本サービスの開発及び導入にご協力を賜りました長谷川治久教授、管理部システム課の皆様、教職員の皆様に感謝申し上げます。