検索拡張生成（RAG）を備えた 生成 AI 対話プラットフォームの性能評価に基づく最適化と全学展開

本間 隼人[1)3)],北 真一[1)3)],松下 有稀[1)3)],高島 咲帆[1)3)],長谷川 治久[2)3)]

1) 日本女子大学 管理部 システム課
2) 日本女子大学 理学部 数物情報科学科
3）日本女子大学 メディアセンター

hommah@atlas.jwu.ac.jp

Optimization and Campus-wide Deployment of a Generative AI Dialogue Platform with Retrieval Augmented Generation (RAG) Based on Performance Evaluation

Hayato Homma[1)3)], Shinichi Kita[1)3)] , Yuki Matsushita[1)3)] , Sakiho Takashima [1)3)]
Haruhisa Hasegawa[2)3)]

1) Information Technology Division, Management Department, Japan Women's Univ
2) Dep. of Mathematics, Physics and Computer Science,Faculty of Science, Japan Women's Univ
3)Media Center, JapanWomen’s Univ.

概要

日本女子大学では、教職員の研究・教育・業務における生成 AI 活用環境の整備を目的に、本学専用の生成 AI 対話プラットフォーム「JWU-GPT」を内製開発した。Microsoft 365 および ChatGPT API を基盤とし、低コスト運用、学内アカウントによるアクセス制御、入力データの学習利用防止を実現している。更に、検索拡張生成（RAG）を活用した学内情報検索機能を追加開発し、20 組織による評価・改善の結果、良好な応答が 90%を超え、全教職員に展開した。

1. はじめに

生成 AI は、その応用範囲の広さから大きな注目を集めており、教職員の研究・教育・業務への適用は、組織の競争力強化に資すると考える。

日本女子大学（以下、「本学」）においても、生成 AI の利活用環境の整備は重要な課題であり、2023 年度より本格的な取り組みを開始した。

しかし、検討開始当初は生成 AI サービスの学内での利用に際してはコストやセキュリティなど、解決すべき課題が複数存在した。これらの課題に対応するため、本学専用の生成 AI 対話プラットフォームを内製し、全教職員に提供した。

さらに、2024 年度には、生成 AI の応答範囲を組織内情報へ拡張する技術である検索拡張生成（RAG: Retrieval-Augmented Generation）が注目を集めた。この技術を適用した生成 AI 対話プラットフォームの機能追加について、組織的な精度評価および改善を実施したうえで、サービスとしてリリースした。加えて、本機能に関しては運用工数のゼロ化を達成するとともに、応答範囲の拡大を容易にする仕組みも構築した。

2025 年度には、本サービスが安定的に運用されていることから、本稿ではその過程や成果、導入効果の振り返りについて報告する。

２．日本女子大学の生成 AI 対話プラットフォーム

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 対話プラットフォーム（以下、JWU-GPT：Japan Women's University - Generative Pre-trained Transformer）を、2024 年 1 月に内製開発、2024 年 2 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 導入の検討背景

2023 年度の検討当初、生成 AI サービスは発展の過渡期であり、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あり、内製開発で本学専用の生成 AI 対話プラットフォームである JWU-GPT を構築する方針とした。

①利用コスト

OpenAI 社 ChatGPT Plus の場合、約 5,040 万円/年※のコストが試算された。

※240 ドル/人・年×約 1,400 人×約 150 円/ドル

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、学習利用の拒否は可能だが、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 のシステム構成要素

JWU-GPT の内製開発で利用した構成要素を表 1 に示す。

アイコン	構成要素	概要
	Teams	生成 AI 対話の ユーザーインターフェース
	Copilot Studio	JWU-GPT のエージェントを構築
	Power Automate	ユーザーの入力を生成 AI に受渡、 生成 AI の出力をユーザーに返す 処理を自動化
	ChatGPT API	OpenAI 社の ChatGPT の LLM に接続 するためのインターフェース
	Share Point	入出力トークン数 を格納するデータベース
	Power BI	利用状況を確認する ダッシュボード
	AD (Active Directory)	学内構成員のアカウントを 管理するシステム
	AADC (Azure AD Connect)	AD と EntraID のアカウント情報を 同期する機能
	Entra ID	クラウド上に存在する 学内構成員のアカウントを 管理するシステム

表 1 ：JWU-GPT の構成要素

2. 3 JWU-GPT のシステム構成

JWU-GPT のシステム構成の概要を図 2 に示す。

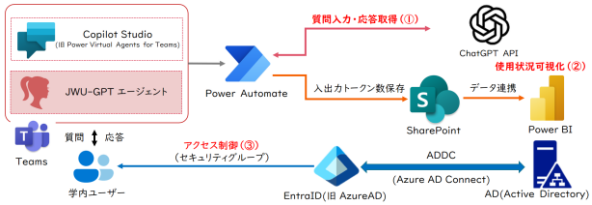

図 1 ：JWU-GPT のシステム構成概要

JWU-GPT のユーザーインターフェースは、本学で導入済みの Teams を採用した。Teams 上に 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、入力を受け取ると、Power Automate を介し、API 経由で OpenAI 社の GPT を大規模言語モデル（以下、LLM）とした応答を生成する（①）。

図 2：JWU-GPT の応答

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）を SharePoint に格納する。本データを用いた 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）。

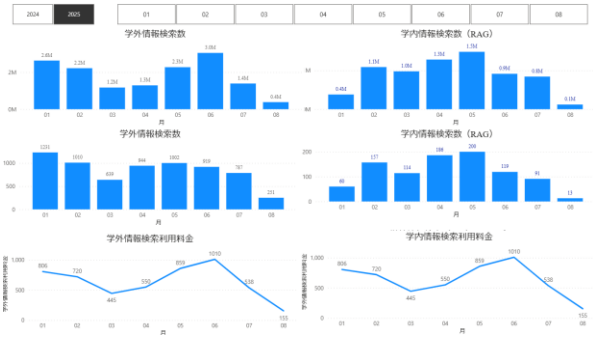

図 3 ：JWU-GPT のダッシュボード

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 で管理されている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）。

2.4 JWU-GPT の特徴

JWU-GPT は次の 3 つの特徴を有することから、当初の課題を解決することができている。

①API 利用による低コストかつ柔軟な運用

JWU-GPT は、すでに導入済みの Microsoft 365 の各種サービスと ChatGPT の API を基盤として構築している。ChatGPT の API は利用者数に依らない入出力トークン数に基づく従量課金方式を採用し、低コストでの運用が可能である。

さらに、ChatGPT の API は常に最新の LLM が提供され、機能向上が図られており、柔軟に活用できることも大きな利点である。コスト、機能向上の観点による ChatGPT の LLM の変更履歴と 100

万トークン当たりの入出力単価は表 2 である。

利用期間	LLM	入力	出力
2024 年 1 月 ～2024 年 5 月	gpt-4-turbo	$10.00 (1,500 円)	$30.00 (4,500 円)
2024 年 5 月 ～現在	gpt-4o	$2.5 （375 円）	$10.00 （1,500 円）
2025 年 2 月 ~2025 年 4 月	o3-mini	$1.10 （165 円）	$4.40 （660 円）
2025 年 3 月 ~現在	gpt-4o-mini search-preview	$0.15 （22 円）	$0.60 （90 円）
2025 年 4 月 ～2024 年 8 月	4.1-mini	$0.40 （60 円）	$1.60 （240 円）
2025 年 8 月 ～現在	5-mini	$0.25 （38 円）	$2.00 （300 円）

表 2 ：ChatGPT の LLM 利用履歴と単価

②セキュリティ管理

JWU-GPT は、Entra ID の情報を用いてアクセス制御を行っている。これにより、本学の教職員アカウントのみに JWU-GPT へのアクセスを限定し、不正利用を防止している。

③学習データへの利用防止

ChatGPT を API 経由で利用する場合、入力データは LLM の学習データとして使用されない。このため、個人アカウントによる ChatGPT の業務利用と比較して、情報漏えいリスクを低減し、安全性を確保できる。

2.4 JWU-GPT の提供機能

2025 年 8 月時点において、JWU-GPT は「基本応答」、「学内情報検索」、「Web 情報検索」の 3 つの機能を提供している。

なお、本稿では、「学内情報検索」機能のリリースまでの取組みについて、第 3 章で詳述する。

機能	処理	提供開始
基本応答	簡単なクリエイティブ作業に 最適な機能	2024 年 5 月～ ※2024 年 1 月 ～トライアル
学内情報 検索	本学の内部的な手続き等に関する 問い合わせに最適な機能	2025 年 2 月～
Web 情報 検索	Web 検索の用途に最適な機能 最新の Web 情報から応答生成し、 出典のリンクを表示	2025 年 5 月～

表 3：JWU-GPT の機能

3．検索拡張生成（RAG）を活用した学内情報検索機能の追加

生成 AI の応答範囲を組織内情報に拡張する技術の検索拡張生成（RAG：Retrieval-Augmented Generation）を活用し、本学の運用に特化した JWU-GPT の機能追加に着手した。

本機能は、教職員からの問合せに対応する一次窓口としての利用を想定している。従来の問合せ対応にかかる工数の削減に加え、時間や場所を問わず迅速に回答できることから、サービスレベルの向上が期待される。

3.1 検索拡張生成による応答

検索拡張生成では、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し、表 4 の順で学内情報に関する応答が可能になる。

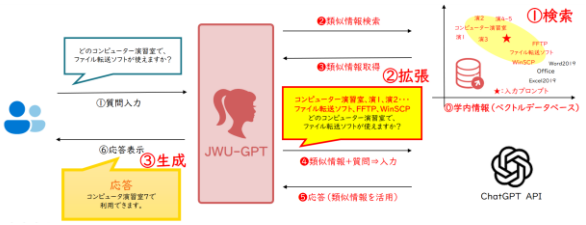

図 4：検索拡張生成による応答概要

項番	手順	処理内容
①	質問入力	利用者が生成 AI に質問を入力する 例：「コンピューター演習室で使えるソフトは？」
②	類似情報検索	ベクトルデータベースから類似度検索
③	類似情報取得	検索結果の関連度の高い情報を取得
④	プロンプト 拡張	取得した類似情報を入力プロンプトに追加
⑤	回答生成	生成 AI が拡張プロンプトを基に回答

表 4：検索拡張生成による応答手順

利用者の問合せに対して、類似情報を検索・取得し、入力プロンプトに追加することで、生成 AI は学内情報に基づいた回答を提供できる。

図 5：検索拡張生成による応答例

3.2 学内情報検索機能の評価検証方法

学内情報検索機能の構築にあたり、本学の 20 組織（以下「参画組織」）による以下の手順で評価検証を行った。「④定量評価の入力」では、想定質問に対する生成 AI の応答を評価する。評価基準として、質問者の疑問を解決できれば「適切な回答（◎）」、質問者の疑問を解決できるが、さらに丁寧な説明があるとよい場合は「参考となる回答（〇）」という評価基準で参画組織が自ら評価した。

項番	手順	処理内容	対応組織
①	想定質問例の作成	学内利用を想定した質問例を作成	参画組織
②	データ格納	想定質問に回答するための 参考情報を共有フォルダに格納	参画組織
③	応答の自動生成	全想想定質問に対して応答を生成	システム課
④	定量評価の入力	応答内容を 5 段階で評価 ◎：適切な回答 〇：参考となる回答	参画組織

		△：誤りを含む回答 ▲：参考にならない・ 誤った回答 ×：問題を含む回答	
⑤	設定の見直し	評価結果に基づき設定を調整 ・プロンプト ・各種パラメーター ・格納データ	参画組織 システム課
⑥	再検証	③、④を再度実施	参画組織 システム課

表 5：応答の検証と評価方法

3. 3 学内情報検索機能の評価及び改善

適切な回答（◎）・参考となる回答（〇）が全応答の 80%以上をリリースの目標値として定め、検証を開始した。 1 回目の結果を表 6 に示す。

評価	件数	割合	評価値
◎ 適切な回答	77	51%	**76.8%**
〇 参考となる回答	39	25.8%	
△ 誤りを含む回答	25	16.6%	(23.2%)
▲ 参考にならない 誤った回答	6	4.0%	
× 問題を含む回答	4	2.6%	
合計	151	100%	-

表 6：検証 1 回目評価結果

評価結果の詳細分析により、生成 AI が臨機応変に対応しようとする傾向が、回答内容の曖昧さを生み出していると判明した。この結果を踏まえ、曖昧さや自由度の高い応答を抑制し、正確性を重視する方向で以下の調整を実施した。

調整項目	詳細
プロンプト調整	プロンプトに「日本女子大学の学内情報を用いて正確性を重視し答えて下さい。」と追加
temperature 調整	1 → 0.5 に変更し、 出力の安定性と正確性を重視
回答元データ 差し替え	不正確または不足していた情報を更新

表 7：調整内容

調整による、具体的な応答の変化の例を次に示す。調整前は、一般的な想定や関連部への問い合わせを推奨する内容であったが、調整後は、本学の規約に準じた内容での回答が得られるようになっている。

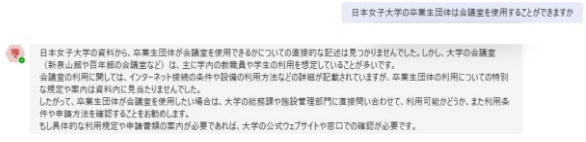

図 6-1.調整前の応答結果

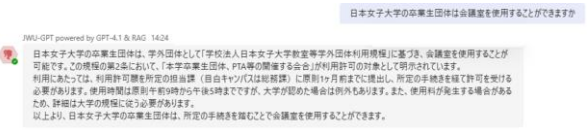

図 6-2.調整後の応答結果

調整後、検証の 2 回目を実施した。その際の評価結果を表 8 に示す。

評価	件数	割合	評価値
◎ 適切な回答	95	57.6%	**90.3%**
〇 参考となる回答	54	32.7%	
△ 誤りを含む回答	12	7.3%	(9.7%)
▲ 参考にならない 誤った回答	2	1.2%	
× 問題を含む回答	2	1.2%	
合計	165	100%	-

表 8：検証 2 回目評価結果

結果として、「適切な回答」と「参考となる回答」の合計割合は 90.3% に達し、目標値を達成したため、全教職員への展開を実施した。

3. 4 学内情報検索機能の運用改善

検証結果から、学内情報検索機能の応答は良好であることが確認された。しかし、運用において精度を維持、さらにサービスレベルを向上させるためには、「回答元データの最新化」および「回答内容の充実」が課題として残された。本課題に対して、以下の取組を実施した。

①ベクトルデータベースの日次更新

応答の精度を保つには、回答元となるデータを常に最新化する必要がある。ベクトルデータベースの更新を自動化し、サービスのダウンタイムを発生させない方式を採用した。

具体的には、既存のベクトルデータベースから古いデータを削除・追加する方法ではなく、新規にベクトルデータベースを作成し、共有フォルダにアップロードされた情報を日次で同期する。検索は常に新しいベクトルデータベースを参照する仕組みとした。

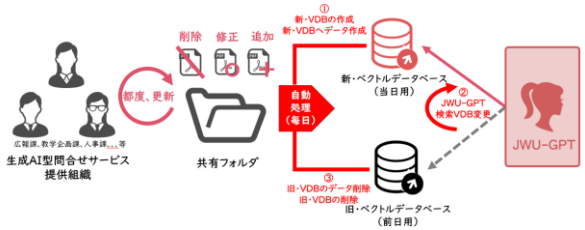

図 7：ベクトルデータベースの日次更新概要

この仕組みにより、システム管理者側での追加作業は不要となり、参画組織は共有フォルダ上のデータを更新するだけで、翌日には最新情報に基づく回答が可能となった。

②検証モードの提供

サービスレベルを向上させるには、回答可能な質問の範囲を拡大する必要がある。そのため、応答の検証モードを提供した。

検証結果が良好であった場合には、所定の共有フォルダにデータを格納するのみで、翌日には本サービスから回答を得ることが可能となる。

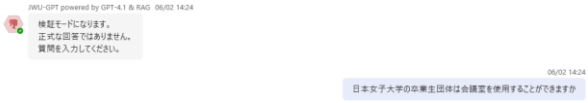

図 8：検証モード

４．JWU-GPT の定量的評価

JWU-GPT は、2024 年 1 月にトライアルを開始し、2024 年 4 月から本運用を開始している。トークン数（≒利用文字数）及び利用件数での定量評価を行う。

＜トークン数＞

全教職員への導入の本運用を開始した 2024 年 5 月以降、大幅に利用トークン数が増加している。さらに、月ごとに増減はあるものの利用の増加傾向があることがわかる。

合計で約 4,700 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。

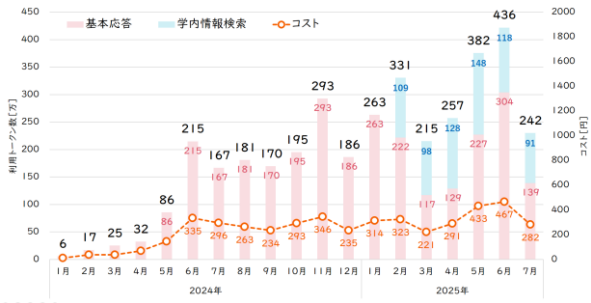

図 9 ：JWU-GPT の月別利用トークン数

また、月平均利用トークン数は、約 200 万トークンの利用となり、コストは月 500 円程度で抑えることができている。

＜利用件数＞

利用件数もトークン数と同様の傾向であるが、学内情報の検索を開始した 2025 年 2 月から大幅に件数は伸びてはいない。これは、潜在的に学内情報への検索のニーズがあり、従来の基本応答の利用から学内情報への利用がシフトしたと考えられる。

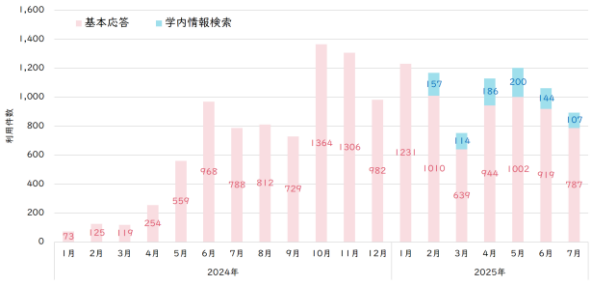

図 10：JWU-GPT の月別利用件数

また、学内情報検索を教職員の問合せの削減として計算した場合、これまでに約 900 件（約 150 時間）の学園全体の問合せ稼働を削減に貢献したと考える。

５．おわりに

本稿では、日本女子大学における生成 AI 対話プラットフォーム「JWU-GPT」の技術的な内容と検索拡張生成（RAG）を活用した学内情報検索機能の開発・評価・改善の過程を報告した。

JWU-GPT により低コストかつ柔軟な運用、学内アカウントによる厳格なアクセス制御、入力データの学習利用防止などの特徴を備えた生成 AI 対話プラットフォーム提供を実現した。

検索拡張生成の技術活用により、学内情報に基づく正確かつ迅速な応答が可能となり、問合せ対応の削減とサービスレベルの向上が確認された。

今後は、利用データの分析に基づく精度向上や、新たな学内システムとの連携による機能拡張、学生への拡大など、さらなる発展を図る予定である。

６．謝辞

本サービスの開発及び導入にご協力を賜りました参画いただいた教職員の皆様に感謝いたします。

参考文献

[１]本間隼人、北真一、松下有稀、高島咲帆、長谷川治久、「検索拡張生成（RAG）で実現する生成 AI 側チャットボット導入に向けた取組」、公益財団法人私立大学情報教育協会 令和 6 年度私情報教 教育イノベーション大会、204 ページ、2024