中小企業での生成 AI の活用に向けた基礎的検討
― マルチモーダル RAG の構築と評価 ―

千家　雅之（情報・生産技術部システム技術グループ）

1.　はじめに

近年，生成 AI の業務での活用は急速に広まっている．しかしながら，生成 AI の利用においては，大規模な計算資源が必要なことから，大手クラウドサービス事業者が提供する生成 AI サービスを利用する形態が主であり，このようなサービスは社外へのデータ送信を伴うため，情報セキュリティの観点から導入に慎重な企業も少なくない．そのため，ローカル環境で動作する生成 AI にも多くの関心が集まっており，自社内に AI サーバーを導入する動きが広がりつつある．そこで筆者は，中小企業への導入を想定して，小規模計算環境におけるローカル AI の実用可能性について考察するために，一例として社内文書に対する質問応答を行う AI チャットボットをローカル環境で構築して評価した [1]．人手評価による正答率は 84%であり，正確性が求められるユースケースには適用が難しいという結論であった．本稿では，近年急速な発展がみられるマルチモーダル LLM (Large Language Model)の一種である VLM (Vision Language Model) とマルチモーダル RAG (Retrieval-Augmented Generation) を用いて同様の評価を実施し，回答精度の向上を検証する．

2.　マルチモーダル RAG

2.1　概要

RAG は，計算負荷が大きい再学習やファインチューニングを行わずに，LLM 等に新たな知識を与える方法であり，外部データベースから関連情報を取得して回答を補強する技術である．マルチモーダル RAG は，テキスト以外のモダリティ（画像や音声等）にも対象を広げたものであり，複数の実現方法がある．本節では代表的なものについて 3 点紹介する．

(a) 全てのモダリティをテキストのモダリティに変換する方法である．テキスト以外のモダリティをマルチモーダル LLM に解釈させてそれをテキストにし，テキストのベクトル空間で検索可能にするものである．しかし，この方法では，画像や音声等が持つ細部の情報は失われてしまう．

(b) 上記の弱点を克服するものであり，ベクトルデータベースにテキストだけでなくテキスト以外のモダリティを参照可能にするメタデータも同時に蓄積し，検索結果とともにテキスト以外のモダリティを参照し，マルチモーダル LLM に与える方法である．しかし，テキスト以外に画像や音声等を保存し，互いに関連付ける必要があり，管理が煩雑になるというデメリットがある．

(c) 画像や音声をテキストと同一のベクトル空間に埋め込む Embedding モデルを使用する方法である．さらにこの方式は細分化できて，画像や音声等をマルチベクトルとして扱うのか，画像や音声等を表す代表ベクトルをつくるのかで方式が分かれる．マルチベクトル方式は Embedding 時の計算負荷が高いが，元データの情報をより多く保持できる．

それぞれの方法には一長一短があり，ユースケースに応じた使い分けが求められる．

2.2　ColPali [2]

ColPali は，文書ページの画像をパッチ（小領域）に分割し，Attention 計算によりパッチ間の文脈を反映したマルチベクトル Embedding を生成する VLM であり，前節(c)のマルチベクトル方式に該当する．クエリとの照合には MaxSim 計算が用いられる．これは，クエリの各トークンに対して全パッチベクトルとの類似度の最大値を取り，その総和をスコアとするものであり，スコアが高い順に検索結果として返される．

3.　AI チャットボットシステムの構成

本稿では，画像情報が失われにくい ColPali 系モデルを採用し，図 1 に示すシステムを表 1 のハードウェア構成で構

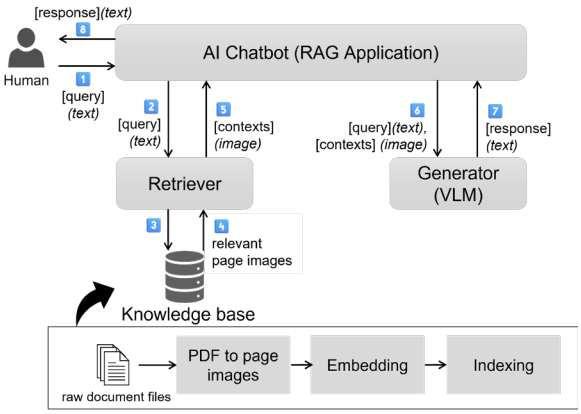

図 1　マルチモーダル RAG の構成

表 1　ハードウェア構成

ハードウェア (Retriever/ Knowledge base)	CPU	Intel Xeon W-1390P
	RAM	64GB
	GPU	NVIDIA RTX 6000 Ada (48GB)
	OS	Ubuntu 24.04 LTS
ハードウェア (Generator)	CPU	AMD Ryzen Threadripper PRO 7955WX
	RAM	128GB
	GPU	NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (96GB) ×2
	OS	Ubuntu 24.04 LTS

築した．ユーザの質問は AI Chatbot を介して Retriever に渡され，Knowledge base から取得したページ画像とともに VLM に入力され，回答が生成される．VLM の推論は多くの VRAM を消費するため, 検索用と生成用で異なるハードウェアを用いた．VLM には，日本語性能が優れているとされる Qwen3.5-122B-A10B-FP8 を採用し，推論エンジンに vLLM 0.20.2 を使用した．検索モデルには ColPali 系モデルである tomoro-colqwen3-embed-8b を採用し，ラッパーライブラリである Byaldi 0.0.7 をカスタマイズして用いた．ナレッジベースの構築は，元となる PDF ファイルを Byaldi に読み込ませるのみであり，前報 [1] で必要であったテキスト抽出, Markdown 整形, チャンク分割等の前処理は一切不要であった．検索時には上位 3 件のヒットページに加え，前後各 1 ページも取得して VLM に入力することで，ページをまたぐ図表にも対処した．

4. 評価

4.1 評価方法

人手で 50 問の質問とその答え（基準回答）を用意した．50 問には，表の読み取りに関する質問 5 問，ページをまたぐ内容に関する質問 5 問を含めた．RAG の性能評価を，人手と自動評価ツール Ragas[3] を用いて行った．

人手による評価は，基準回答と AI の回答を比較し，表 2 の評価基準に従って点数付けを行った．Ragas による評価では, Context recall（基準回答に含まれる情報が検索結果にどれだけ含まれているかを測る指標），Faithfulness（AI の回答が検索結果の情報を正しく反映しているかを測る指標），Factual correctness (precision)（AI の回答が基準回答と矛盾していないかを測る指標）を算出した．評価用 LLM と評価用 Embedding モデルには，Azure OpenAI サービスで提供されている GPT-4.1，text-embedding-ada-002 をそれぞれ使用した．なお, マルチモーダル RAG の検索結果はページ画像であるため, Ragas による評価には VLM を用いて画像からテキストを抽出したものを使用した．

比較対象として, 前報で構築したテキストベース RAG を用いた．比較の公平性を確保するため, 生成用 LLM は前報の Llama-3.3-Swallow-70B-v0.4 からマルチモーダル RAG と同じ VLM に変更し，同じ 50 問で再評価した．

4.2 評価結果

表 3 に人手による評価結果と Ragas v.0.4.3 を用いた自動評価の結果を示す．マルチモーダル RAG での人手評価では，スコア 4 および 5 を正答とみなした場合の正答率は 96%であった．全問で回答内容は正確であったが，条文番号の引用に誤りが 2 件見られた．一方, テキストベース RAG では 88%であり，約 12%には，適切な検索結果を取得できなかったケースや LLM が検索結果から適切な回答を生成できなかったケースが含まれていた．Ragas による評価では，Context Recall，Faithfulness，Factual Correctness のいずれにおいてもマルチモーダル RAG がテキストベース RAG を上回っており，ページ画像から抽出したテキストに基づいて算出しているにも関わらず，自動評価においてもマルチモーダル RAG の優位性が示された．

テキストベース RAG の不正解 6 件を分析すると，特に図表に関する質問では，検索に成功しているにも関わらず正しい回答を生成できないケースが確認された．これは，表をテキストに変換した際に,「同」のような上のセルを参照する省略表現の意味が失われ, LLM が表の内容を正しく解釈できなかったことが原因と考えられる．マルチモーダル RAG では，ページ画像をそのまま VLM に入力するため，表の空間的な構造が保持され，このような省略表現も正しく解釈された．また，検索においても，マルチベクトル方式はパッチ単位でクエリとの照合を行うため，表中の特定の情報に対する検索精度がテキストベース RAG のチャンク単位の検索より高いことが示された．なお，生成用 LLM は両方式とも同一のモデルを使用しており, 回答精度の差は入力形式と検索方式の違いに起因する．

表 2　人手評価の評価基準

スコア	評価基準
5	正確かつ基準回答と比べ過不足がない
4	正確だが基準回答と比べ軽微な過不足がある
3	正確だが基準回答と比べ多くの欠落がある
2	一部に不正確な内容が含まれている
1	大部分が不正確である

表 3　評価結果

		テキストベース RAG	**マルチモーダル RAG**
人手	平均スコア	4.600	**4.780**
	評価 4 と 5 の割合	0.880	**0.960**
Ragas	Context Recall	0.970	**1.000**
	Faithfulness	0.949	**0.967**
	Factual Correctness (precision)	0.765	**0.858**

5. おわりに

本稿では，ColPali 系マルチベクトル方式によるマルチモーダル RAG を用いた AI チャットボットを構築し，前報のテキストベース RAG と同一の社内文書・同一の評価体系で比較評価した．テキストベース RAG で必要であったテキスト抽出やチャンク分割等の前処理を一切行わずに，人手評価で 96%の正答率を達成し，テキストベース RAG の 88%を上回った. これらの結果は 122B クラスのモデルで得られたものである．本方式は回答精度が VLM の性能に依存するため，モデルの進化に応じた精度の向上が見込まれる．今後は，対象文書数を増やし，実運用を見据えた検証に取り組む予定である．

【参考文献】

1. 千家雅之 :「中小企業での生成 AI の活用に向けた基礎的検討　―ローカル環境上で動作する AI チャットボットの試作―」, *KISTEC 研究報告 2025*　(2025)
2. Manuel Faysse, et al. : “ColPali: Efficient Document Retrieval with Vision Language Models”, arXiv:2407.01449 (2025)
3. Ragas, https://www.ragas.io/