RAG を利用した傾聴応答生成の検証

松本 奈々[1] 安藤 一秋[2]
[1]香川大学大学院 創発科学研究科 [2]香川大学 創造工学部
{s24g358, ando.kazuaki}@kagawa-u.ac.jp

概要

近年，日本の総人口に占める高齢者人口の割合は過去最高となり，要介護者数も増加している．介護現場において，高齢者の発言を傾聴することは信頼関係を築くために重要である．しかし，近年の介護士の人材不足や介護負担等から被介護者に十分な時間をかけることが困難である．本稿では，外部情報に基づき応答を生成する RAG（Retrieval-Augmented Generation）を用いて，傾聴応答の生成可能性を検証する．評価実験では，傾聴応答の生成件数や，ファインチューニングの必要性に注目して考察する．

1 はじめに

近年，日本の総人口に占める高齢者人口の割合は29.3%と過去最高を更新し続けているともに，65 歳以上人口は 3,625 万人と過去最多となっている．これに伴い，65 歳以上の要介護者数は増加しており，介護に従事する職員数も増加している[1, 2]．介護現場において，高齢者の発言を傾聴，尊重，共感することは信頼関係を築くために重要であり，一つの実現手段として「バリデーション」がある．この方法は，6 つの基本態度（「傾聴する」，「共感する」など）と，14 の基本テクニック（「はい／いいえ」で答える質問ではなく，「いつ」，「どこで」といった自由な回答が期待できる質問を心がけることなど）を用いて被介護者と丁寧にコミュニケーションすることで，被介護者が抱える悩みを軽減することを目的としている[3, 4, 5]．しかし，介護士の人材不足や介護負担等の問題から個人に十分な時間をかけてケアすることが困難である．

本研究では，介護環境の改善を目指して，バリデーションを活用した対話システムの構築を目的とする．著者らの先行研究[6]では，報酬表現に着目した強化学習による傾聴対話モデルの改善を提案した．ここでは，報酬モデル作成時の報酬を与える表現を改善することで，応答生成における内容の質が向上し，尊重性や共感性がさらに感じられるようになり，傾聴性満足度が向上できた．しかし，傾聴対話モデル構築時における基盤の言語モデルは，GPT2-medium を利用していた．近年，大規模言語モデル（LLM: Large Language Model）の発展に伴い，様々なモデルが公開されている．OpenAI 社は API を用いて利用が可能である．ELYZA 社は Meta 社の Llama3[7]シリーズをベースとして日本語を追加事前学習および事後学習した Llama-3-ELYZA-JP-70B や Llama-3-ELYZA-JP-8B を公開している[8]．また，LLM を追加学習せずに外部情報を組み込むために RAG（Retrieval-Augmented Generation）[9]も注目を集めている．これは，外部情報を入力文章と合わせて LLM に与えることで，外部情報に基づいた生成が可能になる．

本稿では，先行研究から基盤モデルを変更し，LLM を用いて RAG におけるプロンプトの工夫，RAG における検索クエリを改善した手法を用いて，傾聴応答の生成可能性を検証する．さらに，プロンプト作成時において条件を追加することによる応答生成の変化を確認する．最後に，ファインチューニングの必要性について考察する．

2 関連研究

RAG においては，ベクトル DB に格納するデータの工夫，検索前のクエリの工夫，検索時の工夫，検索後の工夫など様々な手法[10][11]が提案されている．検索前のクエリを工夫する手法としては，HyDE（Hypothetical Document Embeddings）[12]がある．これは，入力された質問に対して仮説的な応答を生成させ，その出力を検索に利用する方法である．検索後の工夫としては，RAG-Fusion[13]がある．これは，複数のクエリを生成しそれらの検索結果を RRF（Reciprocal Rank Fusion）で並べ替える手法である．

本稿では，発話-応答形式の対話を想定していることから，HyDE による仮説的な回答を生成させることで，傾聴応答生成への変化を確認する．

図 1：RAG を用いた傾聴応答生成の概要図（ベースライン）

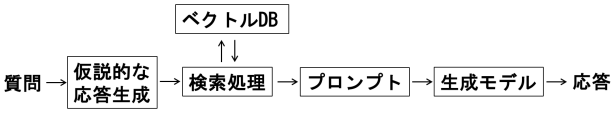

図 2：RAG の検索クエリを工夫した傾聴応答生成の概要図

3 RAG を用いた傾聴応答生成

本稿では，RAG を用いた傾聴応答生成のために主に 2 点の生成フローを作成する．1 点目は，RAG の検索クエリを工夫しない（仮説的な応答を検索前に生成させない）ものである．概要図を図 1 に示す．本稿では，これをベースラインとする．2 点目は，RAG の検索クエリを工夫したものである．概要図を図 2 に示す．以下，2 つのフローの各処理について説明する．

3.1 ベクトルデーターベース（外部情報）

本稿では，Qdrant[i]を採用したベクトルデータベース（ベクトル DB）に，傾聴発話-応答データを格納する．

3.1.1 データ構築

ベクトル DB に格納するデータの構築方法について述べる．まず，日本語日常対話コーパス[14]の各対話データを発話-応答ペアの形式に整形する．次に，先行研究[6]で定義した傾聴応答（掘り下げ質問，尊重応答，共感応答）の各表現に対して，掘り下げ質問は文末での一致，尊重応答と共感応答は文頭からの一致で機械的に判定し，先ほど整形したデータに対してラベリングする．具体的な各傾聴応答の基準例を表 1 に示す．ここで，バリデーション技法のオープンクエスチョンでは，「なぜ～ですか？」や「どうして～ですか？」のような質問は，多くの被介護者が応答に困る傾向があるため，避けたほうがよい表現とされている．したがって，掘り下げ質問に対して，これらの表現を除外している．

表 1：ラベリングの基準例

傾聴の種類	ラベリング基準
掘り下げ質問	？（なぜですか？，どうしてですか？といった 11 種類の表現以外）
尊重応答	「いいですね」，「良いですね」，「素晴らしい」，「すてきです」，「大変ですね」など
共感応答	「そうですね」，「よくわかります」など

表 2：それぞれのテーブルにおけるレコード数

傾聴の種類	レコード数
掘り下げ質問	8,992
尊重応答	562
共感応答	1,364

3.1.2 ベクトル DB へデータ格納

次に，ベクトル DB へのデータの格納について述べる．本稿では，RAG を用いて傾聴応答がどの程度生成できるか検証することを目的としているため，各傾聴応答（掘り下げ質問，尊重応答，共感応答）について，それぞれベクトル DB のテーブル（Qdrant では collection に該当）を作成し，データを格納する．格納するデータの Embedding には，kun432/cl-nagoya-ruri-large[ii]を使用する．構築したそれぞれのテーブルのレコード数の内訳を表 2 に示す．

3.2 プロンプト

プロンプトでは，以下の 4 パターンを作成し，それぞれ比較実験する．

1. 外部情報なし，例示・条件なし
2. 外部情報なし，例示・条件あり
3. 外部情報あり，例示・条件なし
4. 外部情報あり，例示・条件あり

例示には，それぞれの傾聴応答について日本語日常対話コーパスから 2 例ずつ人手で選定したデータを用いる．各条件は，掘り下げ質問は「？」，尊重応答は「いいですね」と「素晴らしいですね」，共感応答は「そうですね」と「よくわかります」，「私もです」をそれぞれ提示してプロンプトを作成する．

[i] https://qdrant.tech/

[ii] https://ollama.com/kun432/cl-nagoya-ruri-large

3.3 応答生成モデル

応答生成モデルを API 利用するために Ollama[iii]を使用する．また，モデルは，ELYZA 社が公開している Llama-3-ELYZA-JP-8B-GGUF[iv]を使用する．

3.4 検索クエリを工夫した傾聴応答生成

検索クエリを工夫する手法として，HyDE を用いる．本稿では，傾聴応答がどの程度生成できるか検証することを目的としているため，仮説的な回答生成･利用による応答生成への変化を確認する．また，HyDE では，仮説的な応答を生成することから，プロンプトとは異なり例示は提示せず，仮説的な応答を生成する際に，以下に示す条件の有無の 2 パターンで検証する．

1. 仮説的な応答生成の際に条件の提示なし
2. 仮説的な応答生成の際に条件の提示あり

4 評価実験

4.1 データセットの構築

評価実験では，先行研究[6]で構築した発話テンプレートから人手で 100 発話を選定し，それらの発話について図 1 と図 2 のそれぞれのフローにしたがって 1 発話につき 1 応答を生成する．評価用のための発話例を表 3 に示す．

4.2 実験設定

4.1 節のデータに対して，3.1.1 項のデータ構築時と同様，各傾聴応答の生成件数について機械的に判定する．ただし，尊重応答&掘り下げ質問や，共感応答&掘り下げ質問のような混合した応答が生成された場合は，掘り下げ質問を優先してカウントする．

表 3：評価用の発話例

発話
ニュースやドラマをよく見ています。
趣味は、絵を描く事です。
秋は気候が良いので遠出したくなりますね。

[iii] https://ollama.com/
[iv] https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B-GGUF

表 4：ベースラインの評価結果（件数）

	掘り下げ質問	尊重応答	共感応答
外部情報なし，例示･条件なし	75	4	2
外部情報なし，例示･条件あり	**85**	81	23
外部情報あり，例示･条件なし	79	6	3
外部情報あり，例示･条件あり	75	**87**	**57**

表 5：検索クエリを工夫した評価結果（件数）

	掘り下げ質問	尊重応答	共感応答
条件の提示なし	77	7	3
条件の提示あり	76	3	1

4.3 評価結果

ベースライン（検索クエリを工夫しない RAG を用いた傾聴応答生成）の評価結果を表 4 に，検索クエリを工夫した傾聴応答生成を表 5 に示す．表 4 より，掘り下げ質問については，どの実験においても約 7~8 割程度生成が可能であることがわかる．一方で，尊重応答については，例示・条件ありのものについては約 8 割程度生成が可能であるが，例示・条件なしのものは 1 割未満の生成となった．共感応答については，例示・条件ありのものに関しても最高で約 5 割の結果となり，「共感」という言葉についての指示方法の工夫や追加学習などが必要である．また，表 5 より，尊重応答や共感応答については，LLM が生成した応答による検索結果において，1 割以下の生成結果となったことから，追加学習が必要である．具体的な生成結果を確認すると，例示・条件なしや検索クエリを工夫したものに関しては，例示・条件ありのものと比較して生成文の長さが長い傾向があった．これについては，生成内容に関して人手評価を実施する必要がある．

4.4 考察

尊重応答や共感応答における例示・条件なしについては，生成数が少ないことから，外部情報の構築法の工夫やモデルベースでの追加学習などの必要性

表 6：条件の追加後の例

傾聴の種類	条件
掘り下げ質問	掘り下げ質問：「？」 含めない：「なぜ」，「どうして」
尊重応答	尊重：「いいですね」，「素晴らしいですね」，「素敵ですね」，「立派ですね」，「楽しそうですね」，「大変ですね」，「つらいですね」
共感応答	共感:「そうですね」，「よくわかります」，「私もです」，「納得です」，「私もそう思います」，「その気持ちはよく分かります」

を確認した．特に，共感応答については，例示・条件ありについても生成数が少ないことから元々の言語モデルの応答性能についても調査が必要である．

そこで，追加実験として，外部情報あり，例示・条件ありに関して，条件の内容を追加したモデルに対する応答生成の変化を確認する．

5 追加実験

5.1 条件の追加

3.2 節で述べたプロンプトの条件に対して，表 6 に示す条件を追加する．特に，バリデーション技法のオープンクエスチョンでは，「なぜ～ですか？」や「どうして～ですか？」のような質問は，多くの被介護者が応答に困る傾向があるため，避けたほうがよい表現とされている．したがって，掘り下げ質問に対して，これらの表現を除外するために，応答に「なぜ」や「どうして」の生成を含めないように条件を追加する．尊重応答や共感応答については元々の言語モデルおよび外部情報を用いた応答生成の可能性について調査するために，条件の例を追加することにより応答の変化を確認する．

表 7：条件の追加後の評価結果（件数）

	掘り下げ質問	尊重応答	共感応答
外部情報あり，例示・条件あり 条件の追加あり	81	49	**63**

5.2 評価結果

評価結果を表 7 に示す. 掘り下げ質問については，条件を追加しても生成数の減少は見られなかった．一方で，尊重応答については生成数が減少した．これは，条件の中に「いいですね」などのようなポジティブな応答と「大変ですね」などのようなネガティブな応答が混じっており，条件を追加することで応答生成が難しくなった可能性が考えられる．共感応答については，条件を追加することで，「共感」に関する情報も増えるため，応答の生成数が増加したといえる．

5.3 考察

掘り下げ質問については，全体的に約 7~8 割応答が生成されており，言語モデルのみでも生成が可能であると考えられる．尊重応答や共感応答については，例示・条件の提示により，生成数は増加するが，外部情報が上手く使われていない可能性が高く，外部情報の工夫が必要である．また，条件を追加することにより尊重応答は生成数が減少した．これらの結果より，複雑な条件に対してはプロンプトだけでなく，言語モデルのファインチューニングも考慮する必要があるといえる．

6 おわりに

本稿では，先行研究から基盤モデルを変更し，RAG におけるプロンプトの工夫，RAG における検索クエリを改善した手法を用いて，傾聴応答の生成可能性を検証した．評価の結果，掘り下げ質問に関しては，基盤モデルのみでの応答生成が可能である結果となった．一方で，尊重応答や共感応答に関しては，外部情報が上手く使われていない可能性が高く，さらなる工夫が必要である．また，プロンプトの条件追加により，共感応答は生成数が増加したが，条件を追加したことにより指示の難易度が高くなったことから応答の生成時間が増加する傾向が見られた．この点については，言語モデルのファインチューニングも検討する必要がある．

今後の課題として，外部情報の工夫の仕方や，言語モデルのファインチューニング，人手評価について検討する．

参考文献

1. 統計からみた我が国の高齢者ー「敬老の日」にちなんでー.（引用日：2025 年 1 月 6 日.）https://www.stat.go.jp/data/topics/topi1420.html.
2. 令和 6 年版高齢社会白書.（引用日：2025 年 1 月 6 日.）https://www8.cao.go.jp/kourei/whitepaper/w-2024/html/zenbun/s1_2_2.html.
3. 介護のコミュニケーションの重要性｜話題作り・大切な技法とコツとは.（引用日：2025 年 1 月 6 日.）https://rehab.cloud/mag/12028/.
4. 都村尚子，三田村知子，橋野建史，認知症高齢者ケアにおけるバリデーション技法に関する実践的研究，関西福祉大学紀要第 14 号，2010.
5. 認知症ケアのコミュニケーション方法「バリデーション」とは｜認知症のコラム．（引用日：2025 年 1 月 6 日.）https://www.sagasix.jp/column/dementia/validation/.
6. 松本奈々，安藤一秋，報酬表現に着目した強化学習による傾聴対話モデルの改善, 第 23 回情報科学技術フォーラム講演論文集，pp.389-390，2024.
7. Dubey, Abhimanyu, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783, 2024.
8. 「GPT-4」を上回る日本語性能の LLM「Llama-3-ELYZA-JP」を開発しました.（引用日：2025 年 1 月 6 日.）https://note.com/elyza/n/n360b6084fdbd.
9. Patrick Lewis, et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems, Vol. 33, pp.9459-9474, 2020.
10. Yunfan Gao et al., Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv preprint arXiv: 2312.10997, 2023.
11. 西見公宏，吉田真吾，大嶋勇樹，LangChain と LangGraph による RAG・AI:-ジェント [実践]入門, 株式会社技術評論社, 2024.
12. Luyu Gao, Xueguang Ma, Jimmy Lin, Jamie Callan, Precise Zero-Shot Dense Retrieval without Relevance Labels, arXiv preprint arXiv:2212.10496, 2022.
13. Zackary Rackauckas, RAG-Fusion: a New Take on Retrieval-Augmented Generation, arXiv preprint arXiv: 2402.03367, 2024.
14. 赤間怜奈，磯部順子，鈴木潤，乾健太郎，日本語日常対話コーパスの構築, 言語処理学会第 29 回年次大会発表論文集，pp.108-113，2023.