F-028

RAG を用いたロボット設計サポートシステムの精度評価
Accuracy Evaluation of Robot Design Support System Using RAG

三浦 颯斗 1)　山内 翔 1)　鈴木 恵二 1)
Hayato Miura　Sho Yamauchi　Keiji Suzuki

1. はじめに

近年，ロボットの社会的利用が増加し，需要に応じたロボットが設計されている．しかし，その機能や形状が多岐にわたることから，ロボットに合わせた設計が必要になっている．また，それらは人手によって行われているため，時間とコストが膨大なものになっている．これを解決するために，大規模言語モデル（以下 LLM）を用いたロボット設計を効率化する研究が行われている[1]．この研究では，LLM を用いた設計の実例として，ChatGPT-3[2]を活用し食糧不足をテーマにしたロボットの設計プロセスを挙げている．これにより，LLM との対話形式によるロボット設計が可能であることが示されている．一方で，LLM は与えられたプロンプトに対して最も確からしい回答を生成する．そのため，これらの回答には不確定性や情報の信頼性などの問題があるとしている．このことから，LLM を用いたロボット設計は実用できるレベルには至っていない．また，LLM に最新情報を随時反映させるにはコストがかかる点や，ロボット設計に特化した LLM は存在していない点も課題となっている．

そこで本研究では RAG[3]を用いたロボット設計サポートシステムの開発を行う．RAG では，情報の索引化と検索により，LLM が出力に使用する情報の補強を行う．これにより，ハルシネーションの軽減や情報の専門性を高め，LLM の課題を解決したロボット設計のサポートが可能となる．

2. 提案システム

提案システムでは，ユーザの質問に対して回答することで，ロボット設計をサポートする．提案システムを開発するに向けて，RAG を用いたロボット設計サポートシステムの構築を行う．図 1 に RAG の流れを示す．RAG では，ユーザの質問に対して LLM がデータベースに関連情報の検索を行う．その後，検索結果をもとに文章をユーザに出力する．

3. 精度評価実験

提案システムでは，RAG を用いてユーザのロボット設計のサポートを行う．この時，ユーザに正しい情報を提示する必要がある．そのため，ユーザの質問に対する回答精度の評価を行った．

3.1 実験詳細

評価実験は，2 つの検証項目からなる．1 つ目は，データベース内のデータ数に応じた回答精度の検証である．2 つ目は LLM の言語モデルの違いによる回答精度の検証である．この 2 つの項目を組み合わせ，各言語モデルに対して，データベースにデータを段階的に保存した場合の各データ数における回答精度の検証を行う．使用する言語モデルは Mistral[4]と Llama3.2 1B[5]の 2 種類である．使用データは IEEE によりまとめられたロボットの仕様データ 10 種類である．また，質問文は「Please tell me about ACM-R5H(ロボット名).」である．

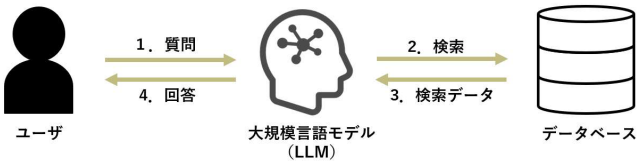

図 1　RAG

実験手順は次の通りである．はじめに，データベースに ACM-R5H(質問するロボット)に関するデータファイルをアップロードする．その後，LLM に質問を行い，回答を記録する．回答記録後，データベースに他のロボットに関するデータファイルを 1 個アップロードする．その後は同様の質問を LLM に行い，回答を記録する．この手順をデータベース内のデータファイル数が 10 個になるまで繰り返す．

3.2 実験結果

各言語モデルに対して，データベースにデータを段階的に追加した場合の各データ数における実際の回答を図 2 に示す．

3.3 考察

入力した質問に対して，データファイル数が 1 個の場合，どちらの言語モデルも正しい情報を出力した．しかし，データファイル数が 2 個の場合は間違った情報が混ざっていた．さらに，データファイル数が 3 個以上になると間違った情報を出力もしくは情報を見つけることができていなかった．この結果から，現時点における提案システムは実用的ではないと言える．

RAG の回答精度の低下の原因として次の 6 点が挙げられている[6].

1. データ不足
- 利用可能なデータ不足していることから，回答に情報の欠落が発生する

2. データの見逃し
- RAG では検索に対して関連性の高い上位数件のデータが選ばれるが，検索エンジンの制約などにより重要なデータが上位に含まれない場合に発生する

3. コンテキストへの統合の失敗
- 検索で返されるデータが多い場合，抽出可能なデータ数の制限などから，重要なデータがコンテキストに取り込まれない場合に発生する

4. コンテキストからのデータ抽出の失敗

1)　Future University Hakodate

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

mistral

データファイル数：1
The ACM-R5H is a snake robot designed for underwater inspections and search-and-rescue missions in hazardous environments. …

データファイル数：2
The ACM-R5H is a snake robot designed for underwater inspections and search-and-rescue missions in hazardous environments. It was developed by … and also features a gripper designed by Robotiq for picking up small objects. …

データファイル数：3～10（3 個時の出力を抜粋）
The ACM-R5H is not a widely recognized or well-known technology related to the context provided in your query. The text appears to …

llama3.2 1b

データファイル数：1
The ACM-R5H is a snake robot designed to perform underwater inspections and search-and-rescue missions in hazardous environments. It was created by…

データファイル数：2
The ACM-R5H is a snake robot designed for underwater inspections and search-and-rescue missions in hazardous environments.… The Adaptive Gripper, on the other hand, is a simple and robust robotic hand designed for industrial applications, created by Robotiq in Canada.

データファイル数：3～10（3 個時の出力を抜粋）
Based on the provided documents, there is no information about a device named “ACM-R5H” in the given context. The documents provide …

図 2　出力結果（一部抜粋）
（赤文字:正しい情報，青文字:間違った情報）

- コンテキストに正しい情報が含まれているにもかかわらず，ノイズや矛盾する情報が含まれていると発生する

5. フォーマットに関するプロンプトの誤った解釈
- LLM がフォーマットに関するプロンプトを適切に解釈せず，特定のフォーマットでの出力を無視する

6. 曖昧な質問
- 入力された質問が曖昧な場合，回答が具体的ではないまたは過剰に具体的になってしまう

7. 一度に複数の質問
- 複数の質問を一度に入力すると，コンテキストのデータを見落としてしまい回答が不完全になる

本実験ではデータファイル数を増やすことで回答精度が悪くなっていったことから，特に 2，3，4 が原因であると考えられる．

これらの原因解消の取り組みとして次の 3 点が挙げられている．

1. データベースへのデータ追加
- データ不足の解消

2. データのチャンク分割
- データを単語やフレーズなどに分割し，扱いやすくする

3. 質問のリライト
- 質問にテキストを追加することで，LLM がより正確に質問を理解できるようにする

また，RAG の回答精度を向上するアプローチとして，質問や要求などユーザからの入力であるクエリの処理について研究が行われている．これらの研究では，クエリに対して複数のプロンプトを使用する手法[7]や，スタック構造を利用したクエリと関連情報の管理手法[8]などが提案されている．

以上のような取り組みを行うことで，RAG の回答精度の向上を図る．

4.　まとめ

本研究では，RAG を用いたロボットの設計サポートシステムを提案し，その回答精度を評価した．精度評価実験では，データファイル数が少ない場合は正しい情報を出力したが，データファイル数が多くなるにつれ，その精度は低下していく結果となった．これを踏まえて，回答精度低下の原因とその解消への取り組みを挙げた．今後はこれらに取り組み，RAG の回答精度の向上を図っていく．

参考文献

[1] Stella F., Della Santina C., Hughes J., How can LLMs transform the robotic design process?, *Nat Mach Intell* 5, 561‑564 , 2023, https://doi.org/10.1038/s42256-023-00669-7
[2] GPT-3，https://openai.com/index/gpt-3-apps/（最終アクセス日：2025 年 5 月 4 日）
[3] Patrik Lewis，Ethan Perez，Aleksandra Piktus，Fabio Petroni，Vladimir Karpukhin，Naman Goyal，Heinrich Küttler，Mike Lewis，Wen-tau Yih，Tim Rocktäschel，Sebastian Riedel，Douwe Kiela，Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks，Advances in Neural Information Processing Systems, 33，9459–9474，2020，https://arxiv.org/abs/2005.11401
[4] Mistral，https://mistral.ai/（最終アクセス日：2025 年 5 月 28 日）
[5] Llama3.2 1B，https://www.llama.com/models/llama-3/（最終アクセス日：2025 年 5 月 28 日）
[6] Scott Barnett，Stefanus Kurniawan，Srikanth Thundumu，Zach Brannelly，Mohamed Abdelrazek，Seven Failure Points When Engineering a Retrieval Augmented Generation System，2024，https://doi.org/10.48550/arXiv.2401.05856
[7] Chengrui Wang，Qingqing Long，Meng Xiao，Xunxin Cai，Chengjun Wu，Zhen Meng，Xuezhi Wang，Yuanchun Zhou，BioRAG: A RAG-LLM Framework for Biological Question Reasoning，2024，https://arxiv.org/abs/2408.01107
[8] Xinke Jiang，Yue Fang, Rihong Qiu，Haoyu Zhang，Yongxin Xu，Hao Chen, Wentao Zhang，Ruizhe Zhang，Yuchen Fang，Xu Chu，Junfeng Zhao，Yasha Wang，TC-RAG: Turing-Complete RAG's Case study on Medical LLM Systems，2024，https://arxiv.org/abs/2408.09199

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.