RAG の生成器における SLM の利用

阿部 晃弥 [1]　新納 浩幸 [2]
[1] 茨城大学大学院 理工学研究科 情報工学専攻
[2] 茨城大学大学院 理工学研究科 情報科学領域
24nm701t@vc.ibaraki.ac.jp　hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

LLM に外部知識を統合する手法である RAG は，情報を検索する「検索器」と，その情報を基に回答を生成する「生成器」から構成されることが一般的である．本研究の目的は，検索器の能力が十分高い場合，SLM を生成器として用いても RAG が有効に機能するかを検証することである．実験では，JQaRA データセットを用いた QA タスクを実施し，検索器が常に正解文書を検索できるという理想的な条件を仮定した．その結果，検索器の性能が高い場合でも，生成器として性能の高いモデルを使用する方が全体の回答精度が向上することが示された．また，生成器の機構に関して，複数の SLM をアンサンブルとして組み合わせることで，全体の正解率を大幅に向上させる可能性が示された．

1　はじめに

近年，自然言語処理の分野では，大規模言語モデル (Larage Language Model，以下 LLM と略す) が高い性能を示している．しかし，特定の分野における詳細な知識や比較的新しい知識といった，学習データに含まれていない情報を扱うことが難しいという欠点がある．この欠点に起因する現象として，言語モデルが不適切な回答を生成する Hallucinations(幻覚) [1] が挙げられる．この問題を解決するため，外部知識を様々な形で LLM に組み込む手法 [2, 3, 4] が研究されている．その一つが **RAG (Retrieval-Augmented Generation)**[5] である．

RAG は，外部知識をベクトル・インデックスとして保存し，入力文をクエリとしてインデックスに問い合わせることで関連する文書を取得し，それを入力文と共に LLM に与えることで，外部知識に基づいた出力を生成する手法である．RAG は一般的に Retriever と Generator の二つの機構に分けられる．Retriever は，Query Encoder と Document Index から構成され，クエリに関連する文書を検索する役割を担う．この機構では，クエリに関連する文書を，高精度で検索することが求められる．Retriever は，RAG の中核的要素といえるため，多くの研究 [6] が行われている．Generator は，Retriever から受け取った Document を用いて文章を生成する役割を担う．つまり，最終的な回答文はこの機構によって生成される．

本研究では，Generator として小規模言語モデル (Small Language Model，以下 SLM と略す) を組み込み，QA タスクを実施することで，使用するモデルが性能や生成結果に与える影響について調査する．また，本実験では，Retriever は完璧であり，必ず入力文に対応する適切な文書を一つ検索できるという理想的な条件を仮定する．RAG において，検索器 (Retriever) の性能が十分高い場合の，SLM の利用可能性について明らかにすることを本研究の目的とする．

2　実験用データセット

本実験には，hotchpotch の Yuichi Tateno 氏が公開している JQaRA(Japanese Question Answering with Retrieval Augmentation) データセット[1)][7] を使用した．

2.1　JQaRA データセット

JQaRA データセットは，一つの質問に対して，検索対象文となる 100 件の文書が対応付けられているデータセットである．データセットの全容を簡略化した様子を図 1 に示す．また，JQaRA データセットの内容の例を付録 A に記す．このデータセットの大元の質問文及び正答として，「JAQKET: クイズを題材にした日本語 QA データセット」[2)][8] の質問と回答が用いられている．JAQKET は，質の高い多様な

1）https://huggingface.co/datasets/hotchpotch/JQaRA
2）https://www.nlp.ecei.tohoku.ac.jp/projects/jaqket/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

図 1　JQaRA データセットの全容

日本語 Q&A データセットであり，Wikipedia の記事タイトル名が回答となる特徴を持っている．

検索対象文の取得方法　検索対象文には，Wikipedia のデータを最大文字数が 400 文字になるようにチャンク分割した，singletongue/wikipedia-utils-passages-c400-jawiki-20230403[3] を使用している．関連する文章の取得には、Embeddings モデルを用いた文ベクトルの類似度が用いられる．このとき，Embeddings モデルによる偏りを防ぎ，多様性を確保するために 5 種類の Embeddings モデルが利用される．そして，各質問文に対し，最も類似する上位 500 件の文書を，各 Embeddings モデルごとに取得する．これら 5 つの結果を RRF(Reciprocal Rank Fusion) を用いてランク付けしなおし，スコアが高い上位 100 文が抽出される．抽出された文書には、それらが含まれる Wikipedia 記事のタイトルも対応付けられており，各質問文にそれぞれタイトル及び検索対象の関連文書が紐づけられる．

正解ラベルの付与方法　質問文に紐付けられた 100 件の文書のうち，文書の本文またはタイトルに，質問に対応する正答の文字列が完全一致で含まれる場合，質問文と検索対象文の間に関連があると判断し，正解ラベルが付与される．

3　実験手法

本実験では，JQaRA データセットの 1,667 件の質問 × 100 件の検索文書からなる 166,700 件のテストデータのうち，正解ラベルが付与された 16,204 件のデータを対象として実験をおこなった．

3.1　実験用 LLM

本実験では，Generator に組み込む SLM として，以下の 8 種類のモデルを用いる．

- line-corporation/japanese-large-lm-1.7b-instruction-sft[4]
- line-corporation/japanese-large-lm-3.6b-instruction-sft[5]
- rinna/japanese-gpt-neox-3.6b-instruction-ppo[6]
- tokyotech-llm/Swallow-7b-instruct-hf[7]
- google/gemma-2-2b-jpn-it[8]
- microsoft/Phi-3-mini-128k-instruct[9]
- llm-jp/llm-jp-3-3.7b-instruct[10]
- elyza/Llama-3-ELYZA-JP-8B[11]

また，実験における正答率の比較のため，OpenAI 社が提供する gpt-3.5-turbo-instruct を追加で使用する．

3.2　手順

2 章で説明した JQaRA データセットのうち，正解ラベルが 1 であるデータ 16,204 件を以下の表 1 に示したプロンプトに組み込み，LLM に入力として与える．

表 1　実験用プロンプト

以下のコンテキストを使用して質問に回答してください。 ————- コンテキスト: title: context ————- 質問: question 回答:

具体的には，JQaRA データセットの text, title, question を，それぞれプロンプトの context, title, question に組み込む．このようにプロンプトを構築することで，あるクエリと，そのクエリに関連するコンテキストを 1 セットとして LLM に入力する．つまり，本実験において検索器は完璧であり，必ず一件の関連文書を検索できると仮定される．プログラムとして実装する際には，フレームワークとして LangChain を用いる．そして，すべての言語モデル

3）https://huggingface.co/datasets/singletongue/wikipedia-utils

4）https://huggingface.co/line-corporation/japanese-large-lm-1.7b-instruction-sft

5）https://huggingface.co/line-corporation/japanese-large-lm-3.6b-instruction-sft

6）https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-ppo

7）https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-hf

8）https://huggingface.co/google/gemma-2-2b-jpn-it

9）https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

10）https://huggingface.co/llm-jp/llm-jp-3-3.7b-instruct

11）https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

において，temperture を 0 に固定する．回答の評価は，生成器のモデルが出力した内容において，テストデータの質問に対する正答が部分一致で含まれるかによって判別する．以上の実験に関して，3.1 節で挙げた 8 種類の SLM と比較用の LLM の，合計 9 種類の言語モデルを Generator に組み換えて繰返し実施し，各モデルのテストデータに対する正答率を調査する．

4 実験結果

3 章の手法を用いて実験をおこなった．その結果は以下の表 2 の通りである．

表 2 各 LLM の正解率

LLM	w/o RAG	w/ RAG	diff
line-1.7b-inst.	0.205	0.578	+0.373
line-3.6b-inst.	0.376	0.683	+0.307
rinna-3.6b-inst.	0.408	0.697	+0.289
Swallow-7b-inst.	0.580	0.854	+0.274
gemma-2-2b-jpn-it	0.062	0.625	+0.563
Phi-3-mini-128k-inst.	0.018	0.689	+0.671
llm-jp-3-3.7b-inst.	0.558	0.809	+0.251
Llama-3-ELYZA-JP-8b	0.425	0.850	+0.425
gpt-3.5-turbo-inst.	0.484	0.844	+0.360

表 2 より，RAG の機構によって正解コンテキストの情報をプロンプトに付与することで，QA タスクの正答率が向上することが示された．

その正答率は，モデルのパラメータ数に大きく左右され，比例して正答率が上昇する様子がみられる．一方で，正答率の向上の幅はパラメータ数に対する影響はあまり見られない．RAG の機構がない場合の正答率が低いモデルほど，正答率の向上の幅が大きいのではないかと考えられる．つまり，言語モデルが持つ知識の量と，プロンプトの指示に対する適応度の高さに関しては，トレードオフの関係にあるのではないかと推察される．

そして，SLM に対する比較対象として，gpt-3.5-turbo-instruct を用意したが，7B クラスのモデルであれば，それを上回る性能を示した．

5 アンサンブルに関する考察

5.1 各 SLM の比較

各テストデータに関して，それぞれどのモデルが正答可能であったかを調査した．そして，gpt-3.5-turbo-instruct を除いた実験対象の SLM から 2 つのモデルを選択し，両方のモデルが正解したデータ，片方のモデルが正解したデータ，両方のモデルが不正解だったデータをそれぞれカウントした．このカウントを，8 つのモデルから 2 つを選択する，全 28 通りの組み合わせについて実施した．その結果に関して，line-1.7b-inst.，gemma-2-2b-jpn-it，line-3.6b-inst. の結果を以下の表 3，表 4，表 5 に示す．なお，その他の比較結果に関してはここでは省略し，付録 B に記す．

表 3 各テストデータについて正解可能か，line-1.7b-inst.（比較元）と各 LLM（比較先）の比較

	line-3.6b	rinna	Swallow	gemma
両方正解	0.461	0.452	0.523	0.405
比較先のみ正解	0.118	0.126	0.055	0.173
比較元のみ正解	0.222	0.245	0.330	0.219
両方不正解	0.200	0.177	0.091	0.203
	Phi-3	llm-jp-3	ELYZA	
両方正解	0.425	0.507	0.518	
比較先のみ正解	0.153	0.071	0.061	
比較元のみ正解	0.265	0.303	0.332	
両方不正解	0.157	0.119	0.089	

表 4 各テストデータについて正解可能か，gemma-2-2b-jpn-it（比較元）と各 LLM（比較先）の比較

	line-3.6b	rinna	Swallow
両方正解	0.462	0.494	0.576
比較先のみ正解	0.162	0.130	0.049
比較元のみ正解	0.221	0.203	0.278
両方不正解	0.155	0.172	0.097
	Phi-3	llm-jp-3	ELYZA
両方正解	0.476	0.565	0.589
比較先のみ正解	0.148	0.059	0.036
比較元のみ正解	0.213	0.244	0.261
両方不正解	0.162	0.131	0.114

表 5 各テストデータについて正解可能か，line-3.6b-inst.（比較元）と各 LLM（比較先）の比較

	rinna	Swallow	Phi-3
両方正解	0.516	0.609	0.490
比較先のみ正解	0.167	0.074	0.192
比較元のみ正解	0.181	0.245	0.199
両方不正解	0.136	0.073	0.118
	llm-jp-3	ELYZA	
両方正解	0.590	0.604	
比較先のみ正解	0.093	0.079	
比較元のみ正解	0.220	0.246	
両方不正解	0.098	0.071	

この結果から，line-1.7b-inst. で正答可能であったにもかかわらず，よりパラメータ数の多いモデルで正答不可能であったデータが，全テストデータ 16,204 件中，それぞれ 5%以上存在することが示された．さらに，その他の全組み合わせに関しても，

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

片方のモデルのみが正答可能であるデータが一定数以上存在することが，分析結果から明らかになっている．つまり，各モデルにはパラメータ数の大小だけでは判別できない，それぞれの得意な領域が存在するのではないかと考えられる．

5.2 アンサンブル RAG の疑似的な構築

各モデルの得意な領域を互いにカバー可能である，アンサンブル RAG について疑似的に構築する．ここでは，3 つの SLM を用いたアンサンブル RAG と，4 つの SLM を用いたアンサンブル RAG について考える．そして，正答の判別方式として，組み込んだ SLM において，いくつのモデルが正答可能であるかに応じて，2 種類の方式を採用する．つまり，検討するアンサンブル RAG は以下の 4 種類である．

1. 3 つの SLM を用いて，あるテストデータに対して，一つ以上のモデルが正答可能であれば，そのデータに正答可能であったとするケース
2. 3 つの SLM を用いて，あるテストデータに対して，二つ以上のモデルが正答可能であれば，そのデータに正答可能であったとするケース
3. 4 つの SLM を用いて，あるテストデータに対して，一つ以上のモデルが正答可能であれば，そのデータに正答可能であったとするケース
4. 4 つの SLM を用いて，あるテストデータに対して，二つ以上のモデルが正答可能であれば，そのデータに正答可能であったとするケース

それぞれのケースに対して，実験で用いた gpt-3.5-turbo-instruct を除く 8 種類の SLM から全組み合わせを試し，その正答率を調査した．なお，ケース 1，ケース 2 に関しては全 56 通り，ケース 3，ケース 4 に関しては全 70 通りの組み合わせである．その結果を表した箱ひげ図を，図 2 に示す．

また，それぞれのケースの最高正答率及び最低正答率を以下の表 6 に示す．

表 6 各ケースの最高正解率（上段）・最低正解率（下段）とその組み合わせ

	acc.	comb.
case1	0.971	Swallow-7b, Phi-3-mini, Elyza-8b
	0.889	line-1.7b, Gemma-2-2b, rinna-3.6b
case2	0.873	Swallow-7b, llm-jp-3, Elyza-8b
	0.658	line-1.7b, Gemma-2-2b, rinna-3.6b
case3	0.984	line-3.6, Swallow-7b, Phi-3-mini, Elyza-8b
	0.936	line-1.7b, line-3.6b, Gemma-2-2b, rinna-3.6b
case4	0.920	Swallow-7b, Phi-3-mini, llm-jp-3, Elyza-8b
	0.792	line-1.7b, line-3.6b, Gemma-2-2b, rinna-3.6b

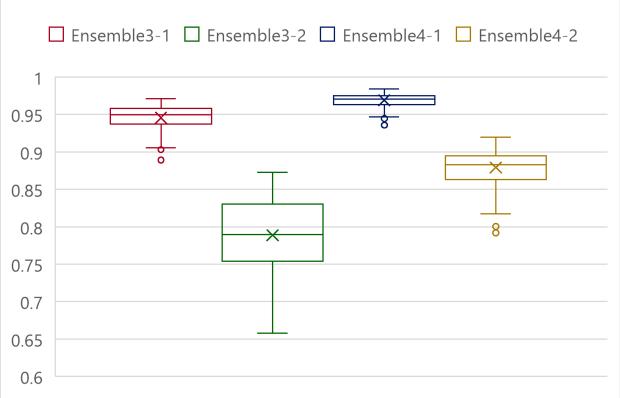

図 2 疑似的なアンサンブル RAG の正答率，左からケース 1，ケース 2，ケース 3，ケース 4

この結果から，SLM を使用した場合であっても，複数のモデルを組み合わせるアンサンブルによる構成にすることで，最大 97%から 98%程度の正答率を達成することが可能であると示された．しかし，今回シミュレートしたアンサンブル RAG において，複数の SLM の出力からどのように正確な回答を取得するのかについて考慮していない．2 つの SLM が同様の回答を生成した場合，それらを最終的な回答とする手法が考えられるが，ケース 2，ケース 4 のように，2 つ以上のモデルが正答可能という条件を付加した場合，アンサンブルでの正答率が低下することが示された．よって，いかに正答を引き出すことができるかが今後の研究の焦点であると考える．

6 結論

RAG の検索器の能力が十分高い場合，小型の LLM でも十分な性能を発揮するか確認することを目的に実験をおこなった．実験では，正解が導けるであろうコンテキストを 1 つプロンプトに与えて回答を出力させる形をとった．この実験によって，検索器の能力が十分高くても，性能の高い LLM を使う方が QA タスクの正解率が高いという結論が得られた．また，それぞれの SLM が各テストデータに関して正答可能であったかどうかを調査することで，各モデルには得意な領域が存在すると推察された．そこで，複数の SLM によるアンサンブル RAG に関して調査を実施したところ，最大 98%程度のテストデータにて，いずれか一つの SLM が正答を生成可能であるという条件を満たすことが可能であることが示された．今後は，SLM を用いた効果的なアンサンブル RAG の手法に関して調査・研究を実施したい．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は国立国語研究所の共同研究プロジェクト「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」及び JSPS 科研費 23K11212 の助成を受けています.

参考文献

[1] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, Vol. 55, No. 12, p. 1–38, March 2023.

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.

[3] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022.

[4] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding, 2021.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[6] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023.

[7] Yuichi Tateno. Jqara: Japanese question answering with retrieval augmentation - 検索拡張 (rag) 評価のための日本語 q&a データセット.

[8] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. Jaqket: クイズを題材にした日本語 (qa) データセットの構築, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　実験用テストデータの例

3.1 章で述べた RAG の実験に用いた評価用のテストデータのうち 3 件を以下の表 7 に示す．

表 7　JQaRA データセットの例

label	text	title	question	answer
1	絶対零度 (ぜったいれいど、アブソリュートゼロ、英: Absolute zero) は、熱力学上の最低温度である摂氏 − 273.15 度。	絶対零度 (曖昧さ回避)	摂氏ではマイナス 273.15 度にあたる、全ての原子の振動が停止する最も低い温度を何というでしょう?	[”絶対零度”]
1	温度は、物質の熱振動をもとにして規定されているので、下限が存在する。それは、熱振動 (原子の振動) が小さくなり、エネルギーが最低になった状態である。この時に決まる下限温度が絶対零度である。古典力学では (以下略)	絶対零度	摂氏ではマイナス 273.15 度にあたる、全ての原子の振動が停止する最も低い温度を何というでしょう?	[”絶対零度”]
0	物理学においては、絶対温度において切りのよい数字である 300K (27° C) が室温とされる場合が多い。	室温	摂氏ではマイナス 273.15 度にあたる、全ての原子の振動が停止する最も低い温度を何というでしょう?	[”絶対零度”]

B　各 SLM の比較結果

5.1 節で述べた全モデルの組み合わせの比較結果に関して，未記載分を以下の表 8 から表 11 に示す．

表 8　各テストデータについて正解可能か，rinna-3.6b-inst.（比較元）と各 LLM（比較先）の比較

	Swallow	Phi-3	llm-jp-3	ELYZA
両方正解	0.635	0.510	0.617	0.636
比較先のみ正解	0.062	0.187	0.080	0.062
比較元のみ正解	0.218	0.179	0.192	0.214
両方不正解	0.084	0.123	0.110	0.088

表 9　各テストデータについて正解可能か，Phi-3-mini-128k-inst.（比較元）と各 LLM（比較先）の比較

	Swallow	llm-jp-3	ELYZA
両方正解	0.606	0.582	0.609
比較先のみ正解	0.084	0.107	0.080
比較元のみ正解	0.248	0.227	0.241
両方不正解	0.063	0.083	0.070

表 10　各テストデータについて正解可能か，llm-jp-3-3.7b-inst.（比較元）と各 LLM（比較先）の比較

	Swallow	ELYZA
両方正解	0.734	0.731
比較先のみ正解	0.075	0.078
比較元のみ正解	0.119	0.119
両方不正解	0.071	0.072

表 11　各テストデータについて正解可能か，Swallow-7b-inst.（比較元）と各 LLM（比較先）の比較

	ELYZA
両方正解	0.763
比較先のみ正解	0.091
比較元のみ正解	0.087
両方不正解	0.059

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).