国語科共通テスト試行調査を用いた RAG による答案生成の評価と再検索 RAG の提案

一柳壮綱 [1]　古宮嘉那子 [2]　石岡恒憲 [3]　中川正樹 [2]
[1] 東京農工大学工学部　[2] 東京農工大学大学院　[3] 独立行政法人大学入試センター
{s236222v@st.go, kkomiya@go, nakagawa@cc}.tuat.ac.jp　tunenori@rd.dnc.ac.jp

概要

近年，大規模言語モデル（LLM: Large Language Model）は発展の一途を辿ってきたが，これまで日本語の国語問題，特に大学入学共通テストのような全国試験レベルの問題に LLM を用いて取り組む研究は十分に行われてこなかった．本研究では，この課題に取り組むため，複数の LLM を用いて大学入学共通テスト試行調査の国語科目における答案を作成し，採点ガイドラインに基づいてその答案を評価する．また，Retrieval-Augmented Generation（RAG）手法を活用し，答案生成の手法を検討する．この際，再検索プロセスを考慮した手法を取り入れることで既存の RAG 手法を改良し，生成される答案の精度や一貫性の向上を目指すとともに，単純な LLM での生成との違いを調査した．実験の結果，再検索プロセスを考慮した RAG 手法を用いることによって，より採点条件を満たした答案を生成できるようになることを確認した．

1　はじめに

「国語」の試験は，受験生の基礎知識を測るだけでなく，多様な素材や議論を通じて思考の過程を深め，その成果を問うことを目的としている．特に，記述式や複数選択の設問を組み合わせることで，文脈の理解力や主体的な学びを評価する．

大規模言語モデル（LLM: Large Language Model）は，言語の理解 [1] および生成 [2] において高い性能を発揮している．これを活用することで，多層的な素材や設問に対応し，文脈を踏まえた答案を生成することが可能になる．さらに LLM を用いることで，受験生の思考過程を可視化し，主体的な学びを支援する新しいアプローチにつながると期待される．

本研究では，大学入学共通テスト試行調査における各設問に対し，その答案生成手法として，単純な LLM，Retrieval-Augmented Generation（RAG），さらには再検索プロセスを考慮しシンプルな改良を施した RAG（以下再検索 RAG と略称する）をそれぞれ用い，公式の採点ガイドラインに基づいて評価を行う．RAG[3] は，外部データベースから動的に知識を組み込むことにより，LLM 単体では解答が難しい訓練時に含まれない専門的知識や，非公開情報を含む問題に対応可能である．さらに，この手法が LLM 内部の知識よりも与えられたコンテキストの情報を優先して利用することを踏まえると [4]，「国語」のような外部情報に依存して複雑な問題を解くタスクにおいて RAG が高い効力を発揮すると考える．再検索 RAG は RAG の改良版であり，シンプルな再検索プロセスを取り入れる．このアプローチにより，より採点条件を満たした答案を生成できるようになることを確認した．

本研究の貢献は以下である．

1. 国語科共通テスト試行調査の公式な採点基準を利用して RAG の性能を定量的に評価したこと
2. 答案生成のために再検索 RAG を提案し，より採点条件を満たした答案を生成できるようになる例を確認したこと

2　関連研究

これまでに，LLM を用いて国語の読解問題を解いている研究として，板橋ら [5] のものがある．彼らの研究では，GPT-3.5-turbo と GPT-4 を対象に，高校入試の国語物語文問題から 108 問の 4 択問題をランダムに選び，解答の正確性と理由の妥当性を人手評価している．この研究は高校入試レベルの選択肢形式問題に焦点を当てており，生成モデルの読解力を調査することに重点を置いている．一方，Oka ら [6] は，手書き認識と自然言語処理を統合し，日本の大学入学共通テストにおける記述問題の採点を完全に自動化しており，記述式問題の採点という応用

に重点を置いている．齋藤ら [7] は，模範解答との比較によって国語の記述問題を採点する手法を提案した．これに対し本研究では，従来では取り扱われていなかった大学共通テスト試行調査の国語問題を採用し，LLM を用いて実際の答案を生成することに加え，RAG および提案手法である再検索 RAG を用いた生成も行い，それぞれの手法を定量的に評価する．

3　国語科共通テスト試行調査による RAG の評価

本研究では大学入学共通テスト試行調査の国語の問題を用いて，RAG の評価を行う．大学入学共通テスト試行調査における記述式問題には，明確な採点基準が設けられている．解答は基本的に，無解答・完答（正解）・条件を満たさない解答（不正解）の 3 つに大別される．不正解については，さらに細分化され，さまざまな条件に基づいて別パターンの不正解として分類されているが，これらの分類には順位づけが存在せず，評価を行う際には正解か不正解のいずれかでしか指標を示せないため，数値的な分析が困難である．

そこで本研究では，各モデルによる解答生成結果を，条件をどれだけ満たしているかに基づいてスコアリングする手法を採用した．具体的には，完答を 1，条件を一つも満たさない解答を 0 とし，条件を一部満たす解答については，解答が満たした条件の数を全体の条件数で割ることで得られた値をスコアとして換算した．例えば，条件が 10 個ある場合，5 つの条件を満たした解答のスコアは 5/10 = 0.5 となる．この手法を用い，各手法・モデル・問題ごとに 5 回ずつ答案を作成させ，得られたスコアの中で最も頻出した値を評価値として採用した．評価手法については，定められた採点基準に基づいて評価値を算出した．採点者は第一著者のみとし，採点の一貫性を確保した．本稿では RAG を伴わない LLM 単体，RAG，また提案手法である再検索 RAG（4 節に後述）を比較し，評価する．

3.1　データ

データは，平成 29 年度・平成 30 年度に行われた，大学入学共通テスト試行調査の国語科目における記述式問題部分を全て抜粋して使用した．内容は，問題冊子を PDF として読み取ったものから，問題文と各設問に該当する部分を抽出してテキスト形式で記録したものとなっており，問題文と各設問はそれぞれ別のファイルで保存されている．なお，問題形式や解答形式によるシステム構成の変更は行わず，どのような形式の問題に対しても同一の方法で解くこととする．

4　再検索 RAG（提案手法）

4.1　RAG

本研究で導入した RAG は，Facebook AI Research（現 Meta AI）によって提案された手法であり，LLM と検索を組み合わせて質問応答や生成タスクを強化することを目的としている．この手法は，まずユーザーからの入力に基づいて，事前に用意した外部データベースから関連情報を検索する．データベースには，元の文章を適当な規則に従って分割したパッセージが格納されており，検索上位 k のパッセージを関連文書として紐づける仕組みになっている．そしてこの関連文書をそのままプロンプトに与え，LLM が文書をもとに生成する仕組みである．国語の答案生成においては，特に文脈に基づく適切な情報の抽出や，日本語特有の繊細な表現や文脈理解を補強することが非常に重要である．このため，RAG を用いることにより，国語の問題解決において重要な要素である「文脈理解」と「適切な情報の提示」の双方を強化できると考える．

4.2　再検索 RAG

答案生成のために，改良を加えた RAG を考案した．本手法は，基本的な構造において 4.1 節で述べた RAG と同一であるが，関連情報の検索過程において単純な改良が加えられている．その改良点は，得られた関連文書群を検索文書として再度検索プロセスを実行する仕組みを追加し，関連情報の拡張を可能にした点である．例えば，1 度目の検索で文書 A を検索文書として用いて，その関連文書である B と C を得たとする．再検索 RAG では，続いて B と C を検索文書として 2 回目の検索を行い，さらに関連文書を取得する．こうすることで，「【ア】に当てはまる適切な文を記述せよ。」のように，解答に必要なヒントとなる単語が明示されていない設問に対しても，本手法は二段階の検索を通じて対応する．こうすることで、初回の検索で得られた関連文書の周辺単語を新たな検索対象として用いることが可能

になり，より広範かつ文脈的な情報を収集することができる．この仕組みにより，データベース内で関連文書を効果的に拡張し，文書内参照が必要で直接的な文書取得が困難な問題に対しても，より高い解答精度を発揮することを期待している．

5 実験

5.1 LLM

本研究では，LLM 単体・RAG・再検索 RAG 全ての答案生成過程において，3 つのモデルを用いる．1 つめは，llama3 を日本語データで継続事前学習した elyza/Llama-3-ELYZA-JP-8B[1]である．Meta の llama アーキテクチャに基づいたデコーダモデルで，約 80 億パラメータを持つ中規模なモデルである．2 つめは，gemma2 を日本語データで事後学習した google/gemma-2-2b-jpn-it[2]である．Google の gemma アーキテクチャに基づいたデコーダモデルで，約 20 億パラメータを持つ比較的軽量なモデルである．3 つめは，OpenAI の開発した gpt-4o[3]である．これら 3 つのモデルについて，LLM 単体・RAG・再検索 RAG により生成された答案を，採点基準を基に採点して評価した．

5.2 プロンプト

図 1 にプロンプトの概要を示す．{context} には問題文や関連文書，{question} には設問が入る．なお，角括弧 [] に囲まれた一部の説明文については，LLM 単体での生成時には含まない．

```
以下の【テキスト】は問題文[の一部から抜粋したパッセージ]である。

【テキスト】
{ context }

つづいて【設問】をよく読み、【テキスト】を参照して解答せよ。

【設問】
{ question }
```

図 1 プロンプトの概要

LLM 単体での生成と，RAG を用いた生成の違いは，入力に関連文書として与えるテキストの構造にある．LLM 単体の場合，問題文をそのまま入力として渡すのに対し，RAG や再検索 RAG では独立した複数のパッセージとして関連文書を入力する．

検索によって拡張された関連文書については，LLM が読み取れる形式にテキストを成形し，プロンプトとして入力する．具体的には，以下のようなテキスト処理を施した：

- 問題文中の「ア」のように空欄部を表す単語について，設問や問題文内で文字間隔が統一されていない場合があったため，空欄部とその前後の文字の間に半角スペースを挿入した．
- 問題文中の傍線部で表される部分について，テキストデータで読み取りを可能にするため，鉤括弧 {} を用いた表示に変更した．
- 問題冊子の PDF データをテキスト化した際に発生する不要な改行や全角空白を除去し，文書を整形した．

これらの調整を行うことで，関連文書がより適切にプロンプト内で利用されるようにした．

5.3 RAG の実験設定

実験を行うにあたり，3.1 章のデータについてチャンキングを実施した．具体的には，1 つの大きなテキストデータを複数のチャンクとして分割し，それぞれ埋め込みを行いハッシュデータベースとして格納する．この過程において，原則として 100 トークン単位で分割を行い，複数改行が含まれる場合や句読点（特に句点）が含まれる場合は，意味的な近さを考慮してチャンクサイズが正確に 100 トークンでなくとも分割を実施可能にするという制限を設けた．また，チャンクごとに 50 トークン分の重複領域 (chunk_overlap) を持たせた．検索手法には Best Matching 25（BM25）を用いた．これはクエリと文書の関連性を計算する情報検索アルゴリズムで，TF-IDF を改良したランキングの手法である．検索時の関連文書取得本数は $k = 10$ とし，再検索 RAG については 2 回目以降の取得本数を $k = 3$ とした．なお再検索 RAG において取得した文書に重複が生じる場合には，入力トークン数を圧迫しないよう関連文書として加えないようにした．設問は問題に対して複数存在するため，検索や関連文書の取得は各設問ごとに行った．

1） https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
2） https://huggingface.co/google/gemma-2-2b-jpn-it
3） https://api.python.langchain.com/en/v0.1/chat_models/langchain_openai.chat_models.base.ChatOpenAI.html

6 実験結果

各手法とモデルによる答案生成結果について付録の表 1 に示す．LLM について比較すると，gpt-4o，llama3，gemma2 の順に最も良い結果となった．これはモデルのパラメータ数の大きさの順と等しく，大きなモデルを選ぶことで単純に答案生成の性能が上がることが分かる．つづいて，gpt-4o を用いた場合だけに着目すると，平成 29 年度は LLM 単体と再検索 RAG の結果が同率一位である．平成 30 年度は設問 01 と設問 02 は LLM 単体と再検索 RAG のどちらも正解しているが，設問 03 については再検索 RAG は 0.8 点なのに対して LLM 単体は 0.6 となり，再検索 RAG の方が良い結果となっている．

さらに RAG の有効性を見るため，それぞれの LLM を利用した場合の LLM 単体，RAG，再検索 RAG を比較すると，llama3 の平成 29 年度の設問 01 については RAG 手法が LLM 単体よりも評価が下がるものの，平成 30 年度の設問 03 においては RAG が有効である．また gemma2 を利用した際は，平成 29 年度の設問 01 が RAG の利用によって未回答から 0.75 に大きく上昇した．さらに，再検索 RAG を用いたスコアを RAG のスコアと比較すると，全設問において同等か上回る結果を出している．特に平成 29 年度においては，gemma2 を除いたほぼ全てのスコアが上昇しており，取得した関連文書群の質の向上がうかがえる．このことから，検索部分の改良の一例における有効性を確認した．

7 考察

本節では，答案の採点基準によって評価する際に問題となりうる点をを実例を参照しつつ述べる．付録の表 2 には，H29 年度試行調査における設問 02 の要件を示す．本問題について再検索 RAG を用いて生成を行ったところ，llama3 では「兼部の規定を緩和」という出力結果が得られたのに対し，同条件での gemma2 は「部活動の終了時間の延長」という出力結果であった．正答条件には『「体育部同士及び文化部同士の兼部」ということが書かれている』という項目が含まれるため，採点者の主観的な評価では，llama3 の生成結果の方が gemma2 の生成結果よりも適切であるように思われる．しかしながら，採点を行う際には正答条件を完全に満たしていなければ，その条件を満たしたとみなすことができず，加点することはできない．llama3 の「兼部の規定を緩和」という表現では，具体的にどのような兼部の条件を指しているのかが明確ではないため，結果として採点者の主観的評価に反する形となり，双方の生成結果ともこの条件に関しては加点されないという採点結果となった．このように，採点基準だけでは LLM が生成する答案を適切に評価しきれない例が複数存在した．ただし，この問題は，人間が回答した答案を採点する際にも生じる可能性があると考えられる．本研究では，人間の答案を採点する際と同じ枠組みを用いて，LLM が生成する答案の評価を試みた結果を示した．「文字数や構成などのルール」のように定量的に判断可能な基準だけでなく，上記のような主観評価で差が生じる「意味的な類似度」の評価をどのように行うべきかは，今後の課題である．現在，LLM を用いた含意関係認識などの意味的読解性能の向上を目指した研究が進められている．そのような技術を活用し，意味的な類似度を適切に評価できる採点ガイドラインや採点システムについて，今後検討していきたい．

8 おわりに

本研究では，大学入学共通テスト試行調査の国語科目における答案を事前学習済みモデルを用いて LLM 単体，RAG，再検索 RAG，それぞれの手法によって作成し，採点ガイドラインに基づいて評価した．結果として，RAG 手法の導入は特に中軽量モデルにおいて一定の向上が見受けられ，さらに再検索 RAG によるデータ拡張は有効である例を確認した．また，定量的な手法での評価には，意味的な近さの差を評価しきれないなどの課題があり，評価指標を改善できることもわかった．今後の方針としては，解答精度を上げるための改善がある．

謝辞

本研究は JSPS 科研費 JP22K12145，JP23K28201，JP24H00738 の助成を受けたものです．答案収集は，本学における人を対象とする研究に関する倫理審査委員会の承認を得て実施しました (No.230402-0411)．

参考文献

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020. https://arxiv.org/abs/2005.14165.

[2] R. Child D. Luan D. Amodei and I. Sutskever A. Radford, J. Wu. Language models are unsupervised multitask learners, 2019.

[3] 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. **arXiv preprint arXiv:2005.11401**, 2021. https://arxiv.org/abs/2005.11401.

[4] Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. **arXiv preprint arXiv:2406.12824**, 2024. https://arxiv.org/abs/2406.12824.

[5] 板橋康知, 松林優一郎. 物語文に対する大規模言語モデルの読解能力の分析. 言語処理学会 第 30 回年次大会 発表論文集. 一般社団法人 人工知能学会, 2024.

[6] Haruki Oka, Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa, and Tsunenori Ishioka. Fully automated short answer scoring of the trial tests for common entrance examinations for japanese university. **AIED (1), pp. 180–192**, 2022.

[7] 齊藤隆浩, 古宮嘉那子, 石岡恒憲, 中川正樹. Jglue データを用いた模範解答との差異に基づく汎用採点モデルの構築. 言語処理学会 第 30 回年次大会 発表論文集. 一般社団法人 人工知能学会, 2024.

表 1　各手法・各モデルごとでの生成結果における定量的評価，無回答の場合はハイフンで表すこととした．

		llm 単体			RAG			再検索 RAG		
問題	設問	llama3	gemma2	gpt-4o	llama3	gemma2	gpt-4o	llama3	gemma2	gpt-4o
H29	01	0.75	-	**1.0**	0.5	0.75	**1.0**	0.75	0.75	**1.0**
	02	0	0.33	**1.0**	0	0.33	0.33	0.33	0.33	**1.0**
	03	-	-	**0.5**	0	0	0.25	**0.5**	0	**0.5**
H30	01	0.67	0.33	**1.0**	0.67	0.33	**1.0**	0.67	0.33	**1.0**
	02	0.33	0	**1.0**	0.33	0	0.67	0.67	0	**1.0**
	03	0	-	0.6	0.2	-	0.6	0.2	-	**0.8**

表 2　H29 年度の試行調査における設問 02 の要件

問題 (一部)	.. 条件の緩和です。これまで認められてこなかった ア という要望です。 島崎 なるほど、分かりました。昨年も体育部・文化部の..
設問	問 2　空欄 ア に当てはまる言葉を、要望の内容が具体的に分かるように、二十五字以内で書け（句読点を含む）
正答条件	1. 25 字以内で書かれている 2. 文末表現が「という要望」に適切に続くように書かれている 3.「体育部同士及び文化部同士の兼部」（又は「体育部と文化部間以外の兼部」,「すべての部活動間での兼部」）ということが書かれている