文書情報構造認識のための AI chatbot プロンプト評価

中渡瀬秀一[1]

[1]大学共同利用機関法人 情報・システム研究機構 国立情報学研究所

概要

生成 AI の一種である AI チャットボットは，対話的に利用者の指示に応えることができる．この機能によって与えられた文書に対する処理(加工･認識など)を行うことも可能である．本稿では，生成 AI で文書の情報構造を認識するタスク（文献抽出）を処理するために用いるプロンプトの実験的評価を行う．実験では異なるレイアウトやフォーマットの論文に対して，先行研究の知見を参考にして設計した複数のプロンプトを適用してそれらの評価と知見の検証を行った．

1 はじめに

ChatGPT 等の AI チャットボットは LLM に基づいて構築されており，自然文による指示(プロンプト)に応えることができる．このシステムは知識を問う質問に対して回答できるだけでなく，与えられた文書の内容を認識して，翻訳・要約・情報抽出などを実行する能力も備えている．

一方，与えられたタスクに対して最適なプロンプトを設計するための理論は確立していない．先行研究ではプロンプトの効果を高めるフレーズの報告[4]や LLM の応答の正確さを向上させるプロンプト設計に関する原則の提案[5]がある．

本研究では，これらの研究を参考にしてプロンプトを設計し，文献抽出タスクに対する実験によってそれらのプロンプトの有効性を評価する．

以下２節では実験の詳細を説明し，３節では実験結果からプロンプトの評価を行う．

2 実験

本節では，プロンプト評価のための実験における文献抽出タスク・その対象文書・実験手順について説明する．

2.1 文書情報構造認識(文献抽出)タスク

このタスクは論文原稿の文書構造を認識して，文献一覧部分から参照文献を抽出するためのものである．それら論文で参照される文献はリスト化されて本文とは別に文献一覧として配置されている．このリストのタイトルには「引用文献」「参考文献」などが用いられている．またリストのフォーマットには箇条書きや改行せずに列挙するスタイルがある．

2.2 対象文書

実験で用いた文書は日本語で書かれた史学論文である．多様なレイアウト等に対するプロンプトの効果を比較するために，3 種のレイアウト(縦書き・横書き・段組)や 3 種の文献フォーマット(参考文献・引用文献・注)で書かれた 3 篇[1,2,3]の論文を選定した．以下に各論文の特徴を示す．論文[1,2]は横書き，論文[3]は縦書きのレイアウトである．段組みは論文[1]が 1 段組み，論文[2,3]が 2 段組みである．参照文献一覧部分のタイトルには引用文献・参考文献・注がある．その一覧は原稿中で末尾または中間に位置している．

表 1　文書の特徴

論 文番号	レイアウト	文献一覧タイトル	文献一覧位置
[1]	横書 1 段組	引用文献	末尾
[2]	横書 2 段組	参考文献	末尾
[3]	縦書 2 段組	注	中間

2.3 実験手順

この実験では ChatGPT（GPT-4）を使用した．プロンプトの使用前には論文の PDF ファイルを ChatGPT にアップロードしておく必要がある．次

に論文[1]から順に単純な初期プロンプト「論文から**引用文献**のリストを作成して下さい」等[i]を適用する．もし文献抽出が不十分な場合には先行研究[4,5]を参考にしたプロンプト改良を行い，再度適用する．この結果からプロンプトの抽出効果を評価する．

プロンプトの改良については，Yang[4]らの報告による特定フレーズ[ii]の追加と Bsharat[5]らが提案するプロンプト設計のための基本原則(全 26 種からタスクに関連するもの)の応用を行う．具体的な内容は次節で説明する．

3 評価

本節では実験結果からプロンプトの評価を行う．

3.1 実験結果

- 論文[1]：横書き 1 段組みの論文
  初期プロンプトを適用した結果，論文末尾にある文献リスト(11 件)が全て正確に抽出された．
- 論文[2]：横書き 2 段組みの論文
  初期プロンプトを適用した結果，論文末尾にある文献リストから文献情報が正確に抽出されたが ChatGPT の出力サイズの制約から一度に全件抽出はできない．そのため追加のプロンプト「N 件目から M 件目を出力して下さい」を複数回用いることで全件抽出が可能であった．ただし N や M の指定に正確に応答しない場合がある．なお元の原稿フォーマットでは図 1 のように同じ著者が連続する場合，2 回目以降が「—」と表記されていたが，ChatGPT はそれらを正確な著者名に置換して出力している．これにより ChatGPT は単純にリスト領域の抽出をする以上の処理を行っていることが確認された．

赤澤史朗（2011）「1950年代の軍人恩給問題（1）」『立命館法学』333/334，1461-1492。
————（2012）「1950年代の軍人恩給問題（2・完）」『立命館法学』341，511-552。

図 1　論文[2]の参照文献フォーマットより

- 論文[3] ：縦書き 2 段組みの論文
  ・この論文に対しては，初期プロンプトだけでは文献リストが認識できず，次の応答メッセージが出力された．
  「文献リストが論文のどの部分に含まれているかを特定するのに苦労し、適切な範囲に到達する前に時間が終了してしまいました」
  ・次に PDF ファイルを分割し文献リストが含まれるページだけをアップロードし，初期プロンプトに「論文は日本語です。レイアウトは縦書きで2段組みです」を追加して再度適用した．その結果，リスト中の英語論文情報だけが正確に抽出された．ただし文献タイトルには原文にない引用符が付加されている．論文[2]と同様に単純な抽出ではなく認識結果に加工が施されている．
  ・抽出が不十分であるためさらに「日本語の文献と英語の文献を出力して下さい」のフレーズを追加して再度適用した．この結果，2 件(全 10 件)の日本語文献が部分的に抽出された．
  ・以上に加え，次に Yang[4]らの特定フレーズ「深呼吸をして、この問題に一歩ずつ取り組んでください」や Bsharat[5]らの原則の応用として「全ての文献が抽出できれば報酬を出します」，「全ての文献が抽出できないとペナルティを与えます」，「###指示」，「###ヒント」を個別に追加したプロンプトによる指示を行った．このうち特定フレーズと報酬・ペナルティに関するフレーズには改善の効果が見られなかったが，指示（やヒント）を構造的に明示した場合に追加で 4 件（5 件）の日本語文献情報が一部不正確な形で抽出された．

3.2 評価

本タスクの処理パフォーマンスは対象とする文書の処理難度とプロンプトの適切さの 2 点に依存する．それぞれを評価する．

- レイアウトによる処理の難度
  横書きの場合，段組み数に関わらず単純なプロ

[i] 「引用文献」の部分は論文に応じて「参照文献」や「注」を使用する

[ii] LLM の学習データに含まれる Q&A フォーラムデータの中で注意深く推論する場合に特定フレーズが使用されることが指摘されている．Telling AI model to “take a deep breath” causes math scores to soar in study. https://arstechnica.com/information-technology/2023/09/telling-ai-model-to-take-a-deep-breath-causes-math-scores-to-soar-in-study/.

ンプトで正確な抽出が可能である．しかし縦書きの場合は，一部の文献だけが認識されているため縦書き文書の処理は難易度が高い．

- プロンプト改良の効果
  改善効果が報告されている Yang[4]の特定フレーズに関して本タスクでは効果が見られなかった．フレーズの効果は適用する文脈に依存する可能性がある．また Bsharat[5]の原則などで指摘されているプロンプト形式の構造化は，本タスクにおいても有効であった．

4 おわりに

生成 AI を用いて文書の情報構造を認識するタスクにおいて，プロンプトの設計を行い，実験でその効果を検証した．具体的なタスクとしては論文からの参照文献抽出を設定した．また先行研究の成果に基づきプロンプトを設計した．対象論文には日本語の史学論文からレイアウト(縦書き・横書き・段組)や文献フォーマット(引用文献・参考文献・注形式)が異なる 3 篇を使用した．

ChatGPT（GPT-4）を用いた実験の結果，横書きレイアウトの論文については単純なプロンプトで正確な抽出が可能であった．しかし縦書きで注形式の論文に対してはプロンプトを改良しても正確で完全な抽出はされなかった．先行研究の知見に基づいたプロンプト改良では，Yang[4]らによる特定フレーズの追加による効果は見られなかったが，プロンプトのフォーマットを構造化することには改善効果が見られた．また文書の特徴をヒントとして示すことも有効であった．

なお ChatGPT の LLM は継続的に新しいデータによる学習が行われている．そのため 2023 年 12 月に行われた本実験の結果が今後も再現するのかは不明である点には留意されたい．

謝辞

本研究は JSPS 科研費 基盤研究(C) 21K02646 の助成を受けたものです．

参考文献

1. 中尊寺文書の基礎的検討．**菅野文夫**，出版地不明 ： 岩手大学平泉文化研究センター年報，2021 年，第 9 巻. 37-52.
2. 戦傷病者戦没者遺族等援護法と更生医療 ： 戦後復興期の京都府を事例として．**山下麻衣**，出版地不明 ： 障害史研究，2023 年，第 4 巻. 13-30.
3. 日本思想史と災厄．**上野太祐**，**石原和**，**殷暁星**，**加藤真生**，**村上晶**，出版地不明 ： 日本思想史学，2021 年，第 53 巻. 37-53.
4. **Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, Xinyun Chen.** Large Language Models as Optimizers. 出版地不明 ： https://arxiv.org/abs/2309.03409, 2023-12 閲覧.
5. **Sondos Mahmoud Bsharat, Aidar Myrzakhan, Zhiqiang Shen.** Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. 出版地不明 : https://arxiv.org/abs/2312.16171v1, 2023-12 閲覧.