E-015

DB 構造化による RAG を用いたトークン数制約下での小型 LLM システムの性能改善
Enhancing the Performance of Small LLM System Under Token Limitations with DB-Structured RAG

仁科 颯† 平田 俊明†
Hayate Nishina Toshiaki Hirata

1. はじめに

LLM は質問応答や生成タスクにおいて高い性能を発揮しているが，学習データに存在しない知識を扱う際には「幻覚（hallucination）」や情報不足といった課題がある．これらの課題を解決する方法として，外部知識ベースと LLM を連携させる RAG は有望な技術の一つである[1]．特に，計算資源の観点から，より小型の LLM を利用する際には，限られたトークン数の中でいかに効率的に外部知識を検索し，LLM に提示するかが重要な課題となる．

従来の RAG システムで一般的に用いられるチャンク分割手法は，実装が容易である一方，情報を固定長で分割するため，文脈が途切れたり，一つのチャンク内に意味的な関連性が低い情報が混在したりする問題がある．これにより，検索精度が低下し，結果として LLM に与える情報の質が損なわれる可能性が指摘されている[2][3].

本研究では，このトークン数制約と情報品質の問題に対し，文書をその意味構造に基づいて YAML 形式で構造化し，ベクトルデータベースと組み合わせることで，トークン数を削減しながら応答性能を向上させる新たなアプローチを提案し，その有効性を検証することを目的とする．

2. 関連研究

(1) チャンク分割

従来の RAG システムにおける一般的な文書処理方法として，チャンク分割がある．この手法では，文書を固定長の単位に分割し，各チャンクをデータベース化する．実装が容易である一方，前述の通り，情報が複数のチャンクに分散してしまったり，意味的な関連性が低い情報が同じチャンクに含まれてしまうことで，検索精度や LLM への入力情報の質が低下する可能性がある．提案手法では，文書の構造を維持する YAML 形式を活用することでこの課題に対処する．

(2) GraphRAG を用いた構造化データ検索

GraphRAG[4]は，複数の文書間の関係性をグラフ構造として捉え，より文脈に即した情報を検索することで応答精度を高める手法である．一方で，グラフ構造の構築や維持に手間がかかり，計算リソースが高くなりがちな課題がある．提案手法では，より軽量で可読性の高い YAML 形式による構造化とベクトルデータベースを組み合わせることで，計算コストを抑えながら情報を取得することを目指す．

3. 提案手法

提案手法では，文書を YAML 形式で構造化し，ベクトルデータベースである Chroma DB に格納することで，トークン数制約下における効率的な外部知識検索と LLM による応答生成を実現する．

† 東京情報デザイン専門職大学
Tokyo Information Design Professional University

3.1 システム構成

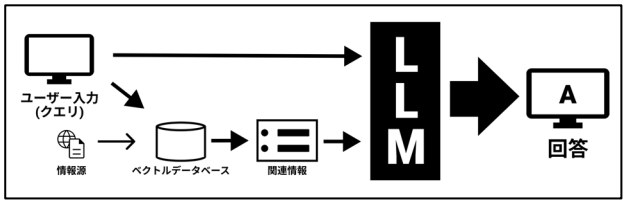

図 1　システム構成図

提案手法のシステム構成を図 1 に示す．ユーザーからのクエリは，まずベクトルデータベースに登録された YAML 構造化文書の検索に使用される．検索された関連情報は LLM にプロンプトとして入力され，最終的な応答が生成される．LLM は Llama3 8B[5]の日本語チューニングされた llama.cpp 量子化版[6]である， Llama3-ELYZA-JP-8B-q4_k_m を使用した．ベクトルデータベースは Chroma DB[7]を，埋め込みモデルとして multilingual-e5-large[8]を使用した．

3.2 情報の YAML 形式による構造化

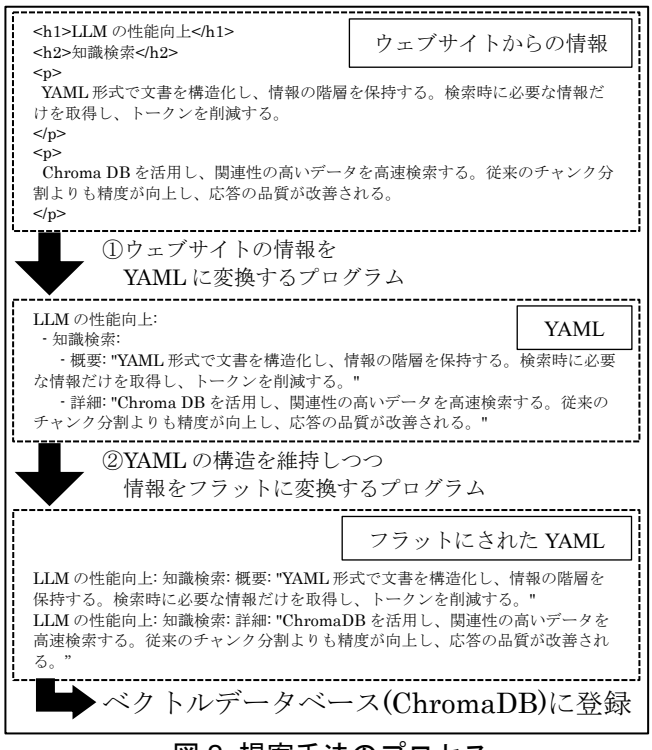

図 2 提案手法のプロセス

提案手法のプロセスを図 2 に示す．提案手法では，まず情報源である Web ページの HTML を Readability.js[9]による本文 HTML の抽出を行う，それを独自に開発したプログラム（図 2 ①）を使用して，見出し（<h1>, <h2>など）やリスト構造を基に，文書の階層構造を維持したまま YAML 形式に変換する．これにより，情報の親子関係や関連性が明示的に保持される．さらに，この階層構造を持つ YAML

データを，LLM が解釈しやすく，かつデータベースに格納しやすいように，プログラム（図 2 ②）を使用して，情報を圧縮した一行の文字列形式に変換する．文字列形式への変換は，YAML の持つ立体的な入れ子構造を，各一行一行が自身のすべて親を持つ，平面的な文字列へと変換する．これらを ChromaDB のカラムとして登録する．

4. 実験

提案手法の有効性を検証するために，既存のチャンク分割手法との比較実験を行った．

4.1 実験設定

実験では，東京情報デザイン専門職大学の Web サイト (https://tid.ac.jp)の sitemap.xml に記載された URL から収集した情報をデータセットとして使用した．このデータセットに対し，提案手法とベースラインとなる従来のチャンク分割手法を適用した．提案手法では，収集した Web ページを独自開発のプログラムで YAML 形式に構造化し，Chroma DB に登録した．一方，チャンク分割手法では，Web ページを固定長のチャンクに分割して登録した．

評価には，収集した Web サイトの内容に基づいて独自に作成した質問データセットを用いた．各質問に対して両手法で応答を生成させ，その品質を人間が 4 段階で評価した．評価基準は，質問の意図を完全に理解し，的確かつ必要十分な情報を過不足なく提供している場合を「4」．質問に的確に回答できているが，些細な情報不足，わずかに冗長な部分，表現の改善の余地など軽微な改善点がある場合を「3」．質問の主要な点には触れているが，看過できない情報不足，冗長性，回答の要点が掴みにくい，または部分的に不正確な点がある場合は「2」．質問の意図を全く理解してせず，見当違いな回答をしている，または明らかな誤情報を含んでいる場合は「1」とした．

4.2 実験結果

提案手法とチャンク分割手法における評価件数の増減率を 表 3 に示す．提案手法はチャンク分割手法と比較して，評価「4」の件数が 16.0 ポイント，評価「3」の件数が 4.6 ポイント向上し，評価「2」の件数が 14.6 ポイント，評価「1」の件数が 6.0 ポイント減少した．

表 3 提案手法とチャンク分割手法における
評価件数の割合と比較

評価	提案手法	チャンク分割	差(pp)
4	53.51%	37.50%	16.01pp
3	33.11%	28.44%	4.67pp
2	10.37%	25.00%	-14.63pp
1	3.01%	9.06%	-6.05pp

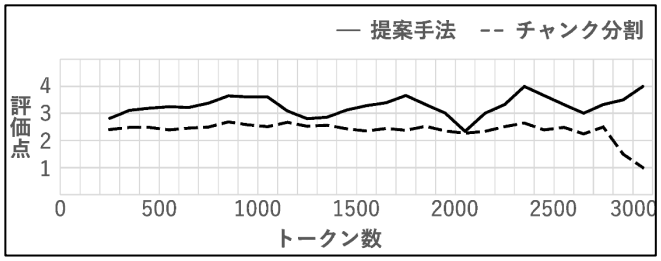

図 4 100 トークンごとの平均評価スコア

また，LLM に与えられた RAG コンテキストのトークン数を 100 ごとの平均評価スコアをプロットしたグラフを 図 4 に示す．このグラフから，提案手法は，チャンク分割手法と比較して，より少ないトークン数で高い平均評価スコアを達成していることが観察される．

5. 考察

実験の結果，提案手法は従来のチャンク分割方式と比較して，総合評価で優位性を持つことが示された．高品質な応答（評価「4」「3」）の割合が増加し，低品質な応答（評価「2」「1」）が減少したことから，応答品質全体が向上したことがわかる．また，小型 LLM の欠点である冗長な繰り返しの生成が[10]，チャンク分割手法と比べ，提案手法では抑制される傾向にあった．これは，構造化されたデータが応答の質を高めるだけでなく，LLM の動作を安定させる効果も持つ可能性を示唆している．

一方で，「キャンパスの所在地」を問う質問では，チャンク分割手法が適切な情報を抽出したのに対し，提案手法は応答できなかった．これは，文書を構造化する過程で特定のキーワード情報が埋もれてしまい，ベクトル検索が効果的に機能しなかったことを示している．情報の階層化が，質問の種類によっては検索精度を低下させる課題があることが明らかになった．

6. おわりに

本論文では，小型 LLM を用いた RAG システムにおけるトークン数制約と応答品質の問題に対処するため，文書を YAML 形式で構造化する手法を提案した．Web サイト情報を用いた実験の結果，提案手法は従来の固定長チャンク分割手法と比較して，LLM への入力トークン数を削減しつつ，応答の精度を向上させることを実証した．これにより，文書の構造情報を活用することが，RAG システムにおける外部知識の効率的な利用と LLM の応答品質向上に有効であることが示された．本手法は，追加学習を必要とせず実装が容易であるため，コストを抑えつつ既存の RAG システムの性能を改善するアプローチのひとつとなりうる．

参考文献

[1] P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401 [cs.CL], 2020.
[2] 江上周作, 福田賢一郎. "文書のチャンクに基づく知識グラフを活用した RAG". 言語処理学会 第 30 回年次大会 発表論文集, pp. 2455-2460 (2024).
[3] Bhat, S. et al., "RETHINKING CHUNK SIZE FOR LONG-DOCUMENT RETRIEVAL: A MULTI-DATASET ANALYSIS". arXiv:2505.21700v2.
[4] Sepasdar, Z. et al., (2024). "Enhancing Structured-Data Retrieval with GraphRAG: Soccer Data Case Study." arXiv preprint arXiv:2409.17580.
[5] Aaron, G. et al., (2024). The Llama 3 Herd of Models. arXiv preprint arXiv: 2407.21783.
[6] https://github.com/ggml-org/llama.cpp
[7] https://www.trychroma.com/
[8] https://huggingface.co/intfloat/multilingual-e5-large
[9] https://github.com/mozilla/readability
[10] Mahaut, M. et al., (2025). Repetitions are not all alike: distinct mechanisms sustain repetition in language models. arXiv preprint arXiv:2504.01100.