議会会議録に対する表構造理解に向けた RAG 及びマルチモーダルシステムの構築について

2021344　前多陸玖

令和 6 年度提出

目 次

第1章　はじめに

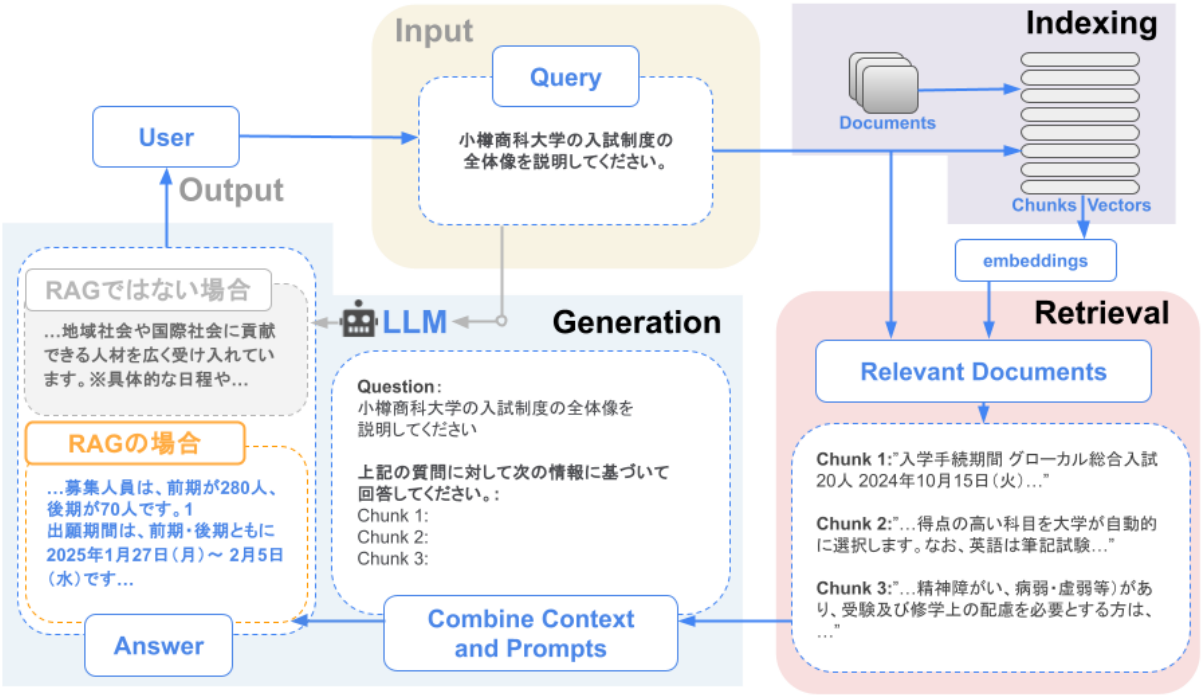

図 1.1: RAG の処理の流れ．青い導線が RAG，灰色の導線が RAG ではない場合の処理を表す．

近年，大規模言語モデル（LLM）における追加学習なしに外部知識を扱う手法として，RAG (Retrieval-Augmented Generation) が用いられている [1]．図 1.1 は RAG の大まかな処理の流れである．RAG の処理は大きく Indexing（インデックス作成），Retrieval（検索），Generation（生成）の 3 つの段階に分割される [2]．例えば，ユーザーが「小樽商科大学の入試制度の全体像」について LLM に対して RAG を用いて質問した場合と，そうでない場合を比較する．

RAG を用いる場合，Index の段階では小樽商科大学の入試制度にかかわりのある文書の PDF や HTML などの情報を Documents として扱う．Documents は任意に設定することのできる chunk Size ごとに分割され，embedding model によりベクトル化されたのち，ベクトルデータベースに格納される．次に Retrieval の段階では，ベクトル化された各 chunk とユーザーの質問に対して意味的類似度を求め，質問に最も高い上位 k 個の chunk を検索する．最後の Generation の段階では，元の質問と検索されたチャンクを一緒に LLM に入力し，最終的な答えを生成する．一方で，RAG を用いない場合では，ユーザーの質問は LLM に対して直接渡され，回答を出力することになる．RAG は情報を検索し生成時に活用することで，LLM による誤った回答（ハルシネーション）のリスクを軽減させ，信頼性の高い出力を可能としている．

一方で，RAG で使用する文書に HTML や表・画像等が含まれている場合，出力の精度が低下することが確認されており [3]，適切な前処理が必要である．またベクトルインデックス作成の際

に使用する embedding model や chunk Size，Overlap Size が出力に影響を与える [4].

多くの企業が社内情報や顧客情報を RAG に用いているが，それらの文書に含まれている表への対応に苦慮している．また今後活用が見込まれる行政や金融，教育分野などにおいても，議会会議録（予算表）や有価証券報告書，学術論文に含まれる表が問題となり得る．そのため表への適切な対応は，RAG における大きな問題だといえる．

本研究では，MBLink（Mineutes-to-Budget-Linking）で提供されている小樽市の令和 4 年度の議会会議録と予算表を対象として，表の含まれる文書の前処理において，マルチモーダル LLM(M-LLM) を使用することによる，検索における精度への影響を検証する．また，embedding model 間の性能差や chunk Size，Overlap Size の値の影響を検証する．

本研究の貢献は，以下の 3 点である．

- M-LLM を用いた表構造理解の手法を提案した．
- HTML 形式で記述された表と，M-LLM を通してマークダウンに近い形式に変換した表のそれぞれをデータベースとして，RAG を適用した場合にどのような差があるのかを検証した．
- 複数の embedding model や chunck Size, Overlap Size において，それぞれの差を比較した．

以下，2 章では議会会議録に対する自然言語処理の現状の関連研究と，表構造理解に関するセル分類の関連研究について述べる．また LLM を用いた表構造理解の難点についても述べる．3 章では，本研究で提案する RGA のシステムについて説明し，従来手法との違いを述べる．4 章では，MBLink のデータセットに対して，HTML 形式で記述された表に対して RAG による検索を行う従来手法と，本研究で提案する手法とを比較する実験について説明を行う．また提案手法に対して，chunk Size 及び Overlap Size の複数の組み合わせにおける，OpenAI が提供する OpenAIembeddings の text-embedding-ada002，text-embedding-3-small，text-embedding-3-large ごとのモデル間の違いを検証する．そして，それらに対する評価手法についても説明を行う．5 章では各実験の結果についてまとめ，それらの実験結果からわかることを述べる，6 章ではこれら実験に対する考察を行い，提案手法の優位性や適切なモデルやハイパーパラメータ設定について述べる．最後に 7 章では全体をまとめる．

第2章　関連研究

2.1　議会会議録に関する研究

議会会議録とは，地方議会において議員や首長が，予算や条例の制定などの審議を行う際の発言をすべて文字に書き起こし，地方自治法第 123 条に基づき保存したものである．インターネットの普及により，議会会議録をウェブ上に PDF ファイルや HTML ファイル，テキストファイルなどのフォーマットで公開されている．

議会会議録を用いた取り組みとして「ぎ〜みる[1]」があげられる.「ぎ〜みる」は都道府県議会の本会議録を検索・視覚化できるウェブアプリケーションである．全文検索エンジンに Elasticsearch を採用し,「発言検索」や「TF・IDF キーワード検索」「時系列グラフ」などの機能を提供している [5]．この取り組みの問題点としては，全文検索であるためユーザーが欲しい情報を入手するのに手間が掛かってしまう点，生成 AI を組み込んでいないためユーザーとの柔軟なやり取りを行えない点があげられる．

2.2　表構造理解に関する研究

奥山ら（2023）は有価証券報告書に含まれる表において，どのようなセルが機械判読であるかを，一般的な機械学習のアプローチから明らかにした [6]．その中で，有価証券報告書内の機械判読が困難な表を「小見出し行」「複数の Header（セル結合を含む）」「空白セル」「非スカラー値」「特殊な形」の 5 つのタイプに分類した．

Ye ら（2023）は LLM を用いた表形式データの推論の向上を目的とし，表形式データを小さな集合に分割し，複雑な質問をより単純な部分質問に分割することで，より正確な検索が可能であることを示した [3]．その中で，LLM が表構造データの推論を不得意とする理由について,「LLM のトークン制限を超過する巨大な表形式データ」「比較，集計，演算などを必要とする複雑な質問」「テキストと表形式データの複雑な相互理解」の 3 点にあると指摘した．また，従来の表形式推論の手法が，列やヘッダーなどの表の構造的な情報を主に利用し，セルのテキスト情報が重視されていなかった点を指摘し，LLM によりセルテキストの意味を抑えた表構造理解が可能であることを示した．

Wang ら（2021）は GTR（Graph-based Table Retrieval）を用いて，表形式グラフを入力とすることで，表のセルや行列の特徴を抽出しする手法を提案した [7]．これにより，既存の BERT4TR（BERT for Table Retrieval）や TaBERT と比較して表の構造的な依存関係を保持することで，複雑な表のレイアウトをとらえることが可能となった．

[1] http://local-politics.jp/

第3章　提案システム

3.1　対象データ

本研究では MBLink に含まれる小樽市の令和 4 年度の議会会議録と予算表を対象として実験を行う．MBLink は NTCIR-17 QA-Lab PoliInfo-4[1]のサブタスクである．議会の予算審議において，ある予算に関する発言に対して，関連する予算表のセルを紐づけるタスクである [8] [9]．NTCIR-17 は，NII が主催する評価型ワークショップであり，2022 年 7 月から 2023 年 12 月まで開催された[2]．MBLink で使用されるデータセットは，議会会議録中の市長の発言文と予算説明書などに含まれる表，それぞれの HTML からなり，GitHub 上で公開されている[3]．MBLink において HTML にはそれぞれ識別用の ID が割り振られており，入力されたテキストに対して関連した複数の表を紐づけている．図 3.1 は MBLink における識別 ID を用いたテキストと表の紐づけ例である．入力されたテキストに対して関連のある表を紐づけている．

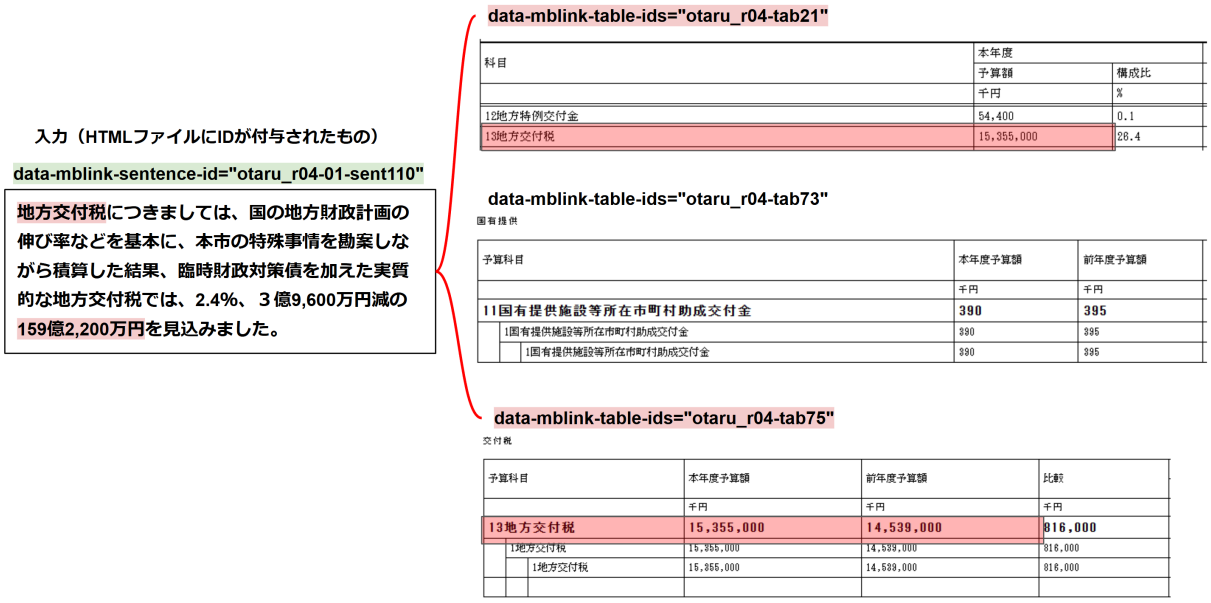

図 3.1: MBLink において識別 ID を用いてテキストと表を紐づける例

本研究では，MBLink のデータセットの中でも，小樽市の令和 4 年度の議会会議録から作成された HTML ファイルを使用する．Sentence ID の付与された議会会議録のテキストと，そのテキストに対して正解の Table ID の付与されている表の HTML ファイルをのそれぞれ 46 個を対象とする．

[1] https://sites.google.com/view/poliinfo4
[2] https://research.nii.ac.jp/ntcir/ntcir-17/
[3] https://github.com/poliinfo4/PoliInfo4-FormalRun-Minutes-to-Budget-Linking

3.2 表の前処理

3.2.1 前処理における基本的な流れ

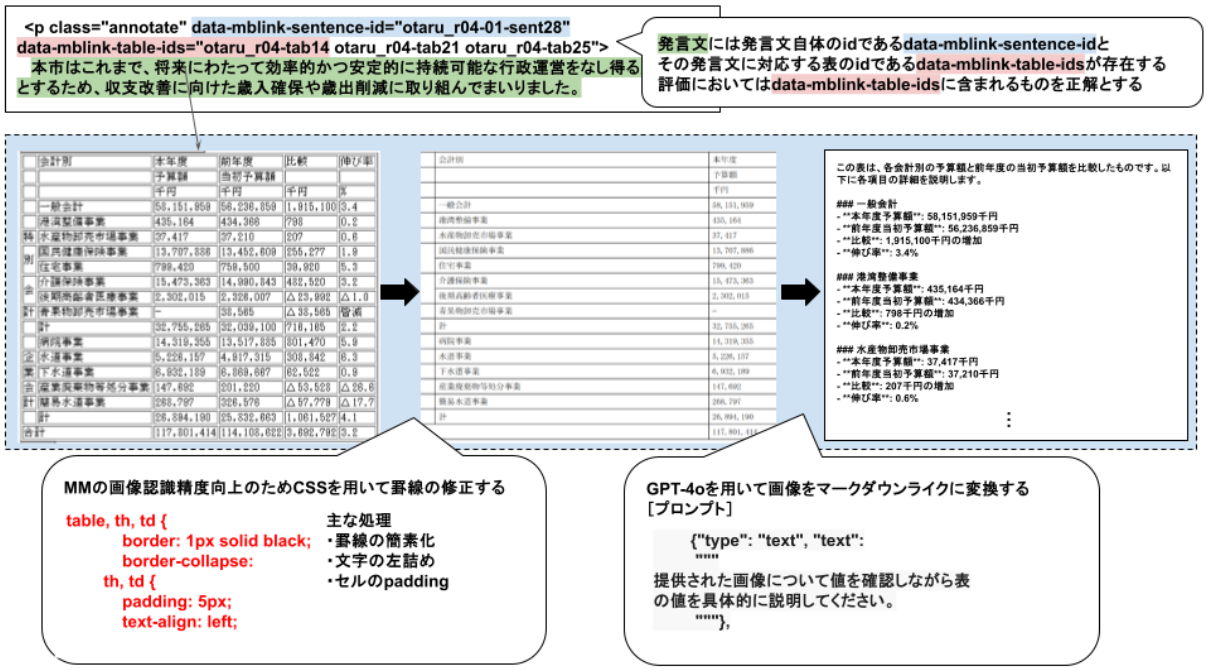

図 3.2: 表の前処理の流れ

本研究では，表の前処理において OpenAI が開発した M-LLM である GPT-4o（gpt-4o-2024-05-13）を使用する．図 3.2 は表の前処理の具体例である．GPT-4o を用いて表をマークダウン形式に変換する上では，各 Table ID ごとに対象となる表の画像が必要となる．まず，MBLink に含まれる令和 4 年度の予算説明書は一つの HTML ファイルに収められているため，table タグごとに HTML ファイルを分割する．その後対象となる表かどうかを判別し，使用する 46 個の HTML ファイルを作成した．

また表の HTML ファイルを分割する作業と並行して，HTML のスタイルシート（CSS）を一部書き換えた．CSS を変更したことにより，表の罫線が二重線から単純な直線へと置換される．これにより M-LLM による画像認識の精度を大きく向上した．そのように Table ID ごとに分割された HTML ファイルに対して，Selenium を用いて HTML ファイルをブラウザ上で開いたのち，表のスクリーンショットを撮影した．

3.2.2 前処理における特殊な処理

図 3.3 は表の前処理におる特殊な例である．HTML ファイルの分割に際して，表がページを跨ぐ場合，表の Header 情報が欠損することがある．このような場合には，対応する Header を持つ表からヘッダー部分をキャプチャし，Header 画像と表画像の両方を GPT-4o に入力した．

表の画像を GPT-4o に入力し，マークダウンに近い形式（以下，マークダウン形式）に変換する際に GPT-4o の出力トークンの問題から，表のすべての要素を出力することができなかった場合には，プロンプトに前回の出力を含めて入力することで，表の続きを生成した．また出力ごとに表形式データを生成し終えたかを確認し，すべての生成が終了するまで同様の処理を繰り返した．

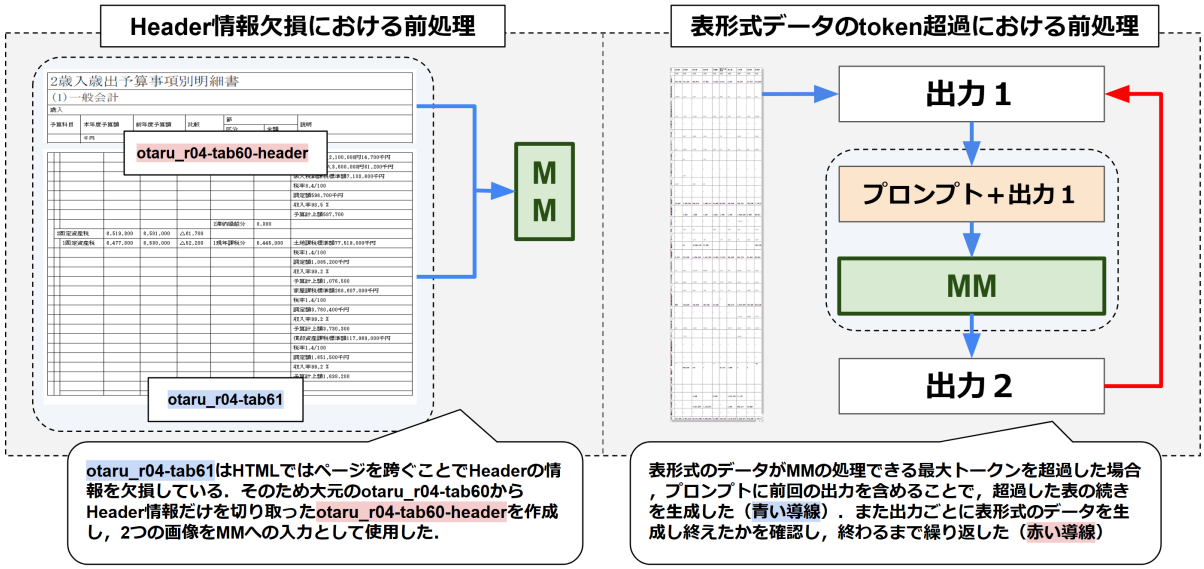

図 3.3: 表の前処理における特殊な例

具体的には生成ごとに，表形式データをすべて生成できたか M-LLM に判定してもらい，生成できなかった場合には「続きがある」「続きがあります」という言葉をともに生成させた．M-LLM の出力に対して正規表現でこれらの言葉を検知することができた場合は繰り返し処理を，そうでない場合は処理を終了させた．

3.3 提案手法の概要

表の前処理を経て，マークダウン形式に変換した表形式データを使用し，RAG を作成する．RAG の作成に関しては Python のライブラリである LangChain[4]を用いる．LangChain は LLM に基づいたアプリケーションを構築するためのオープンソースフレームワークであり，プロンプトテンプレートやチェーン，インデックス，メモリーといった RAG 作成における基本的な機能を提供している．

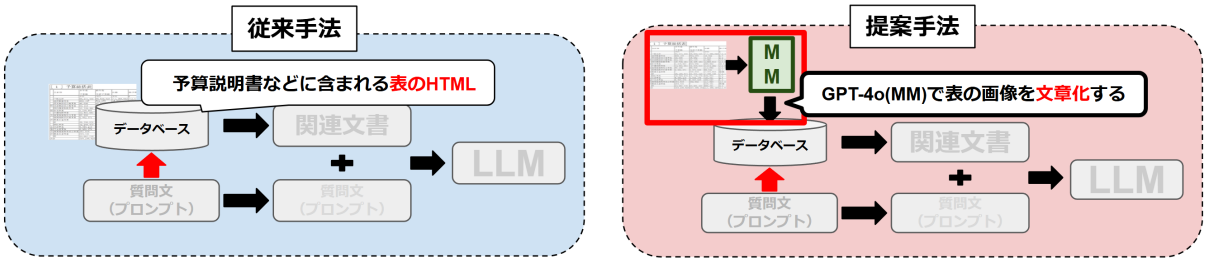

図 3.4: 従来手法と提案手法の RAG の違い

図 3.4 は従来手法と提案手法の RAG の違いを表したものである．基本的な RAG の処理は変わらないが，データベースに大きな違いがあることがわかる．従来手法では，データベースに加工を施してしない各表の HTML を用いる．各 HTML は，langchain が提供する UnstructuredHTMLLoader の機能を用いて，データの読み込みを行い，ベクトルデータに変換される．一方で提案手法では，

[4]https://www.langchain.com/

M-LLM によって表の画像をマークダウン形式に変換した文章をテキストとして使用する．各テキストは，同じく langchain が提供する TextLoader の機能を用いて，データの読み込みを行い，ベクトルデータに変換される．従来手法と提案手法における各ベクトルデータは chunk Size ごとに分割され，入力との意味的類似度を算出したのち，上位 k 件を LLM に渡されることになる．

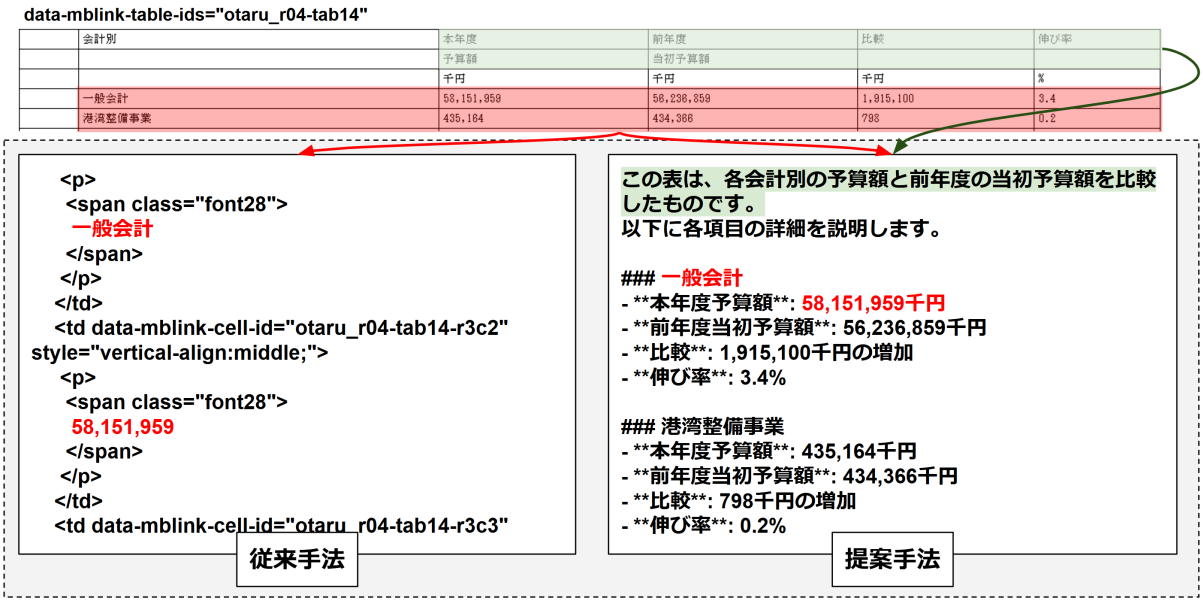

図 3.5: 従来手法と提案手法におけるデータベースの違い

図 3.5 は従来手法と提案手法における使用するデータベースの違いの例である．示された表の赤で囲われた部分に対して，従来手法では HTML 形式で，提案手法ではマークダウン形式で記述されていることがわかる（赤い導線）．

また提案手法において，M-LLM を用いたことによる，マークダウン形式だけではない，独自の記述がなされていることがわかる（緑色の導線）．提案手法の冒頭の「この表は，各会計別の予算額と前年度の当初予算額を比較したものです。」という記述は，M-LLM に入力する HTML には含まれない情報である．そのため，表の画像を受け取った M-LLM が，表の緑色で囲われた Header 情報を参照したうえで，対象の表が『各会計別の予算額と前年度の当初予算額を比較』した表であると，意味的に理解したうえでマークダウン形式を記述したことがわかる．

このような M-LLM を使用することによる，表の要約・意味的解釈は，従来手法と提案手法を比較する上で，明確な違いであるといえる．これにより従来手法では HTML で記述された表の情報をテキストコンテンツの集合としてとらえていたのに対して，提案手法では表の意味を理解したうえでの検索が可能になると考える．

第4章　実験

4.1　提案手法の検証実験の設定

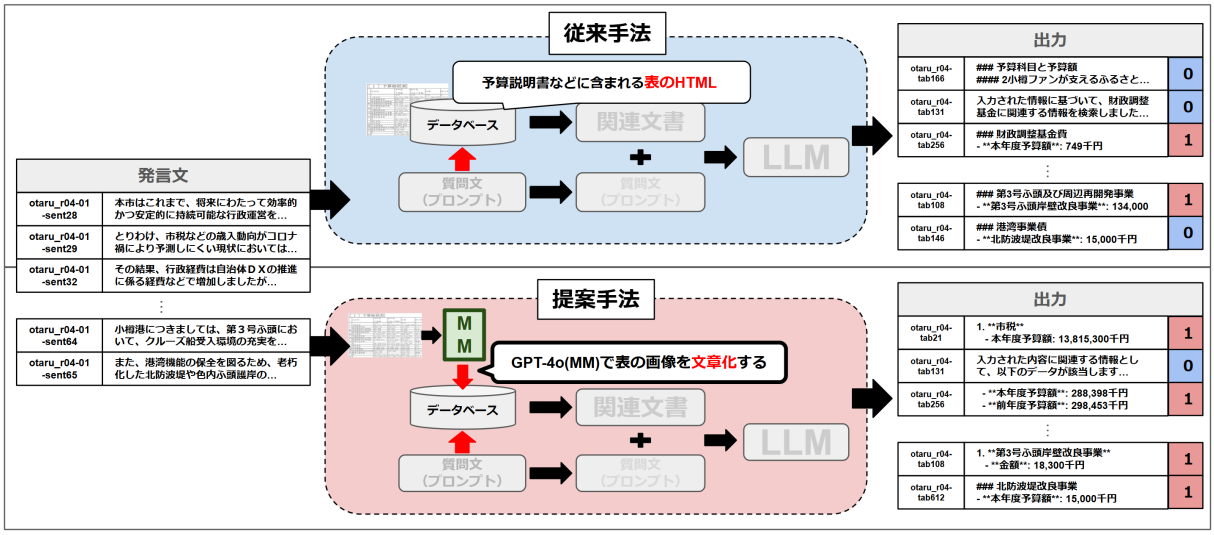

図 4.1: 提案手法の検証実験のイメージ

提案手法の検証においては，embedding Model として OpenAI embedding（text-embedding-ada-002）を用いる．図 4.1 は提案手法の検証実験のイメージである．未処理の表の HTML を OpenAI embedding を通したベクトルデータベースと，マークダウン形式に変換した表のテキストを OpenAI embedding を通したベクトルデータベースを作成する．作成したデータベースに対して，議会会議録のテキストを入力として，ベクトル検索を行い，類似度の高い上位 3 件の表を LLM に渡し回答を生成する．この際 chunk Size は 200 から 1000 までを 200 区切りで求め，Overlap Size は設定しない．回答を生成する LLM には gpt-4o-2024-05-13 を用いる．

4.2　モデル・パラメータ検証実験の設定

図 4.2 はモデル・パラメータ検証のイメージである．embedding Model や chunk Size，Overlap Size といったパラメータは，Indexing の段階で用いる．embedding Model は文字列をベクトル変換する際に用いる．モデルごとに次元数や学習データが異なり，変換後のベクトルも異なる．chunk Size は Documents をどのくらいの単位で分割するかを定めるものである．chunk Size が小さい場合，検索精度は高まるが，関連性の低い断片も増える可能性がある．一方で大きい場合，情報の全体像をとらえやすいが，不要な情報が含まれる可能性がある．Overlap Size は chunk 分割の際に，隣り合う chunk 間で重複させる情報の量を指し，これにより情報の文脈や連続性を維持し検

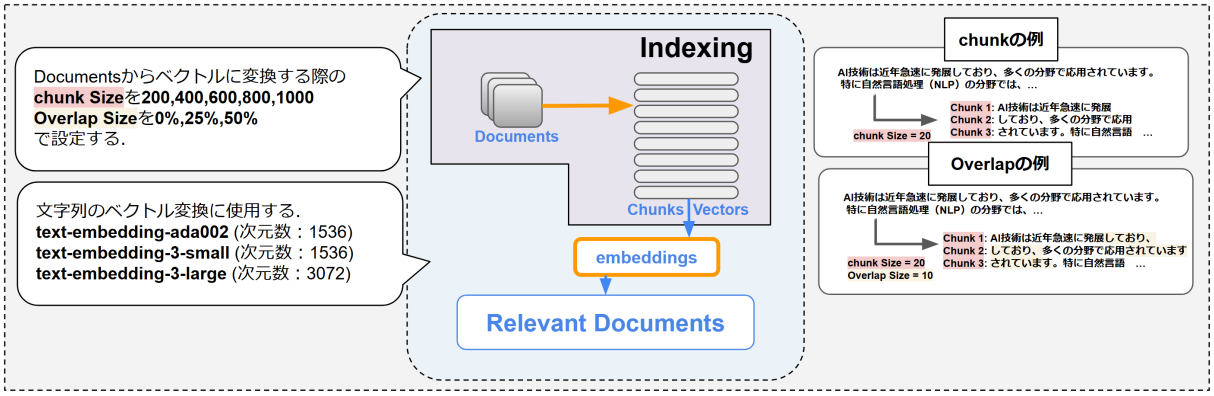

図 4.2: モデル・パラメータのメージ．chunk Size は，文書やデータをどのような単位に分割するかを決めるパラメータの一つ．Overlap Size は，文書をチャンクに分割する際に隣り合うチャンク間で重複させる情報の量．情報の文脈や連続性を維持しながら検索生成を行う場合に重要．

索・生成を行うことができる．Overlap Size が小さすぎる場合，文脈が途切れ情報が断片となるが，大きすぎる場合データ量が増え冗長性が生じる．

embedding Model には，text-embedding-ada002，text-embedding-3-large，text-embedding-3-small を用いる．M-LLM を用いて前処理を行った表のテキストデータを，各 embedding Model に通してベクトルデータベースを作成する．作成したデータベースに対して，議会会議録のテキストを入力として，ベクトル検索を行い，類似度の高い上位 3 件の表を LLM に渡し回答を生成する．この際 chunk Size は 200 から 1000 までを 200 区切りで求め，Overlap Size は chunk Size に対して 0%，25%，50%で求める．回答を生成する LLM には gpt-4o-2024-05-13 を用いる．

4.3 評価方法

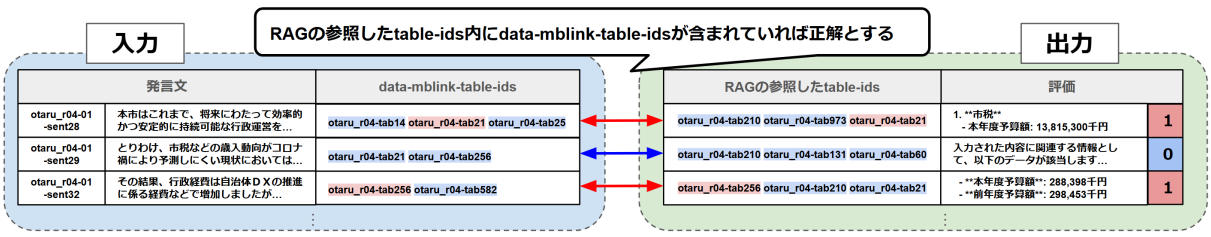

図 4.3: 正解率算出の例

図 4.3 は評価の例である．どちらの実験においても，正解率を用いる．文書検索時に参照した上位 3 件の表 ID の中に，正解となる表 ID が正解とする．参照した表 ID と正解の表 ID に一致しているものがある場合は正解，ない場合は不正解として扱われる．これらの作業を発言文全 46 個に対して行い，全発言文に対して正解の割合を求める．評価において正解率を用いる理由は，MBLink において発言文に紐づけられた正解の表が複数存在する場合があり，RAG で参照する表の件数がどの発言文においても固定となるためである．また同様の理由から，適合率は算出しない．

第5章　結果

5.1　提案手法の実験結果

表 5.1 は従来手法と提案手法の正解率を比較した表である．提案手法ではどの chunk Size においても 55% 近い正解率を算出したのに対して，従来手法ではどの chunk Size においても正解率は 41.35% であった．また chunk Size が 600，800 の場合において正解率が 56.52% と最大で 15% 以上の性能改善が見られた．

表 5.1: 提案手法の検証実験

chunk Size	提案手法	従来手法
200	54.35%	41.30%
400	52.17%	41.30%
600	**56.52%**	41.30%
800	**56.52%**	41.30%
1000	54.35%	41.30%

5.2　モデル・パラメータ検証の実験結果

各 chunk Size，Overlap Size ごとの text-embedding-ada002，text-embedding-3-small，text-embedding-3-large の結果を表 5.2 に示す．正解率が最も高かったのは，embedding Model が text-embedding-3-small で Overlap Size が 25%，chunk Size が 200 の時で 63.04% であった．

表 5.2: 各モデル・パラメータ間の比較

model	**text-embedding-ada-002**			**text-embedding-3-large**			**text-embedding-3-small**		
chunkSize	**overlap Size0%**	**overlap Size25%**	**overlap Size50%**	**overlap Size0%**	**overlap Size25%**	**overlap Size50%**	**overlap Size0%**	**overlap Size25%**	**overlap Size50%**
200	54.35%	**56.52%**	47.83%	45.63%	45.65%	41.30%	58.70%	**63.04%**	56.52%
400	52.17%	50.00%	50.00%	47.83%	43.48%	**52.17%**	52.17%	56.52%	58.70%
600	**56.52%**	43.38%	47.83%	41.30%	36.29%	32.61%	56.52%	50.00%	50.00%
800	**56.52%**	50.00%	47.83%	45.65%	43.48%	34.78%	54.35%	50.00%	50.00%
1000	54.35%	50.00%	50.00%	45.65%	47.83%	39.13%	52.17%	54.35%	56.52%

第6章　考察

6.1　提案手法の実験結果に対する考察

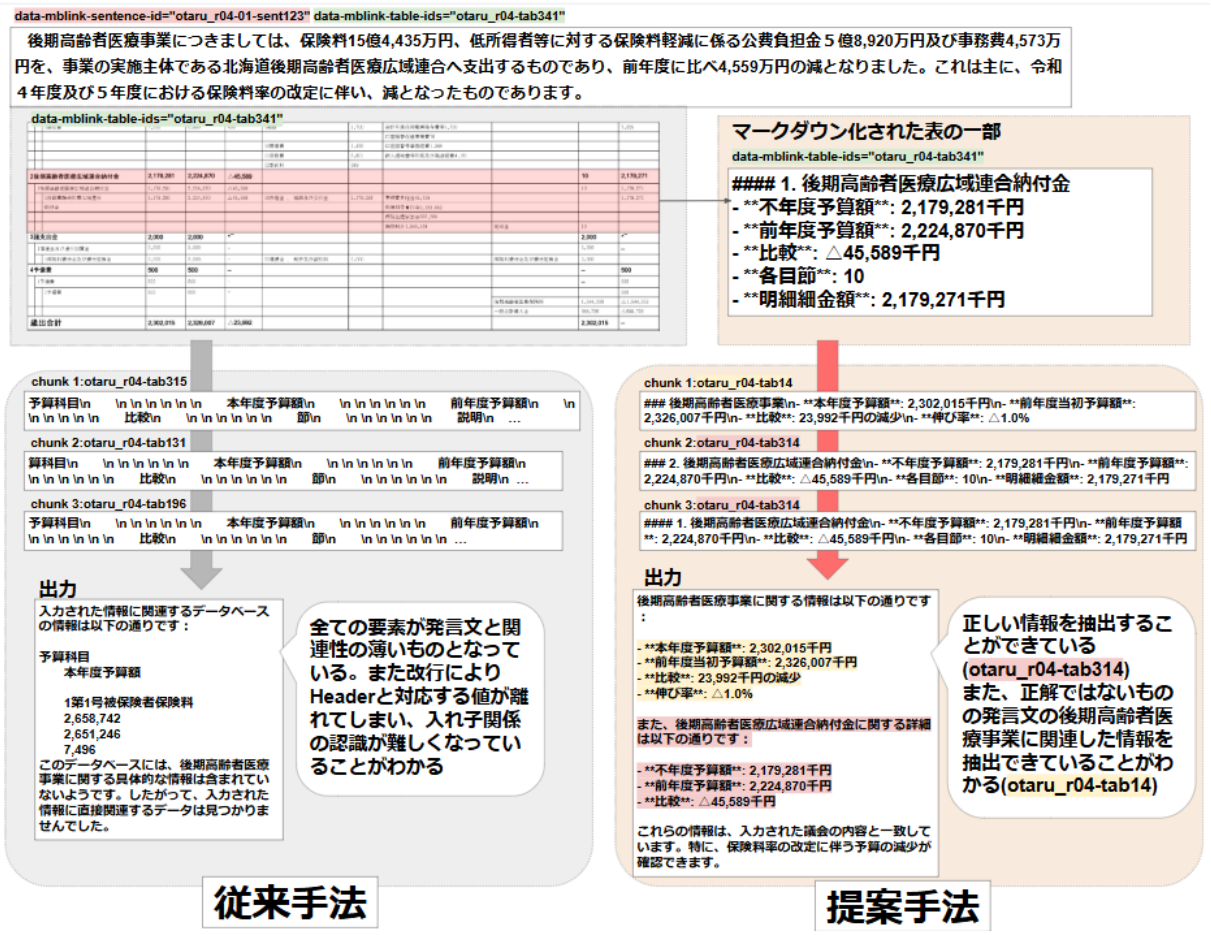

図 6.1:「提案手法のみ全て正解」を記録した事例

表 6.1 は，提案手法と従来手法における発言文ごとの正解・不正解の分類である．提案手法と従来手法の中でどの chunk Size においても，正解を検索出来ていた場合には「すべてのケースで正解」のラベルを，正解を検索できなかった場合には「すべてのケースで不正解」のラベルを付けた．提案手法と従来手法において，それぞれの手法でのみ正解が存在しどの chunk Size で正解である場合は「提案・従来手法のみすべてで正解」のラベルを付与した．ラベル付けにおいては重複を含めずに行った．

提案手法のみに正解があったケースは 13 件あり，従来手法のみに正解があったケースは 1 件あることがわかる．そのことから，ケースごとに見ても提案手法が従来手法と比較して優れていることがわかる．

表 6.1: 提案手法と従来手法における正解の分類

分類	合計
全てのケースで正解	12
全てのケースで不正解	14
提案手法のみ全て正解	5
従来手法のみ全て正解	1
提案手法のみ一部正解	8
従来手法のみ一部正解	0
どちらのケースでも正解	6

図 6.1 は「提案手法のみ全て正解」を記録した事例である.『後期高齢者医療事業につきましては、保険料 15 億 4,435 万円、低所得者等に対する保険料軽減に係る公費負担金 5 億 8,920 万円及び事務費 4,573 万円を、事業の実施主体である北海道後期高齢者医療広域連合へ支出するものであり、前年度に比べ 4,559 万円の減となりました。これは主に、令和 4 年度及び 5 年度における保険料率の改定に伴い、減となったものであります。』という発言文に対して，データベースから類似する chunk を 3 つ抽出する. 従来手法における上位 3 件の chunk をそれぞれ見ると，どれも予算科目から始まり，多数の改行が含まれることがわかる. 生成された出力に関しても，発言文の中で言及されている後期高齢者医療関連の情報は含まれていなかった.

一方で，提案手法における上位 3 件の chunk をそれぞれ見ると，chunk1 は後期高齢者医療事業に関する予算額のデータ，chunk2，3 は後期高齢者医療広域連合給付金に関する予算額のデータが含まれている.chunk1 は正解の表から抽出したものではないものの，発言文との関連がかなり高い.

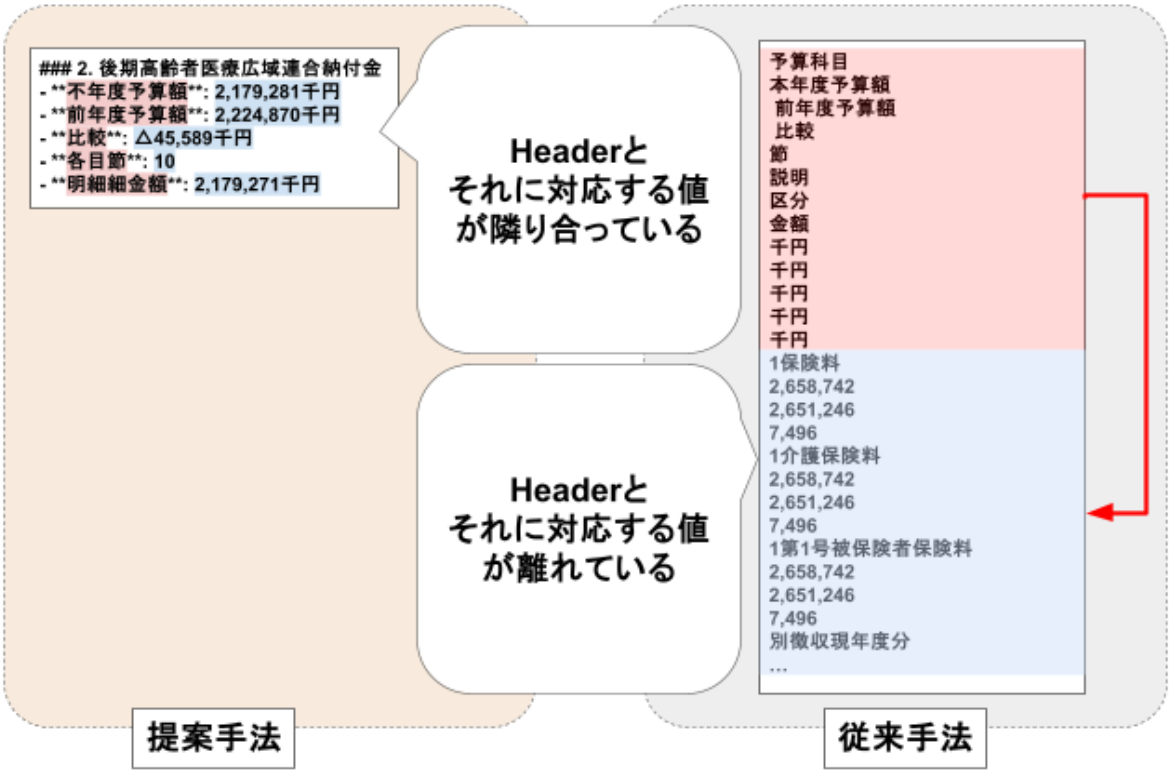

図 6.2: chunk における Header の問題

図 6.2 は，従来手法と提案手法の chunk の比較である．従来手法と提案手法における正解率の差は，例に示すような chunk における Header の問題に起因すると考えられる．提案手法の chunk を見ると「**前年度予算額**: 2,224,870 千円」のように，前年度予算額という Header に対して，2,224,870 千円という値が与えられていることがわかる．一方，従来手法の chunk では，前年度予算額という Header に対して，2,658,742 千円という値が与えられていると認識するには，chunk の先頭にある Header とそれ以下の値を対応させる必要がある．これは Header からより離れた値ほど対応付けの難易度が大きいと考えられ，検索・生成どちらにも影響を及ぼしている．

6.2 モデル・パラメータ検証の実験結果に対する考察

6.2.1 embedding Model について

表 6.2 は，embedding Model によるケースごとの正解の分類である．chunk Size が 200，Overlap Size が 0% 時における，text-embedding-ada002，text-embedding-3-small，text-embedding-3-large ごとの正解をまとめた．それぞれのケースにおいて，各 embedding Model のみが正解であった場合と，不正解であった場合を計測した．

全てのケースで正解が 13 件，不正解が 10 件であった．text-embedding-ada002 のみ正解が 3 件，不正解が 2 件，text-embedding-3-small のみ正解が 4 件，不正解が 2 件，text-embedding-3-large のみ正解が 4 件，不正解が 8 件であった．これらから text-embedding-3-small が 3 つの embedding Model の中で最も適していると考えられる．

表 6.2: embedding Model における正解の分類

分類	合計
全てのケースで正解	13
全てのケースで不正解	10
ada-002 のみ正解	3
ada-002 のみ不正解	2
3-small のみ正解	4
3-small のみ不正解	2
3-large のみ正解	4
3-large のみ不正解	8

6.2.2 chunk Size について

図**??**は text-embedding-3-small における overlap Size が 0%であった場合の chunk Size ごとの正解・不正解をまとめた表である．この図から chunk Size ごとに大きく 3 つにわけることができる．赤枠で囲われた chunk Size が 200 の場合，オレンジ枠で囲われた chunk Size が 400 から 800 までの場合，緑枠で囲われた chunk Size が 1000 の場合である．それぞれ各質問に対して，それぞれの分類ごとで，正解・不正解が決まるケースが見られた．

表 6.3 は chunk Size ごとに，唯一正解のケースと唯一不正解のケースをまとめたものである．chunk Size が 200 の場合，唯一正解が 3 件，唯一不正解が 1 件であった．また chunk Size が 200・400・600 はそれぞれ存在せず，chunk Size が 1000 の場合は唯一正解が 4 件，唯一不正解が 2 件であることがわかった．

chunk Size	200	400	600	800	1000
otaru_r04-01-sent28	1	1	1	1	0
otaru_r04-01-sent29	1	0	0	0	0
otaru_r04-01-sent30	0	0	0	0	1
otaru_r04-01-sent31	0	1	1	1	0
otaru_r04-01-sent32	1	0	0	0	1
otaru_r04-01-sent33	0	1	1	1	1
otaru_r04-01-sent38	0	0	0	0	0
otaru_r04-01-sent40	0	0	0	0	0
otaru_r04-01-sent46	1	1	0	0	0
otaru_r04-01-sent56	1	0	1	0	0
otaru_r04-01-sent59	0	0	0	0	1
otaru_r04-01-sent61	1	1	1	1	1
otaru_r04-01-sent64	1	1	1	1	1
otaru_r04-01-sent65	1	1	1	1	1
otaru_r04-01-sent67	1	1	1	1	1
otaru_r04-01-sent68	1	1	1	1	1
otaru_r04-01-sent71	0	0	0	0	0
otaru_r04-01-sent72	0	0	1	1	0
otaru_r04-01-sent77	1	1	1	1	1
otaru_r04-01-sent78	1	0	0	0	0
otaru_r04-01-sent81	1	1	1	1	1
otaru_r04-01-sent83	1	1	1	1	1
otaru_r04-01-sent88	1	1	0	0	0
otaru_r04-01-sent89	0	0	0	0	0
otaru_r04-01-sent91	1	1	1	1	1
otaru_r04-01-sent94	0	0	0	0	0
otaru_r04-01-sent108	1	1	1	1	1
otaru_r04-01-sent109	1	1	1	1	1
otaru_r04-01-sent110	1	1	1	1	1
otaru_r04-01-sent111	0	1	1	1	1
otaru_r04-01-sent112	0	0	0	0	1
otaru_r04-01-sent113	1	0	0	0	0
otaru_r04-01-sent114	1	1	1	1	1
otaru_r04-01-sent115	1	1	1	0	0
otaru_r04-01-sent116	1	1	1	1	0
otaru_r04-01-sent118	0	0	0	0	0
otaru_r04-01-sent119	0	0	0	0	1
otaru_r04-01-sent120	0	0	1	1	1
otaru_r04-01-sent121	1	1	1	1	0
otaru_r04-01-sent122	0	0	0	0	0
otaru_r04-01-sent123	1	1	1	1	1
otaru_r04-01-sent124	0	0	1	1	0
otaru_r04-01-sent128	0	0	0	0	0
otaru_r04-01-sent129	0	0	0	0	0
otaru_r04-01-sent132	1	1	1	1	1
otaru_r04-01-sent133	1	0	0	1	1

図 6.3: chunk Size における分類

表 6.3: chunk Size ごとの正解の分類

chunk Size	唯一正解	唯一不正解
200	3	1
400	0	0
600	0	0
800	0	0
1000	4	2

6.2.3 Overlap Size について

表 6.4 は text-embedding-3-small における Overlap Size ごとの正解・不正解をまとめた表である．Overlap Size は chunk Szie に依存するように設定しているため，chunk Size ごとの正解・不正解を計測した．各 chunk Size において,「すべてのケースにおいて正解・不正解」「Overlap Size が 0%，25%，50%の場合のみ正解・不正解」の計 8 つのラベルに分類した．すべてのケースにおいて正解・不正解にラベル付けられた件数が，全体の約 88%であることがわかった．このことから，Overlap Size が検索に与える影響は，embedding Model や chunk Size と比較すると小さいことが推測できる．

表 6.4: Overlap Size ごとの正解の分類

chunk Size	**200**		**400**		**600**		**800**		**1000**	
	正解	不正解	正解	不正解	正解	不正解	正解	不正解	正解	不正解
すべて	24	15	23	18	21	18	22	20	23	19
0%のみ	0	0	0	2	3	0	2	0	0	1
25%のみ	3	1	1	1	1	1	0	0	1	1
50%のみ	1	2	1	0	1	1	1	1	1	0

第7章　おわりに

本研究の貢献は，以下の 3 点である．

- M-LLM を用いた表構造理解の手法を提案した．
- 従来手法と提案手法の比較において，正解率が最大で 15%近く向上することを明らかにした．
- embedding Model では text-embedding-3-small が，chunk Size では 200 が適切で，Overlap Size は性能に大きな影響を与えないことを示した．

本研究では，RAG に使用するデータベース内に含まれる表の html に対して，画像をもとに M-LLM を使用してマークダウン形式に変換することによる，精度向上を目的としたシステムの構築・検証を行った．検証においては，議会会議録内の発言文と予算表の結び付けを行うタスクである MBLink から，令和 4 年度の小樽市の議会会議録と予算表のそれぞれ 46 件を対象としたデータセットを用いた．また，HTML ファイルから Selenium を用いることで，表の画像のスクリーンショットを撮影し，表をマークダウン形式に変換することによる提案手法の RAG を構築した．

従来手法と提案手法における正解率の比較を行ったところ，従来手法の正解率が 41.30%であったのに対して，提案手法ではどの chunk Size においても 55%付近の正解率をとり，最大で 15%程度の精度向上を確認した．また，モデル・パラメータの検証では，embedding Model では text-embedding-ada002，text-embedding-3-small，text-embedding-3-large の 3 つのモデルと，chunk Size が 200，400，600，800，1000 の場合，Overlap Size が各 chunk Size に対して 0%，25%，50%の場合をそれぞれ検証した．その結果，正解率が最も高かったのは，embedding Model が text-embedding-3-small で Overlap Size が 25%，chunk Size が 200 の時で 63.04% であった．

提案手法と従来手法における正解率の差は，検索時にベクトル検索を通じて類似度の上位とされる chunk が影響していると考えられる．提案手法では，表に含まれる各値とそれに対応する Header の値が近い距離で記述されるのに対して，従来手法では Header は最初に記述されるのみで，対応する値との距離が遠いことがわかった．これにより従来手法では，chunk 内に Header の要素が含まれないことが多くなり，検索・生成に悪影響を与えていると推測できる．

embedding Model に関しては，各モデルの正解･不正解を分類した．その結果，text-embedding-3-small が唯一正解であったケースが最も多く，唯一不正解であったケースが最も少ないことがわかった．このことから，text-embedding-3-small が適切な embedding Model であると考える．chunk Size においては，chunk Size が 200 の場合，chunk Size が 400・600・800 の場合，chunk Size が 1000 の場合，それぞれにおいて正解・不正解において異なる傾向が見られた．正解率が最も高かったのは chunk Size が 200 の場合であったが，他の chunk Szie に大きな差をとるわけでないため，使用するデータによって柔軟に対応する必要がある．Overlap Size においては，各 chunk Size における正解・不正解をまとめた．その結果，すべて Overlap Size において正解・不正解にラベル付けられた件数が，全体の約 88%であることがわかった．これにより，Overlap Size が検索・生成に与える影響はごく少数であると考える．

今後は，実際にユーザーが議会会議録に対して RAG を用いることを目的として，より大きなデータセットに対して提案手法を実装する予定である．

参考文献

[1] Patrick S. H. 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. CoRR, Vol. abs/2005.11401, , 2020.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[3] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning, 2023.

[4] 阿部晃弥, 新納浩幸. Rag における小説データベースの chunk size と overlap size と embedding モデルの効果. 言語処理学会 第 30 回年次大会 発表論文集, Vol. 2024, pp. 3215–3220, 2024.

[5] 高丸圭一. 地方議会会議録コーパスと地方議会会議録を用いた学術研究の現状. 知能と情報（日本知能情報ファジィ学会誌）, Vol. 31, pp. 25–33, 2019.

[6] 前多陸玖, 奥山和樹, 佐藤栄作, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表のセル分類に向けて. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3M5OS12b03–3M5OS12b03, 2024.

[7] Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. Retrieving complex tables with multi-granular graph representation learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, p. 1472–1482. ACM, July 2021.

[8] 木村泰知, 梶縁, 乙武北斗, 門脇一真, 佐々木稔, 小林暁雄. 議会会議録と予算表を紐づける minutes-to-budget linking タスクの提案. 言語処理学会 第 29 回年次大会 発表論文集, Vol. 2023, pp. 2427–2431, 2023.

[9] 小川泰弘, 木村泰知, 渋木英潔, 乙武北斗, 内田ゆず, 高丸圭一, 門脇一真, 秋葉友良, 佐々木稔, 小林暁雄. Ntcir-17 qa lab-poliinfo-4 のタスク設計. 言語処理学会 第 29 回年次大会 発表論文集, Vol. 2023, pp. 611–615, 2023.

付録

embedding model	text-embedding-ada002									text-embedding-3-small									text-embedding-3-large								
chunk Size	200			400			600			800			1000			200			400			600			800			1000		
Overlap Size	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%	0%	25%	50%
otaru_04-01-sent28	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent29	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent32	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent33	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent40	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent58	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent61	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent64	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent65	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent67	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent68	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent72	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent77	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent78	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent81	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent83	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent91	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent108	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent109	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent110	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent111	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent112	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent113	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent116	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent118	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent119	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent120	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent123	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
otaru_04-01-sent128	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent129	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent132	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
otaru_04-01-sent133	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1