統計データを用いた事実確認支援のための統計データ内の関連箇所抽出

宮﨑　隆豪†　　宮森　　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2254711,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，ネット上の言説に対する統計データを用いた事実確認支援のための関連箇所抽出の問題に取り組む．省庁などが提供している統計データは，一定の信頼性が担保された情報と捉えることができるため，それらを用いた事実確認はネット上の言説の真偽判断支援に有用と考えられる．これを実現するには，関連する統計データの検索，統計データ内の関連箇所抽出，ネット上の言説と統計データ内関連箇所との整合性検証の 3 段階の処理が必要となるが，本稿では，この第 2 段階の問題を扱う．統計データ内の関連箇所抽出を実現するには，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットが必要となるが，従来研究ではそのようなデータセットは提供されていない．そこで本稿では，この課題を解決するため，大規模言語モデルを活用したデータセット構築手法を提案する．具体的には，統計データとその内部の関連箇所を入力とし，それに整合する，あるいは，整合しない言説を LLM に生成させることで，大規模かつ質の高いデータセット構築を試みた．実験では，構築したデータセットの妥当性を検証するとともに，このデータセットを活用して関連箇所抽出モデルの性能を評価する．

キーワード　事実確認支援，統計データ，関連箇所抽出，ソーシャルメディア，説明文生成

1　はじめに

近年，SNS やオンラインプラットフォームは日常における主要な情報源として定着しており，日常的に多種多様な情報がやり取りされている．これらの情報は社会や個人の意思決定に大きな影響を与える一方で，その信頼性が常に保証されているわけではない．特に，誤った情報や悪意のある情報が拡散されることで，社会的混乱を招くリスクが高まっている．例えば，パンデミック時には，誤った医療情報が広がり，正しい予防策の普及や治療への適切な対応が遅れる事例が報告されている．また，政治的な偽情報が選挙結果や政策決定に影響を及ぼす可能性も指摘されている．このような背景から，情報の真偽を正確に判断するための「事実確認支援（fact-checking support）」が社会的に重要な課題となっている．

こうした状況において，信頼性の高い情報源として，政府や公共機関が提供する統計データの重要性が増している．統計データは客観的な数値に基づく情報を提供し，主観や憶測が含まれやすいネット上の言説に対し，客観的かつ確定的な証拠能力を持つ．したがって，統計データを用いた事実に即した検証は，言説の真偽を客観的に判断するための強固な基盤となる．

統計データを活用した事実確認支援を実現するためには，以下の 3 段階の処理が必要と考えられる．まず，事実確認対象の言説に関連した統計データ検索 [1]，次に，検索された統計データ内から言説の事実確認に利用できる関連箇所の特定，最後に，特定した関連箇所と言説との整合性の検証である．例えば，「日本の人口は増加している」という言説に対して，適切な統計データを用いてその真偽を判断するには，統計データ内の総人口に関するセクションを特定し，近年の人口推移を確認した上で言説内容との整合性を検証する必要がある．

本稿では，事実確認支援の第 2 段階である「統計データ内の関連箇所抽出」に焦点を当てる．この処理を効果的に実現するためには，統計データの内容を適切に解釈し，言説に対応する箇所を的確に特定する必要がある．しかし，従来研究では，このようなタスクに必要なデータセットが整備されておらず，適切な関連箇所抽出の実現に向けた基盤が十分ではなかった．また，統計データは数百から数千行に及ぶことも珍しくなく，それらをすべて大規模言語モデルの入力ウィンドウに収めることは，技術的な制限や情報密度の過剰な上昇を招く．これにより，LLM が表の構造を正しく把握できず，存在しない数値を参照したり，行を読み間違えたりするハルシネーションを引き起こし，抽出精度が著しく低下するという課題も存在する [2].

本稿では，これらの課題に対応するため，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを構築し，その有用性を示すことを目的とする．本稿で対象とする統計データには，公的機関が提供する信頼性の高い CSV 形式のデータを用いる．まず前処理として，複雑な構造を持つ CSV データを，解析が容易な 2 次元のデータフレーム形式へと変換する処理を行う．これにより，表形式データ特有の行列構造を保持したまま，計算機での効率的な処理を可能とする．その後，統計データの背景情報を示すメタデータと統計データ内の特定の一行である関連箇所を LLM への入力とし，その関連箇所に基づいた言説を生成することで，統計データ，言説，関連箇所の対応関係が明確なデータセットを構築する．この際，ソーシャルメディアなどで見られる多様な言説の真偽の度合いを反映させるため，生成される言説には真偽の度合いが異なる 4 種類のラベルを付与する．

本稿では，構築したデータセットを用い，LLM による関連箇所抽出の性能を評価するための実験を実施する．前述の通り，

統計データは膨大な数値情報を含むため，一度の推論ですべての情報を精査することは困難であるという課題がある．そこで本稿では，データを 10 行単位のバッチに分割して段階的に絞り込みを行う「2 段階推論システム」と，独自の「関連度スコア」を用いた手法を提案し，抽出精度と処理効率の両面からその有効性を検証する．本稿の主な貢献は以下の通りである．

(1) 言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを構築し，統計データを活用した事実確認支援の基盤を提供する．

(2) 大規模な統計データを効率的に処理するバッチ分割型の 2 段階推論システムと，独自スコアリングによる高精度な関連箇所抽出手法を提案し，統計データ内の関連箇所抽出を実現した．

(3) 構築したデータセットと提案手法を用いた実験を通じて，統計データを用いた事実確認支援における LLM の適用可能性と，本データセットの妥当性を示した．

本稿の構成は以下の通りである．第 2 節では，事実確認支援および表形式データの理解に関する関連研究について述べる．第 3 節では，統計データ内の関連箇所抽出における問題設定を定義し，本稿で提案する 2 段階推論システムと関連度スコアリング手法の詳細を説明する．第 4 節では，提案手法の検証に用いるためのデータセット構築手順とその諸元について述べる．第 5 節では，構築したデータセットを用いた実験結果を示し，提案手法の有効性および今後の課題について考察する．最後に，第 6 節において本稿のまとめと今後の課題について述べる．

2 関 連 研 究

2.1 関連箇所抽出

機械読解（Machine Reading Comprehension; MRC）は，与えられたテキストから質問に対する回答を抽出するタスクであり，特にスパン抽出（Span Extraction）は，回答を文中の一部分として特定する手法として広く用いられている．代表的なデータセットとして SQuAD [3] があり，文中の開始位置と終了位置を特定する形式での応答が求められる．MRC におけるスパン抽出手法は，BERT [4] や SpanBERT [5] など事前学習モデルを用いることで性能の向上が進んでいる．

しかし，従来の抽出型機械読解は，回答が一つの範囲に限定される単一回答 QA が中心であった．これに対し，DROP [6] などのデータセットでは，複数の範囲における回答抽出が必要な複数回答 QA が新たに追加されており，複数の回答がコンテキストに散在する場合に対応するための研究も進められている．本稿は，言説の根拠となる統計データ内の「関連箇所」を特定するものであり，テキストベースの MRC で培われたスパン抽出技術を，表形式データに適用するための基礎研究として位置づけられる．

2.2 表形式データの学習

表や統計データを活用した自然言語処理 (NLP) の分野では，テキスト情報と表データを統合的に扱う手法が注目されている．

TaBERT [7] は，表データの構造を考慮し，BERT を基盤としたエンコーダを用いて，自然言語テキストと表の情報を統合的に学習するモデルである．これにより，表の構造やセル情報を考慮しながら，質問応答 (QA) や検索タスクに応用可能な表現を学習できる．

また, TAPAS [8] も, 表形式データ上の推論に特化した BERT ベースのモデルとして提案されている．TAPAS は，表のキャプションや記事のタイトルなどを質問文の代わりとして用い，Wikipedia の大規模な表形式データで事前学習を行うことで，表のセルを選択したり，集計操作を行ったりするタスクを解くことが可能である．これらの研究は，表形式データの構造を理解し，その内容を自然言語で扱うための基盤技術である．しかし，TaBERT や TAPAS といった既存の表形式データ専用エンコーダを用いた手法は，モデルの入力制限により，本稿が対象とするような数百行を超える大規模な統計データ全体を一度に処理することが困難である．これに対し，本稿では特定の表形式データ用エンコーダは採用せず，汎用的な大規模言語モデルを推論エンジンとして用いる．これにより，専用モデルの追加学習コストを抑えつつ，第 3 節で提案する「2 段階推論システム」を通じて，大規模な表構造の中から必要な情報を柔軟かつピンポイントに特定することが可能となる．

2.3 事実確認支援

事実確認支援 (fact-checking support) は，オンライン上の誤情報や悪意のある情報の拡散を防ぐための重要な研究領域である．特に，ソーシャルメディアやニュースサイトにおける情報の信頼性を検証するため，多くの手法が提案されている．代表的なデータセットとして FEVER [9] があり，Wikipedia を基に 18 万件以上の主張と，それに対応する証拠文のペアを提供する．FEVER では，事実確認を「主張が支持されるか，反証されるか，情報不足か」を判定するタスクとして定義し，自然言語推論 (NLI) や検索技術と組み合わせた多くの手法が提案されている．例えば，文書検索と文選択を組み合わせたパイプライン手法 [10] や，グラフ構造を用いて複数の証拠文を統合し推論を行う手法 [11] などが提案され，高い精度を達成している．

また，近年では Transformer ベースのモデルを活用した誤情報検出技術も発展している．DisinfoBERT [12] は，ソーシャルメディア上の誤情報を識別するために設計された BERT ベースのモデルであり，文のコンテキストや言語的特徴を考慮して誤情報を分類している．このような技術は，ニュース記事や SNS の投稿の信頼性を判断する上で有用であり，誤情報対策として実用化が進められている．

本稿は，これらの事実確認の枠組みと共通する課題を扱うが，対象とするデータが異なる点に特徴がある．既存の研究は主にニュース記事や Wikipedia を対象とし，テキストベースのファクトチェックを行うのに対し，本研究では公的機関が提供する信頼性の高い統計データを直接的な根拠とする言説の収集・整理を可能にする．ネット上の書き込みと統計データの関連箇所を特定することで．主張と数値データの整合性を評価し，ファクトチェックを補助する仕組みを構築することを目的としている．

3 提案手法

本節では，第 1 節で述べた統計データ内の関連箇所抽出という課題に対し，本研究が提案する 2 段階推論システムおよび関連度スコアリング手法の詳細について述べる．

3.1 問題設定

本稿で取り組む統計データ内の関連箇所抽出タスクは，自然言語で記述された言説と対象となる統計データを入力とし，その言説の事実確認を行う上で参照すべき統計データ内の関連箇所を特定することを目的とする．

本タスクの入出力を次のように定義する．言説の集合を C，統計データの集合を D とする．言説 $c \in C$ に対し，対象とする統計データ $d \in D$ は，背景情報を示すメタデータ m，表の構造を定義するヘッダー h，および表内の n 行の行集合 $R = \{r_1, r_2, \ldots, r_n\}$ の組として次のように表される．

$$d = (m, h, R) \tag{1}$$

ある言説 c と統計データ d に対して，各行 $r_i \in R$ が言説の根拠としてどの程度関連しているかを示す関連度スコアを s_{r_i} とする．本タスクの目的は，c, m, h, r_i を入力として s_{r_i} を算出する関数 f を用い，スコア s_{r_i} が最大となる行 r^* を関連箇所として特定することである．

$$s_{r_i} = f(c, m, h, r_i) \tag{2}$$

$$r^* = \arg\max_{r_i \in R} s_{r_i} \tag{3}$$

本稿では，LLM に対して「言説 c」「メタデータ m」「ヘッダー h」「行 r_i」の要素を提供し，これらを統合的に解釈させることで関数 f を実現し，関連箇所の特定を行う．なお，本稿では計算機による一貫した管理と言説生成の確実性を担保するため，抽出の最小単位を統計データ内の特定の一行として定義する．詳細なデータセットの構成および各ラベルの定義については第 4 節で述べる．

3.2 2 段階推論システムによる抽出プロセス

大規模な統計データから効率的かつ高精度に関連箇所を特定するため，本稿では図 1 に示すような「2 段階推論システム」と「関連度スコアリング」を用いた手法を提案する．具体的な抽出プロセスを以下に定義する．

3.2.1 バッチ分割

まず，入力となる統計データの行集合 R を，k 行（本稿では $k = 10$）ずつの連続する部分集合であるデータバッチ $B = \{b_1, b_2, \ldots, b_m\}$ に分割する．ここで，$m = \lceil |R|/k \rceil$ である．各データバッチ b_j は，LLM が表構造を識別しやすいよう Markdown 形式に変換され，各行には明示的な行番号が付与される．

3.2.2 バッチフィルタリング

第一段階では，言説 c，メタデータ m，ヘッダー h，および各データバッチ b_j を LLM に入力し，当該バッチと言説の関連度を示すバッチスコア s_{b_j} を算出する．

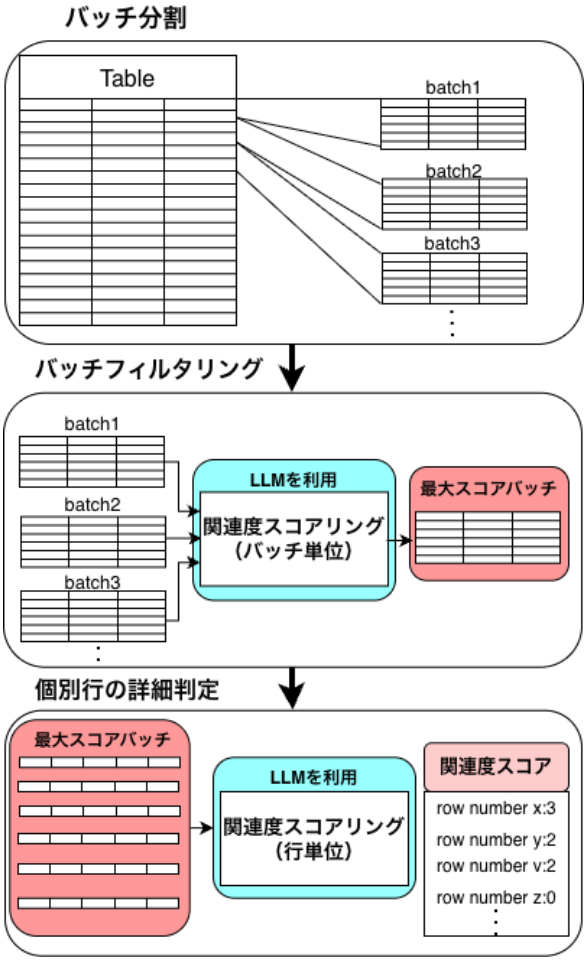

図 1　関連箇所抽出の概要図

$$s_{b_j} = f_{stage1}(c, m, h, b_j) \quad (s_{b_j} \in \{0, 1, 2, 3\}) \tag{4}$$

次に，スコア集合 $\{s_{b_1}, \ldots, s_{b_m}\}$ の中から最大スコア（閾値 $s \geq 1$ ）を獲得したバッチ b_{target} を詳細解析の対象として選別する．

$$b_{target} = \arg\max_{b_j \in B} s_{b_j} \tag{5}$$

全てのバッチのスコアが 0 であった場合は，該当箇所なしと判定し処理を終了する．

3.2.3 個別行の詳細判定

第二段階では，b_{target} に含まれる各行 $r_i \in b_{target}$ に対してより緻密な解析を行う．ここでは，項目名と数値の対応を強調するため，各行を「フィールド名：値」のテキスト形式に整形して提示する．LLM はこれに基づき，各行の関連度スコア s_{r_i} を算出する．

$$s_{r_i} = f_{stage2}(c, m, h, r_i) \quad (s_{r_i} \in \{0, 1, 2, 3\}) \tag{6}$$

最終的に，あらかじめ設定した閾値（$s \geq 2$ ）を満たし，かつ最大スコアを持つ行 r^* を言説の根拠となる関連箇所として特定する．

$$r^* = \arg\max_{r_i \in b_{target}} s_{r_i} \tag{7}$$

表 1　関連度スコアの定義と判断基準

スコア	定義
3	データ行の情報のみで，言説の主要な要素の真偽が完全に確定する．
2	データ行の情報が，言説の主要な要素の真偽を部分的に判断できる証拠を提供する（一部の要素は不明だが，重要な手がかりとなる）．
1	データ行の情報が，言説の背景，文脈，または間接的なヒントとして関連するが，真偽判定には不十分である．
0	データ行の情報と，言説の主要な要素の間に意味のある接点がない．

3.3　関連度スコアリング

本手法における各段階の推論では，LLM に対して言説とデータの関連性を数値化させる「関連度スコア」を導入する．スコアは 0 から 3 の 4 段階の整数値で定義され，その具体的な判断基準を表 1 に示す．評価にあたっては，主観的な解釈や推測を排除し，提供されたデータ行が言説の真偽を判定するための直接的な根拠となり得るかという観点から判断を行うよう，プロンプトにて指示する．

各段階におけるスコアの活用方法は以下の通りである．第一段階のバッチフィルタリングにおいては，閾値を 1 と設定し，スコア 1 以上のバッチの中から最大スコアを獲得したバッチを次段階の解析対象として選出する．第二段階の詳細推論においては，閾値を 2 と設定し，スコア 2 以上の行を抽出対象とする．このように，段階的に閾値と提示形式を調整することで，膨大なデータの中から真偽判定に寄与する核心的な情報の特定を試みる．

4　データセット構築

4.1　データセット構築の概要

本稿では，統計データを用いた事実確認支援のための基盤を構築することを目的とし，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを作成する．この 3 つ組のうち，統計データについては，NTCIR-15 [13] で提供される公的データを利用する．NTCIR-15 の統計データは，多様な分野のデータを網羅し，標準化された形式で提供されていることから，データの信頼性が高く，事実確認支援の基盤として活用するのに適している．一方，ネット上で事実確認支援の対象候補となる言説は，統計データと紐付けられておらず，データセットとして利用可能なデータが不足している．このため，本稿では，NTCIR-15 の統計データを活用し，それに関連する言説を生成するために LLM を用いた言説生成手法を採用する．

本稿では，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットの構築を問題として設定し，その具体的な要件を以下のように定める．

まず，事実確認の対象となる自然言語の「言説」に対し，その根拠となる統計データ内の「関連箇所」が明確に紐付けられている必要がある．本稿では，言説生成の確実性とデータセットの構造的一貫性を保つため，関連箇所を統計データ内の特定の一行として定義する．統計データにおける関連箇所は，本来「単一のセル」「行内の一部の列」「離れた複数行」など多様なパターンが存在するが，本稿で行単位の定義を採用した主な理由は，以下の 2 点に集約される．

まず，統計データを用いた事実確認支援の第 3 段階である整合性検証において，LLM は抽出された特定行のみならず，その周辺行をコンテキストとして参照することで，統計データの時系列的な変化やカテゴリ間の比較といったメタ情報を踏まえた高度な推論を行うことが可能である．したがって，第 2 段階において行単位の特定が実現できれば，整合性判定に必要な情報は十分に保持されると言える．

次に，実用上の柔軟性が挙げられる．バッチ分割による 2 段階推論と関連度スコアリングを組み合わせた本手法では，個別の行に対して独立にスコアリングを行うため，関連箇所が離れた複数行に及ぶ場合でも，各行に対して高い関連度を付与することで，漏れなく第 3 段階の整合性検証へと情報を引き渡すことができる．また，特定のセルに依存する言説であっても，そのセルを含む「行」全体を入力として与えることで，第 3 段階における数値の特定とその意味解釈を同時に行うことが可能となる．

以上の理由から，本稿では計算効率と推論に必要な情報のバランスを考慮し，抽出の最小単位を行として定義する．これにより，どの行の情報と言説が対応しているかを厳密に管理しつつ，後続の処理へ十分な情報量を担保することが可能となる．

次に，現実の事実確認の状況を反映させるため，生成される言説は多様な真偽の度合いを持つ必要がある．ソーシャルメディアなどで見られる言説は，必ずしも完全に正しいものや誤っているものばかりではなく，一部に事実を含むものや，提示された情報だけでは判断できないものも存在する．このような現実の状況を再現するため，本データセットでは統計データに基づいて生成される言説に対し，表 2 に示す 4 種類のラベルを定義し，付与することとする．

これらの要件を満たすデータセットの構築は，手動アノテーションの莫大なコストや専門知識の必要性といった課題を伴う．本稿は，これらの課題に対し，特に LLM を活用したデータセット構築アプローチを提案することで，統計データを用いた事実確認支援のための関連箇所抽出技術の発展に寄与することを目指す．

4.2　データセット構築の手順

本節では，統計データ，言説，および関連箇所の 3 つ組から構成されるデータセットの構築手順について述べる．本手法では，人手による大規模なアノテーションコストを削減しつつ，質の高いデータを効率的に生成するため，LLM を活用した自動構築プロセスを採用する．具体的な手順は以下の 2 段階から

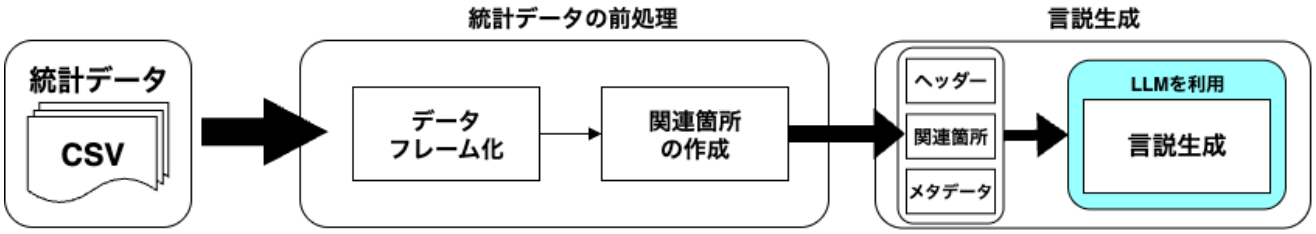

図 2　データセット構築の概要図

表 2　言説に付与するラベルの定義

ラベル	定義
True（真）	言説の内容が，対応する統計データによって完全に支持される．
False（偽）	言説の内容が，対応する統計データによって明確に反証される．
Partially True（一部真）	言説の一部は統計データによって支持されるが，他の一部は裏付けられない，あるいは矛盾する内容を含む．
Undeterminable（判別不能）	言説は統計データに関連する内容ではあるが，そのデータのみでは真偽を判断することができない．

なる（図 2）．

(1) 統計データの前処理とデータフレーム化

まず，NTCIR-15 から提供される CSV 形式の統計データを取得し，前処理を行う．複雑な構造を持つ CSV データを，LLM が論理的に解釈しやすい 2 次元のデータフレーム形式に変換する．この際，行列ヘッダの階層構造を整理し，各行が独立した意味を持つ単位として抽出可能な状態に整える．

(2) LLM による言説の自動生成とラベル付与

次に，データフレーム化された統計データからランダムに抽出した「関連箇所（特定の 1 行）」と，表の構成を定義する「ヘッダー」，およびその統計データの背景情報である「メタデータ」を LLM に入力する．ヘッダー情報を併せて入力することで，LLM は関連箇所に含まれる数値情報の意味を正確に把握し，文脈に即した言説を生成することが可能となる．LLM には，与えられた数値情報に基づき，表 1 で定義した 4 種類のラベル（True, False, Partially True, Undeterminable）のいずれかに合致する自然言語の言説を生成させる．これにより，関連箇所と言説の対応関係が厳密に担保されたデータセットを構築する．

4.3　統計データの前処理

本節では，データセット構築の基盤となる「統計データの前処理」の詳細について述べる．本工程の目的は，多様な形式で存在する公的統計データを，一貫した構造を持つデータフレーム形式へと変換することにある．

まず，元の CSV データに対して，不要な連続した数値列や特定の文字列 (改行文字，カンマなど)，および空白行の削除といった前処理を適用し，データのクリーンアップを行う．次に，クリーンアップされたデータの中から，表の数値領域の左上部分の位置を特定し，その情報を用いて行ヘッダおよび列ヘッダを抽出する．一部の統計データに見られる多階層構造のヘッダに対しても，直前までの行・列の状態を保持・比較することで適切に結合し，単一のヘッダ行として展開する処理を適用する．最後に，抽出されたヘッダ情報と数値領域を対応付け，最終的なデータフレームを構築する．前述の通り，本稿では関連箇所の最小単位を「行」として定義しており，このデータフレームの各行が言説生成の根拠として抽出される．このように構造化された形式で入力を行うことで，LLM は数値情報の意味を正確に把握し，文脈に即した言説生成が可能となる．なお，本工程で構築されたデータフレームは，第 5 節で述べる提案手法の関連箇所抽出タスクにおいて，LLM が論理構造をより解釈しやすいよう Markdown 形式へと変換した上で入力に利用される．

4.4　言 説 生 成

本節では，データセット構築の中核となる，大規模言語モデルを用いた言説生成手法について詳述する．言説生成には，高い推論性能と多岐にわたるタスクへの適応性を持つ OpenAI 社の GPT-4o [14] を利用する．

具体的な生成プロセスとして，LLM には，統計データの背景情報を提供するメタデータ，統計データのヘッダー，および，統計データ内の数値領域（ヘッダー行を除くデータ行）からランダムに抽出された「特定の 1 行（関連箇所）」を入力し，これらに基づいた主張文を生成させる．行をランダムに選択することで，統計データ内の特定の項目に依存しない，網羅的かつ客観的な言説の収集を可能にしている．

この際，関連箇所の該当行のみを入力するのではなく，その周辺前後 5 行を含めたコンテキストを提示する工夫を施している．これにより，LLM は関連箇所の数値が統計表全体の中でどのような位置付けにあるのかを正確に把握することが可能となり，表の構造を誤解することなく，より自然かつ妥当性の高い言説の生成が期待できる．

また，本稿では，4.1 節で定義した 4 種類のラベルに基づき，プロンプトを通じてそれぞれ異なる要件を含めることで，特定の性質を持つ言説を生成させる．これにより，単に事実と整合する言説を生成するだけでなく，虚偽の言説や判別不能な言説など，現実の事実確認シナリオに対応した多様な言説を体系的に作成する．

4.5　データセットの諸元

本手法を用いて構築したデータセットの具体的な統計量を表 3 に示す．本データセットは，NTCIR-15 から選定した 200 件の

統計データに基づき構築された．各統計データに対して 4 種類の真偽ラベル（True, False, Partially True, Undeterminable）の言説をそれぞれ 3 件ずつ，合計 12 件生成しており，データセット全体では 2,400 件の言説で構成される．

対象とした統計データの規模は，最小 2 行から最大 7,479 行と極めて幅広く，平均行数は 1,142.4 行に達する．表 3 の行数分布に示す通り，500 行を超える大規模なデータが全体の半数以上（54.5%）を占めており，多くの統計データにおいて LLM の最大トークン制限を超えるサイズが含まれている．この結果は，本稿で提案するバッチ分割型の 2 段階推論システムの必要性を裏付けるものである．

表 3　構築したデータセットの統計量および内訳

項目	件数・数値	割合
統計データの行数分布		
1 ～ 100 行	42 件	21.0%
101 ～ 500 行	49 件	24.5%
501 ～ 1,000 行	25 件	12.5%
1,001 ～ 2,000 行	25 件	12.5%
2,001 行以上	59 件	29.5%
統計データ全体		
対象ファイル総数	200 件	100.0%
累積総データ行数	228,474 行	-
平均行数 / ファイル	1,142.4 行	-
生成言説の内訳（ラベル別）		
True / False / Partial / Undet.	各 600 件	各 25.0%
合計言説数 (N)	2,400 件	100.0%

5 実　　験

5.1 実験目的

本稿では，第 4 節で構築したデータセットの妥当性を検証するとともに，統計データ内の関連箇所抽出における提案手法の性能を評価するための実験を実施する．

第一の目的は，提案手法の構成要素である「バッチ分割による 2 段階推論」および「関連度スコアリング」の有効性を検証することである．具体的には，これらの要素の有無を組み合わせた複数の手法を比較評価するアブレーション実験を行い，従来手法である一括入力方式に対する優位性を，推論の成功率と抽出精度の両面から明らかにする．

第二の目的は，提案手法を用いた際の言説の性質やモデルの違いによる挙動の差を分析することである．ここでは提案手法に固定した上で，言説のラベル別に関連箇所抽出の精度を算出する．あわせて，バックエンドに用いる LLM として商用モデルとオープンソースモデルを比較し，ラベルごとの抽出難易度やモデル間の性能差について詳細な知見を得ることを目的とする．

5.2 実験設定

実験に用いるモデル，比較手法，データセット，および評価指標について説明する．

5.2.1 使用するモデル

推論エンジンには，高い推論能力を持つ商用モデルである OpenAI 社の GPT-4o，およびオープンソースモデルである Alibaba Cloud 社の Qwen2.5-14B-Instruct [15] を採用した．Qwen2.5-14B は，パラメータ数が比較的軽量でありながら高い日本語処理能力を保持しており，ローカル環境やプライベートクラウドでの運用を想定した実用的なモデルとしての性能を検証するために採用した．これにより，クラウド型の巨大モデルと、運用コストやデータ秘匿性に優れた軽量モデルとの間における性能差を明らかにする．

5.2.2 比較手法

本稿では，提案する「2 段階推論」および「関連度スコアリング」の有効性を明らかにするため，これらの構成要素の有無を組み合わせた 4 つの手法を設定し，比較評価を行う．

まず，バッチ分割を行わず，統計データの全行およびメタデータを単一のプロンプトに集約して入力する「一括方式」として，スコアリングを行わずに関連行の直接的な抽出のみを指示する「一括方式（スコアなし）」と，各行に対するスコアリング結果に基づき抽出を行う「一括方式（スコアあり）」を設定する．これらの一括方式においては，提案手法との公平な比較を期すため，入力データは提案手法と同様の Markdown 形式に変換し，各行に行番号を付与して提示する．なお，一括方式は統計データが長大である場合に LLM のトークン制限を超過する恐れがあるため，本実験ではこのような入力制限に対する手法の堅牢性についても評価の対象とする．

これらに対し，第 3.2 節で述べたバッチ分割および 2 段階推論を採用した手法として，各段階での判定を関連の有無のみの二値判定とする「2 段階方式（スコアなし）」と，これに各段階での関連度スコアリングを統合した「提案手法（2 段階＋スコアあり）」を比較に用いる．

5.2.3 実験データセット

本実験では，第 4 節で構築したデータセットの中から，評価用として以下の条件に基づき抽出したサブセットを利用する．

まず，構成要素の有無によるアブレーション実験においては，各ラベルから 30 件ずつ抽出した合計 120 件の言説を用いる．また，GPT-4o と Qwen2.5-14B を用いたモデル別の詳細比較実験においては，各ラベルから 50 件ずつ抽出した合計 200 件の言説を用いる．

データセットの選定にあたっては，実験結果が特定のデータ規模に依存することを防ぐため，統計データの行数分布が元のデータセット全体の平均的な分布（表 3）と整合するよう配慮した．この条件を満たす範囲内でランダムに抽出を行うことで，大規模なデータから小規模なデータまでを網羅し，かつ客観的な評価が可能なサブセットを構築した．

5.3 評価方法と評価指標

本実験では，提案手法による関連箇所抽出の性能を定量的に評価するため，各言説に対し正解となる関連箇所は常に 1 行であると定義し，以下の手順と指標を用いて評価を実施する．

5.3.1 評価の手順

提案手法については，プログラムの処理ステップに合わせ，Stage 1（バッチフィルタリング）および Stage 2（個別行の詳細判定）の 2 段階で評価を行う．まず，Stage 1 ではデータバッチの中から正解行を含むバッチを正しく選別できているかを検証する．次に，Stage 2 において，選別されたバッチ内から最終的に正解行を特定できたかを検証する．

一方，一括入力方式については，単一の推論結果において正解行が最大スコアを獲得しているかを検証する．なお，適合率，再現率，および F1 スコアの算出にあたっては，正常に推論を完了した試行のみを評価対象とし，実行に失敗した試行は集計から除外する．

5.3.2 評 価 指 標

本稿では，実験の目的に応じて以下の指標を使い分けて評価を行う．構成要素の有無によるアブレーション実験では，手法の堅牢性と基本性能を測るため，実行成功率，適合率，再現率，F1 スコア，および平均実行時間を用いる．ラベル別・モデル別の詳細分析では，提案手法の内部動作を詳細に評価するため，実行成功率に代わりバッチ特定成功率を導入し，これに適合率，再現率，F1 スコア，平均実行時間を加えて評価を行う．各指標の定義は以下の通りである．

1. 実行成功率：入力データが LLM のトークン制限などに抵触せず，正常に推論を完了した割合を示す．大規模な統計データに対する各手法の「堅牢性」を評価する指標である．

2. バッチ特定成功率（BSR; Batch identification Success Rate）：Stage 1 において，正解行を含むバッチが最大スコアを獲得し，次段階へ正しく引き継がれた割合である．提案手法の有効性を内部的に評価する指標として用いる．

4. 適合率 (Precision)・再現率 (Recall)・F1 スコア (F1-score)：モデルが「関連あり」と判定した行を対象に算出する．適合率は抽出された行のうち正解が占める割合，再現率は正解行を漏らさず抽出できた割合，F1 スコアはその調和平均である．

5. 平均実行時間：1 言説あたりの処理に要した時間の平均値であり，手法やモデル間の実用的な処理効率を比較するために用いる．

5.4 実 験 結 果

本節では，5.3 節で定義した各指標に基づき，関連箇所抽出の評価結果を述べる．まず，アブレーション実験を通じて各構成要素の有効性を検証し，次に複数の LLM を用いたラベル別の詳細な性能比較を行う．

5.4.1 アブレーション実験による構成要素の評価

表 4 に，GPT-4o を用いた提案手法および各比較手法の性能評価結果を示す．

まず，実行成功率に着目すると，一括方式（スコアなし・あり）がいずれも 70%台に留まっているのに対し，2 段階方式を採用した手法はいずれも 100.0%の成功率を記録した．大規模な統計データに対しても，バッチ分割による 2 段階推論を用いることで，トークン制限を回避し安定して推論を実行できることが確認された．実行時間については，一括方式と比較して大幅に増加する傾向にあるものの，データの規模に関わらず安定した処理が可能であるという利点がある．

次に，抽出精度（F1 スコア）を比較すると，一括方式（スコアあり）が 0.817 と高い値を示している一方で，提案手法（2 段階＋スコアあり）は 0.784 であった．ただし，一括方式の結果は実行に成功した 70%のデータのみを対象とした集計値であるのに対し，提案手法は失敗事例を含む全データ（$N = 120$）を完遂した上での数値である．また，スコアリングの有無による影響を見ると，それぞれスコアありの手法がスコアなしの手法の F1 スコアを上回る結果となった．

5.4.2 モデル別のラベル別抽出精度

各ラベル 50 件（合計 $N = 200$）のデータセットを用い，提案手法を GPT-4o および Qwen2.5-14B に適用した際の性能比較を表 5 に示す．

全体平均の性能を確認すると，バッチ特定成功率（BSR）については，GPT-4o が 0.969，Qwen2.5 が 0.943 といずれも高い数値を示した．これは，バックエンドのモデルの種類に関わらず，提案手法の第一段階（バッチフィルタリング）が，正解を含むバッチを極めて高い確率で次段階へ引き継げていることを示している．

一方で，ラベル別の抽出精度を確認すると，モデルによって精度に顕著なばらつきが確認された．GPT-4o では「True」ラベルにおいて F1 スコアが 0.860 と最も高かったのに対し，Qwen2.5 では「Partially True」が 0.640 と最も高く，「True」は 0.542 であった．特に Qwen2.5 の「True」ラベルでは，再現率（Recall）が 0.930 と高い一方で適合率（Precision）が 0.382 と低く，他のラベルと比較して適合率の低下が顕著であった．

また，「Undeterminable」ラベルにおいても，両モデルで異なる傾向が見られた．GPT-4o は再現率（0.692）と適合率（0.729）が近い値となったのに対し，Qwen2.5 は適合率が 0.810 と全項目中で最高値を示した一方で，再現率は 0.462 と低い値に留まった．

平均実行時間については，GPT-4o（374.51 秒）に対し，Qwen2.5（1738.39 秒）は約 4.6 倍の時間を要しており，商用モデルとオープンソースモデルの間で処理効率に顕著な差が見られた．

5.5 考　　察

本実験の結果に基づき，提案手法の有効性とデータセットの妥当性，および今後の課題について考察する．

第一に，提案手法の妥当性について考察する．表 4 の結果より，提案手法は実行成功率 100.0%を達成した．一括入力方式（スコアあり）の F1 スコア（0.817）は数値上は提案手法（0.784）を上回っているが，これは全データの約 3 割におよぶエラー事例を除外した，成功事例のみの平均値である点に留意する必要がある．本手法のようにデータをバッチ分割して段階的に絞り込むアプローチは，一度に処理する情報密度を最適化し，長大なデータに対しても安定した出力を可能にした．実行時間の大幅な増加という課題はあるものの，データの規模に関

表 4　提案手法および各比較手法の性能比較（$N = 120$）

比較手法	実行成功率	Precision	Recall	F1 スコア	平均時間 (s)
一括方式（スコアなし）	71.0%	0.709	0.833	0.766	**6.78**
一括方式（スコアあり）	70.0%	0.764	**0.876**	**0.817**	9.15
2 段階方式（スコアなし）	**100.0%**	**0.813**	0.732	0.771	262.93
提案手法（2 段階＋スコアあり）	**100.0%**	0.798	0.771	0.784	316.73

表 5　モデル別のラベル別抽出精度の詳細比較（$N = 200$）

ラベル	モデル	BSR	適合率	再現率	F1 スコア	平均時間 (s)
True	GPT-4o	**1.000**	**0.761**	**0.989**	**0.860**	**390.72**
	Qwen2.5	0.930	0.382	0.930	0.542	1306.85
False	GPT-4o	0.921	**0.569**	**0.842**	**0.679**	**375.43**
	Qwen2.5	**0.923**	0.489	0.718	0.582	2268.03
Partially True	GPT-4o	0.955	**0.617**	**0.864**	**0.720**	**396.46**
	Qwen2.5	**0.971**	0.511	0.857	0.640	1498.66
Undeterminable	GPT-4o	**1.000**	0.729	**0.692**	**0.710**	**335.43**
	Qwen2.5	0.949	**0.810**	0.462	0.588	1880.00
全体平均	GPT-4o	**0.969**	**0.669**	**0.847**	**0.742**	**374.51**
	Qwen2.5	0.943	0.548	0.742	0.588	1738.39

わらず処理を完遂できる「堅牢性」は，実務における事実確認支援システムとして不可欠な要素であると言える．

第二に，抽出精度における指標間の特性差について考察する．表 5 の結果より，全体として適合率よりも再現率が高い傾向が見られた．これは，モデルが正解行を確実に含めるために，正解の周辺行や類似した属性を持つ行に対しても高いスコアを付与する「過剰検知」が発生したためと考えられる．特に，輸出入統計のように類似した数値や項目名が並ぶデータにおいて，LLM が直接的な根拠だけでなく比較対象となる背景情報まで「関連あり」と広義に捉える傾向が確認された．

第三に，ラベル別の精度差とモデルごとの特性について考察する．本実験では「False」や「Undeterminable」の F1 スコアが低迷する傾向が見られたが，その要因はモデルによって異なると推察される．GPT-4o においては，言説と統計データ内の数値が一致しない場合に「関連なし」と判断する傾向が強く，数値の表層的な一致を優先したことが精度低下を招いたと考えられる．一方，Qwen2.5 においては，「True」ラベルで再現率が極めて高い（0.930）一方で適合率が極端に低い（0.382）という特徴が見られた．これは Qwen2.5 が数値の一致・不一致に関わらず，特定のキーワード等に基づいて過剰に関連箇所を抽出する傾向があることを示唆している．このように，モデルの種類によって，真偽ラベルに応じた抽出の挙動に明確な差異が存在することが確認された．

第四に，実用上の運用形態と今後の課題について述べる．本実験の結果，一括入力が可能な小規模なデータに対しては一括入力方式が精度面で優位であり，大規模なデータに対しては 2 段階推論が堅牢性の面で不可欠であることが明らかになった．したがって，実システムにおいては，入力される統計データの行数やトークン量に応じてこれら 2 つの手法を動的に切り替えるハイブリッド型の抽出アルゴリズムが有効であると考えられる．

今後は，どの程度のデータ規模を閾値として手法を切り替えるべきかの最適化に加え，過剰検知を抑制するためのプロンプトエンジニアリングや，関連度スコアに対する適切な閾値の設定が必要である．また，数値が不一致である場合でも，それが「否定の根拠」であることを LLM に正しく認識させるための推論プロセスの改善も，事実確認支援の精度向上において不可欠な課題であると言える．

6 ま と め

本稿では，ネット上の言説に対する統計データを用いた事実確認支援の 3 段階のうち，第 2 段階である「統計データ内の関連箇所抽出」に焦点を当て，その実現のために必要となる書き込みテキスト，統計データ，関連箇所の 3 つ組から構成されるデータセット構築を試みた．データセット構築においては，統計データ内の特定の一行を関連箇所とし，当該箇所の数値情報と統計データの背景情報であるメタデータを LLM に入力することで言説を生成した．その際，生成される言説には実際のソーシャルメディアで見られるような多様な真偽の度合いを反映した 4 種類のラベルを付与し，現実の事実確認の状況を再現する基盤を構築した．

本データセットを用いた関連箇所抽出実験では，データのバッチ化と独自の「関連度スコア」に基づく段階的な絞り込みを行う「2 段階推論システム」を提案し，その有効性を検証した．実験の結果，一括入力方式ではトークン制限により実行成功率が 70%台に留まったのに対し，提案手法は 100.0%を達成し，大規模な統計データに対する堅牢性を示した．また，関連度スコアリングを用いた判定により，提案手法の第一段階におけるバッチ特定成功率（BSR）においても高い精度を記録し，

膨大なデータの中から参照すべき箇所を効率的に絞り込めることを示した．

一方で，適合率が再現率を下回る傾向が見られたことから，類似した数値が並ぶデータにおいて正解以外の行にも高い関連度スコアを付与してしまう「過剰検知」が課題として明らかになった．さらに，モデルごとの特性差として，GPT-4o では数値の表層的な不一致を優先して関連なしと判断する傾向が，Qwen2.5 では特定のラベルにおいて過剰に抽出を行う傾向がそれぞれ確認された．

今後の課題としては，プロンプトの改良や関連度スコアの閾値の精査により，過剰検知を抑制し抽出の厳密性を向上させることが挙げられる．また，本実験で明らかになった手法間の特性を踏まえ，入力データの規模に応じて一括入力方式と 2 段階推論を動的に切り替えるハイブリッド型システムの構築も検討すべき重要な課題である．今後は，関連箇所が複数行に跨るケースや離れた数行に点在する複雑なデータセットへの拡張を行い，より高度な事実確認支援システムの実現を目指す．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] 黒川博生, 宮森恒. 大規模言語モデルを用いた文書補強とリランキングによる統計データ検索. 情報処理学会論文誌データベース (TOD), Vol. 18, No. 3, pp. 20–34, 2025.

[2] 宮﨑隆豪, 宮森恒. 統計データを用いた事実確認支援のための統計データ内の関連箇所抽出. In *DEIM2025*, pp. 8f–02, Kyoto, Japan, 2025. Information Processing Society of Japan.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 64–77, 2020.

[6] Dheeru Dua, Odin Wang, Aniruddha Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

[7] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, Online, July 2020. Association for Computational Linguistics.

[8] Jonathan Herzig, Paola Spangher, Jonathan Bogin, Ronen Chen, and Jonathan Berant. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 11270–11285. Association for Computational Linguistics, 2020.

[9] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6859–6866, 2019.

[11] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Chao Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 892–901, 2019.

[12] Pritam Deka and Ashwathy Revi. PD-AR at ArAIEval shared task: A BERT-centric approach to tackle Arabic disinformation. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pp. 570–575, Singapore (Hybrid), December 2023. Association for Computational Linguistics.

[13] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. Overview of the NTCIR-15 data search task. In *Proceedings of the NTCIR-15 Conference*, 2020.

[14] OpenAI. Gpt-4o technical report. https://openai.com/research/gpt-4o, 2024. Accessed: 2024-08-05.

[15] Qwen Team. Qwen2.5: A party of foundation models, September 2024.