事実確認支援のための言説と統計データ関連箇所との整合性検証

樫山　和貴†　　宮森　　恒†

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
E-mail: †g2253280@cc.kyoto-su.ac.jp

あらまし　本稿では，SNS 等における言説の真偽判定を支援するため, 統計データを活用して言説と統計データの関連箇所の整合性を検証する問題に取り組む. 従来の手法では，統計データの構造や暗黙的な意味関係を捉えることができず，キーワード検索や一般的なテキストベースの自然言語推論に依存しており正確な判断が難しいという課題があった．本研究は，統計データを活用し, 言説と統計データ間の意味的関係を捉えた整合性を検証する． 本稿では，言説テキストと統計データ関連箇所を用いた多クラス含意関係認識の問題として定式化する．言説テキスト, 関連箇所, 含意関係ラベル（含意/矛盾/部分的に真/判定不能）の 3 つ組からなるデータセットが不可欠であるが, 既存研究では提供されていない．そこで, 統計データと関連箇所から LLM を用いて SNS 投稿を模した言説テキストを生成し，人手による精査と修正を加えて含意関係ラベルを付与することで，数千件規模のデータセットを構築する．実験では, 構築したデータセットを用いて統計的推論に対応した含意関係認識モデルの性能を評価する．その結果，本手法が統計データに基づく事実確認において高い整合性検証能力を持つことを示す.

キーワード　事実確認支援，統計データ検索，整合性検証，含意関係認識，ソーシャルメディア

1　はじめに

現代社会において，SNS では，情報共有やコミュニケーションの主要な手段として広く利用されている．その利便性と迅速性から，個人や組織が情報を発信し，受け取るスピードは劇的に向上した．その一方で，誤情報やデマが拡散しやすいという問題が顕在化しており，統計データや事実関係を歪めた情報が拡散されることで，社会的な混乱や誤解が生じ，時には政策決定や個人の判断に重大な影響を与えることがある．近年，SNS 上では，新型ウイルスの感染拡大に関する誤情報が急速に拡散し，ワクチン接種率の低下や不安の増大を招いた事例がある．誤情報は信頼性の低い出所や意図的な操作による場合が多く．正しい情報との区別が難しいことが課題であり，こうした問題を解決するため，事実確認支援システムの開発が必要である．特に, 統計データを基にした客観的な検証は，情報の信頼性を高めるために重要である．以上の背景から，本研究では，SNS 等の投稿内容と公的な統計データとの整合性を検証するためのデータセットを構築し，構築したデータセットを用いて整合性検証を検証する.

統計データを用いた整合性検証は，事実確認の客観性を確保するための重要な手法である．統計データは，政府機関や研究機関などの公的機関によって収集・公開される信頼性の高い情報であり，一定の方法論に基づいて収集・分析されているため，検証基準として活用することで主観的な判断を排除することができる．また，経済指標や人口動態などの統計データは，信頼性が高いだけでなく，広範囲にわたるテーマを包括しているため，さまざまな分野での整合性検証に適用可能である．統計データを基盤とした検証は，個人ユーザーが情報の信憑性を判断する際の大きな助けとなる．SNS やニュースサイト上で多種多様な情報が飛び交う中，信頼性の高い統計データを基準として活用することで，ユーザーは主観的な判断に頼ることなく，情報の正確性を効率的に確認できる.

データセット構築の基盤として，NTCIR-15 が提供する統計データを活用する．NTCIR-15 は信頼性の高い統計データを提供しており，整合性検証のためのベースとなるデータとして適切である．SNS 上の投稿内容には，膨大かつ多様な情報が含まれており，それらの正確性を検証することは大きな課題である．本研究では，現実の SNS 投稿を直接使用するのではなく，NTCIR-15 の統計データを基に，SNS 上に見られるような書き込みテキストを生成するアプローチを採用した．この手法により，統計データとの直接的な整合性検証が可能となり，研究の効率性と信頼性を高めることができる.

本稿の目的は，SNS 等における言説に対し，統計データを用いてその真偽を判断する支援を行うことである．SNS は情報共有の重要な場である一方で，誤情報や誤解を招く投稿が拡散されるリスクも高い．これにより社会的混乱が生じることから，真偽を客観的かつ効率的に検証する仕組みが求められている．本研究では，公的に信頼のおける統計データを基盤とし，SNS 投稿の内容と統計データの間に整合性があるかを検証するモデルを構築することを目指す.

本提案手法では，信頼性の高い NTCIR-15 の統計データを基盤とし，大規模言語モデルを活用して，SNS 投稿を模倣したデータセットを生成する手法を採用する．具体的には，統計データを基に「統計データを支持する言説（True)」「統計データと矛盾する投稿（False)」「統計データを部分的に支持する言説 (Partly_true)」「統計データから判定不能な言説（Undeterminable)」の 4 つのカテゴリに分類される投稿データを構築する．この生成プロセスにより，統計データとの整合性を直接的

に検証できるデータを生成することが可能となる．

次に，生成された言説を評価し，その内容の適切性および信頼性を詳細に分析する．評価基準としては，情報の正確性，論理的一貫性，表現の明瞭性を設定し，特に統計データとの整合性を重視する．具体的には，言説が統計データに基づいて正確に記述されているか，ならびに誤解を招く表現や矛盾が含まれていないかを精査し，検証を行う．この評価を通じて，生成された言説の品質を把握し，統計データに基づく適切な情報提供の可能性を検討する. さらに，評価結果を分析し，言説の信頼性向上に向けた課題を明確にすることで，より精度の高い事実確認支援の実現を目指す.

本研究の特徴は，NTCIR-15 の統計データを基に，「言説テキスト」「統計データ関連箇所」「含意関係ラベル（True/False/partly_Ture/Undeterminable）」の 4 つからなるデータセットを構築し，4 つの多クラスラベルに対して，表形式データを維持したまま整合性検証を行うことである．このデータセットは，含意関係認識モデルの訓練および評価の基盤として機能し，SNS 投稿が統計データを支持する（True），矛盾している（False），部分的に支持する（Partly_True）, 判定不能（Undeterminable）を分類する能力を正確に測定することを可能にする．これにより，提案手法が SNS 投稿の真偽判定に有効であるかを明らかにし，SNS 上の情報信頼性向上に貢献することを目指している．

2 関連研究

2.1 自動化ファクトチェック

近年フェイクニュースの拡散防止を目的とした自動ファクトチェックの研究が活発化している. [11] が提唱した枠組みをはじめとして自動ファクトチェックのプロセスは，主に 4 つのタスクで構成されている．第一に「主張検出」であり，ドキュメント内から検証すべき重要な主張を抽出する． 第二に「証拠の取得」であり信頼性の高いデータベースや，統計情報，Wikipedia などから主張に関連する証拠を収集する．この段階では，従来のテキスト情報のみならず，画像や動画などのマルチモーダルデータを証拠として扱う研究も進展している． 第三に「判定予測」であり，収集した証拠と主張を突き合わせ含意関係認識などの技術を用いて真偽を判断する． 最後に「根拠生成」であり，なぜそのような判定に至ったかの理由を提示する．この一連のプロセスにより，説明可能で透明性の高いシステムの実現が目指されている．自動化ファクトチェックでは多くの研究が「支持」「反論」「情報不足」の 3 値分類あるいは単純な「真」「偽」の 2 値分類に焦点を当ててきた．しかし，現実では必ずしも白黒が明確ではなく，一部の情報が正しいものの誤解を招く表現や文脈に依存する複雑な構造を含んでおり，単純なラベル付では微細なニュアンスに対応できない．また，既存のデータセットには特定の単語が含まれると「偽」になりやすいといったバイアスが含まれておりこれがモデルの汎化性能に悪影響を与える可能性も示唆されている [5] こうした背景からよりきめ細やかな判定を可能にする新たなラベル付け手法やフレームワークが求められている． そこで本研究では，言説と統計データの整合性を検証する際に単純な二値分類ではなく，「含意」「矛盾」「部分的含意」「判定不能」の 4 クラスを用いたラベル付けを行う．これにより統計数値とテキスト間の複雑な意味関係を精緻に捉え，より高精度かつ実用的なファクトチェック判断の実現を目指す. .

2.2 表を用いたファクトチェック

従来のファクトチェック研究の多くは，Wikipedia の記事やニュース記事などの非構造化テキスト（自然言語文）を主要な証拠源として利用してきた．しかし，世界中の Web 上には，統計データ,，スポーツの試合結果など，豊富な情報が表形式やデータベースなどの構造化データとして存在している．これらの構造化データは従来のテキスト中心のアプローチでは検証が困難であった数値的な主張や比較を含む主張に対して不可欠な証拠を提供する．そのため表形式データを活用したファクトチェックは検証範囲の拡大と精度の向上を実現する新たなアプローチとして注目されている. [3] 表形式データを用いた検証は通常のテキスト処理よりも高度な推論能力を必要とする．テキストデータが主に言語的な意味理解を要するのに対し表データは行や列の構造を理解した上で記号的推論を組み合わせる必要があるためである．Wikipedia の表に基づいた真偽判定タスクである TabFact や，テキストと表の両方を証拠として扱う FEVEROUS の提案は，構造化データの理解における新たな挑戦領域を生み出している. [12] [1] [4] こうしたデータセットの登場に伴い検証モデルの研究も急速に進展し，SQL クエリの実行履歴を通じて表構造を学習する TAPEX を提案し，TabFact における SOTA 達成を通じて，表をフラットなテキストとして扱う従来手法に対する構造的理解の重要性を実証した. [7] 一方で，膨大なテキストデータによる事前学習を通じて高度な汎用推論能力を獲得した LLaMA や GPT シリーズに代表される大規模言語モデル（LLM）の発展も，この分野に大きな影響を与えている. [10] しかし既存手法の多くは依然として「真」か「偽」かの二値分類に焦点を当てており，統計データが持つ複雑な含意関係を十分に扱えていない場合がある．本研究では TAPEX のような構造理解に優れたモデルや Llama 等の LLM が持つ高度な言語生成・理解能力を活かしつつ，統計データに基づいて生成されたテキストとの整合性を多クラスラベルを用いてる検証する.

2.3 統計データを用いたデータセット構築

統計データなどの構造化データを自然言語記述に変換する Data-to-Text タスクは，情報の信頼性と説明性を向上させる重要な手段として注目されている．特に，大規模なデータセットから得られる有用な情報を，一般のユーザーが理解しやすい形で表現できる点は大きな利点である． しかし，表データの多様な形式への対応や，文脈に応じた適切な数値情報の選択，過不足のない要約，さらには生成文の流暢さと事実への忠実性の両立など，解決すべき課題は多岐にわたる．「WikiStatCells」では，Wikipedia 記事と統計データとの正確な対応関係をアノ

テーションすることにより，統計データを基にしたテキスト生成モデルの学習や評価，および統計データ検索タスクに貢献している.. [13] [2] また，「ToTTo」では選択されたセル情報に対して忠実な説明文を生成するタスクを提案しており，構造化データの活用に関する研究は深化している. [9] しかし，既存のデータセットの多くは Wikipedia のような解説的な長文テキストを主な対象としており，SNS 上の投稿に見られるような主観的，あるいは断定的な短文言説とは性質が異なる． そこで本研究では，NTCIR-15 の統計データを基盤として，SNS 投稿を模倣したデータセットの構築を行う． 具体的には，統計データとの整合性に基づき，単純な二値分類ではなく，「含意」「矛盾」に加え，「部分的真」「判定不能)」の計 4 クラスのラベルに対応する言説を生成する．これにより，より複雑で現実的なシナリオにおける整合性検証を目指す.

3 データセットの構築

本研究の目的である統計データに基づいた言説の整合性検証を行うためには，信頼性の高い統計データと，それらを参照して生成された多様な言説（含意・矛盾・部分的に真・判定不能)，および正解ラベルが紐付いた大規模なペアデータが必要不可欠である．しかし，既存のデータセットの多くは Wikipedia のような解説文を対象としており，SNS 上で見られるような主観的かつ断定的な短文言説と統計データを体系的に結びつけたものは存在しない．そこで本研究では，NTCIR-15 の統計データをソースとして，大規模言語モデルを用いた自動生成プロセスにより，SNS 投稿を模した言説データセットを構築する．本データセットの構築プロセスは，主に「統計データの前処理」,「関連箇所の特定」,「言説生成とラベリング」の 3 つのから構成される．以下に各手順の詳細を述べる.

3.1 統計データの前処理

本研究における言説生成では，大規模言語モデルを活用する．しかし，現行の LLM には一度に入力可能なトークン数に厳格な上限が存在する．そのため，NTCIR-15 で提供されるような M 行 N 列の巨大な統計データ全体を，そのまま LLM に入力して分析や言説生成を行うことは不可能である．このような技術的制約の下では，入力情報の質と量を最適化するための前処理が不可欠となる．単にデータを切り捨てるのではなく，限られた入力枠内で，いかにして元データの本質的な情報を LLM に伝達するかが極めて重要である．特に，行数 M が大きい場合，後段の処理における計算コストの増大や，情報過多による多重共線性といった問題を引き起こす可能性がある．そこで本提案手法では，これらの課題に対処し，データに内在する構造を抽出するために，前処理として次元削減を適用する．具体的には，元の M 行 N 列のデータ行列を，N 行 N 列程度のサイズに縮小・分割し，複数の扱いやすい統計データとして再構成する.

3.2 関連箇所の特定と文脈化

大規模言語モデルを用いた言説生成において，入力プロンプトに含まれる情報の質と密度は，生成されるテキストの正確性および論理的整合性を決定づける最も重要な要因である．広範かつ多岐にわたる統計データ全体を入力するのではなく，分析の核となる「関連箇所」を的確に特定し，LLM が焦点を絞って解釈できる形式で提示するプロセスが不可欠となる．本研究では，LLM に「どのデータを見て」「どのような文脈で」判断すべきかを明確に指示するため，以下の手順を用いて関連箇所を特定し，言説生成のためのコンテキストを構築する.

1. **データチャンクの選定**
   前節の前処理によって分割・縮小された複数のデータチャンク群（例：「統計データ名_列数_連番.csv」）の中から，分析対象とするファイルをランダムサンプリングによって選定する．ファイル名に含まれるメタデータは，そのデータが元データのどの部分に由来するかを示す情報として保持する.
2. **対象行の指定**
   次に，選定されたデータチャンク内に含まれるレコードの中から，言説生成の主たる対象となる特定の 1 行をランダムに指定する．統計表には多数の項目が含まれるが，LLM に対して「この行のデータに着目せよ」という明確なアテンションを与えることで，生成される言説の主題を固定する.
3. **周辺行による文脈の付与**
   指定されたターゲット行単体の数値情報のみでは，その値が全体の中で高いのか低いのか，あるいは時系列変化の中でどのような位置にあるのかといった「相対的な傾向」を LLM が理解することは困難である．実際，予備調査において周辺行を含まない設定（$k = 0$ または $k = 1$）で生成を試みたところ，順位関係や前後の文脈を無視した言説が生成される傾向が見られた．そこで本手法では，ターゲット行を中心として，その前後 5 行ずつ（計 11 行）を一つの意味的なまとまりとして切り出す処理を行う.
4. **メタデータの結合と構造化**
   切り出した行データに対し，表のヘッダー情報や単位といったメタデータを明示的に結合する．これにより，単なる数値の羅列を，人間が読むのと同等の「意味のある統計情報」へと変換する.

3.3 言 説 生 成

LLM を用いた言説生成においては，生成されたテキストに対し，「True」「false」「Partly_True」「undeterminable」の 4 種類の正解ラベルを付与する．ここで，各ラベルはそれぞれ「含意」「矛盾」「部分的含意」「判定不能」に対応するし，ラベリングは言説生成と同時に LLM に行わせる．プロンプトを通じて，統計データの内容と言説の整合性を生成時に検証させることで，各言説に対して論理的に適切なラベルが分類・付与されるように設計する.

4 提案手法

4.1 問題設定

本研究におけるタスクは，統計データ（表形式データ）の関連箇所 T と，それを根拠としているとされる言説テキスト S を入力とし，両者の意味的な整合性ラベル y を予測する多クラス分類問題として定式化される．

具体的には，関連箇所 T と言説テキスト S のペアに対して，最適なクラス $y \in \mathcal{Y}$ を出力する予測モデル f を構築することを目的とする．

$$y = f(T, S) \tag{1}$$

ここで，分類対象となるクラスラベルの集合 $\mathcal{Y}$ は，以下の 4 つのカテゴリにより定義される．

- **True (含意)** (y_{true}): 言説 S の内容が統計データ の関連箇所 T によって論理的に支持される．
- **False (矛盾)** (y_{false}): 言説 S の内容が統計データの関連箇所 T と矛盾する，あるいは事実と異なる記述を含む．
- **Partly_True (部分的真)** (y_{part}): 言説 S の一部は統計データの関連箇所 T によって支持されるが，一部は支持されない，あるいは不正確な記述を含む．
- **Undeterminable (判定不能)** (y_{und}): 統計データの関連箇所 T の情報のみからは，言説 S の真偽を論理的に導出できない．

検証対象となる関連箇所 T には，NTCIR-15 に含まれる信頼性の高い公的統計データを用いる [6]．本来，この予測モデル f の学習には，実際の Web 上の言説と統計データが紐付いた大規模なデータセット $\mathcal{D} = \{(T_i, S_i, y_i)\}_{i=1}^{N}$ が必要となる．しかし，現状では統計データを引用した Web 上の言説は体系的に整理されておらず，教師データとして利用可能なリソースが著しく不足している．

そこで本研究では，NTCIR-15 の統計データを基に大規模言語モデルを用いて生成した「疑似的な SNS 投稿」を検証対象とする．すなわち，本研究の問題設定は，生成された言説 S が元の統計データの関連箇所 T に対して正しい含意関係 y を持っているかを，提案モデルがいかに正確に識別できるかを評価することにある．

4.2 整合性検証

4.2.1 TAPEX の採用

統計データのような構造化データと自然言語テキストの間の整合性を検証するためには，表の構造を正しく理解し，数値の比較や集計といった推論を行う能力が不可欠である．一般的な大規模言語モデルは高い言語能力を持つ一方で，複雑な表構造の認識や厳密な数値計算を苦手とする場合がある．そこで本研究では，整合性検証モデルの基盤として，表形式データの理解に特化した事前学習済みモデルである TAPEX を採用する．TAPEX は，表に対する SQL クエリの実行結果を予測するというタスクで事前学習されており，これにより表の構造や数値間の関係性を深く理解する能力を獲得している．この特性は，統計データに基づいた事実確認タスクにおいて極めて有効であると考えられる．

4.2.2 学習データの入力形式

TAPEX は Transformer ベースのエンコーダ・デコーダモデルであり，入力として 1 次元のテキストシーケンスを要求する．そのため，2 次元構造を持つ統計データを構造情報を欠損させることなく自然言語のシーケンスへと変換する処理が必要となる．本研究では，各セルを「ヘッダー名 : セルの値」のペアとして記述し，それらを特殊トークンで連結する．

さらに，モデルの推論精度を向上させるため，検証対象の言説が直接言及しているターゲット行の全セルに対して特定の記号を付与し，表内の重要なコンテキストを明示的に強調した．第二に，ターゲット行のみならず，その前後各 5 行を周辺コンテキストとして含めることで，最大 11 行の局所的な表構造を保持した状態で入力を行うこととした．

4.2.3 ファインチューニング

構築した学習データセットを用いて，TAPEX に対して 4 クラス分類タスクのファインチューニングを行う．モデルは，入力された統計データと言説のペアから，両者の意味的関係が「True」「False」「Partly_True」「Undeterminable」のいずれであるかを予測するように学習する．汎用的な LLM にゼロショットで推論させるのではなく，統計データに基づく含意関係認識という特定のタスクに特化した学習を行うことで，特に「Partly_True」や「Undeterminable」といった複雑なニュアンスを含む関係性に対しても，高精度な判定能力を獲得することを目指す．以上のデータセット構築から TAPEX による整合性検証に至る，本提案手法の全体的な処理概要を図 1 に示す．

5 評価実験

5.1 実験目的

本実験の目的は，構築したデータセットおよび TAPEX を用いた提案手法が，統計データに基づく事実確認タスクにおいてどの程度有効であるかを定量的に検証することである．従来の一般的な言語モデルを用いた手法では，2 次元的な構造を持つ表データを 1 次元のテキスト列に変換して処理するため，行や列の対応関係や，セル間の階層構造といった重要な情報が失われる傾向にあった．これに対し，本研究で採用する TAPEX は，表構造を理解するための事前学習が行われている．本実験では，このモデルが統計データの構造的特徴を適切に認識し，数値の比較や集計といった高度な推論を必要とする整合性検証において，高い精度を発揮できるかを確認する．既存の多くのファクトチェック研究は「真」か「偽」かの二値分類に焦点を当ててきた．しかし，現実の言説には，一部は正しいが一部は誤っている「部分的真」や，提示されたデータだけでは判断できない「判定不能」といった複雑なケースが多々存在する．本実験では，提案手法がこれらの曖昧性を含む 4 つのラベル（True, False, Partly_True, Undeterminable）をどの程度正確に識別できるかを検証し，より精緻で実用的な事実確認支援が可能であることを示す．

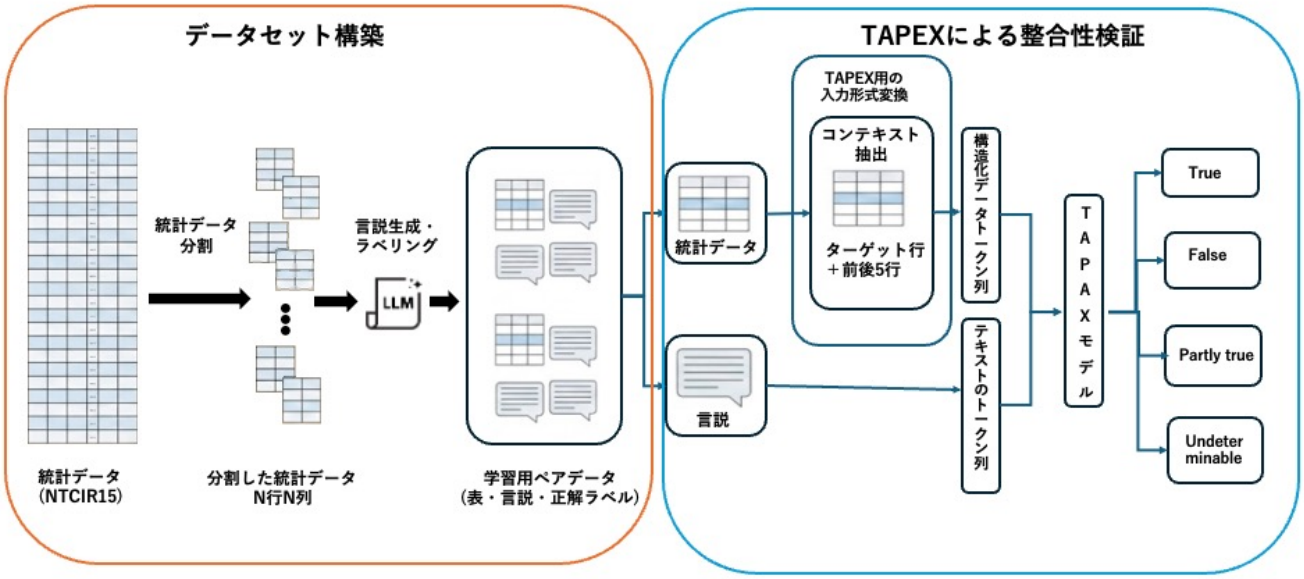

図 1　データセット構築および整合性検証の処理フロー

5.2 実験方法

本実験では，NTCIR-15 のデータセットから収集したデータに対し，第 3 章で述べた前処理を施した 500 件の統計データチャンクを使用した．これは，巨大な元データをそのまま使用するのではなく，モデルの入力制限および検証の焦点化を考慮し，意味的なまとまりを持つ単位（$N' \times N'$ サイズ）に再構成したものである．

これらの統計データチャンクをソースとして，GPT-4o [8] を用いて言説生成を行った．生成にあたっては，各統計データが示す数値情報に加え，関連箇所およびメタデータをプロンプトとして入力し，統計データとの整合性が 4 つのラベル（True, False, Partly_True, Undeterminable）のいずれかに該当する言説を出力させた．最終的に，合計 12,000 件の言説テキストを生成し，これを本実験のデータセットとした

5.3 比較手法

本研究の提案手法である TAPEX の有効性を検証するため，比較手法として汎用的な大規模言語モデルである Llama を採用する． 両手法とも入力データをトークン列として処理する点は共通しているが，表構造の取り扱い方に決定的な差異がある．

TAPEX は，入力トークンに対して行および列のインデックス情報を埋め込み表現として明示的に付与することで，表の構造情報を保持したままエンコードが可能である．

対照的に，Llama などの一般的な LLM は，表データを構造化データとして扱うための専用の入力層を持たない．そのため，比較実験における Llama への入力では，統計データを Markdown 形式や「ヘッダー名: 値」の形式等のテキストシーケンスに変換して入力する手法をとる．この場合，モデルはテキスト上の区切り文字や改行などのパターンから表構造を暗黙的に推論する必要がある．

本実験では，表の各セルの座標（行・列）を直接モデルに入力する TAPEX のアプローチと，テキスト形式に変換して汎用モデルの推論能力に委ねる従来的な LLM アプローチとの性能差を明らかにする．

5.4 データの分割と評価手順

構築した合計 12,000 件のデータセットを，学習用データとして全体の 80%にあたる 9,600 件，検証用データとして 10%にあたる 1,200 件，そして評価用データとして残りの 10%にあたる 1,200 件に分割した．

この際，学習データと評価データの間で，元となる統計データチャンクが重複しないように配慮し，統計データチャンクから生成された言説群は，すべて学習用か評価用のどちらか一方にのみ含まれるように分割を行った．これにより，モデルが統計データの内容そのものを丸暗記するのではなく，未知の統計データに対しても構造を理解し，汎化性能を発揮できるかを厳密に検証できる設定とした．

さらに，データの偏りによる影響を排除し，実験結果の信頼性と堅牢性を担保するため，上記の「データのランダム分割」および「学習・評価」の一連のプロセスを 4 回繰り返した．最終的な評価には，これら 4 回の試行における評価指標の平均値を採用する．

5.5 学習と評価指標

学習フェーズでは，前節で分割した学習用データ 9,600 件を用いて各モデルをファインチューニングし，4 クラス分類の推

論能力を学習させた．また，検証用データ 1200 件は，学習過程におけるモデルの挙動監視およびハイパーパラメータの調整に用いた． データのランダム分割によるバイアスを排除するため，この学習プロセスは分割手順と同様に 5 回独立して実施した．

続く評価フェーズでは，学習および検証には一切使用していない 600 件の評価用データを各試行のモデルに入力し，推論を行わせた． モデルが出力した予測ラベルと，データセット生成時に付与された正解ラベルとを比較し，各ラベルごとの適合率（Precision），再現率（Recall），F1 値を算出した． 最終的な性能評価には，これら 4 回の試行で得られた各指標の平均値を用い，提案手法の有効性を定量的に評価した．

5.6 評価結果

5.6.1 混同行列を用いた誤分類の傾向分析

モデルの誤分類における傾向を詳細に分析するため，表 1 および表 2 に混同行列を示す．本行列は，行に正解ラベル，列に予測ラベルを配置し，4 回の試行における各ラベルの平均件数を算出したものである．

分析の結果，TAPEX は Llama と比較して，テーブル構造に基づいた極めて精緻な推論を行っていることが明らかになった．

正解が「True」の事例において, TAPEX が「Undeterminable」と誤認した件数はわずか 0.5 件にとどまった．これは，対照的な結果となった Llama（7.0 件）の 14 分の 1 という極めて低い数値である．Llama は検証対象が表内の多岐にわたる場合に情報の集約に失敗し，Undeterminable を選択する傾向があったのに対し，TAPEX は広範囲のセルを参照する必要がある場合でも，必要な根拠を的確に抽出できていた．特に，2 箇所以上のセルを横断的に参照する推論において，Llama は推論の複雑化に伴い確信度を低下させ「Partly_True」や「Undeterminable」へと分類する傾向が見られたが，TAPEX はテーブル構造に特化した事前学習の恩恵により，安定した判定を維持していた．

一方で，両モデルに共通して見られた課題として，数値の桁数が大きい場合に正解ラベルを正しく識別できず，他のラベルへ誤分類してしまう傾向が挙げられる．特に，正解が「False」であるにもかかわらず他のラベルへと判断してしまう事例において，対象となる数値が 6 桁以上の大きな値である場合，モデルが一部の桁の不一致を無視し，全体的な整合性のみで判定を下してしまう傾向が確認された．

TAPEX においてはこの傾向が顕著であり，正解が「False」の事例に対して True（27.8 件）といった他のラベルへ誤分類するケースが一定数発生している．これは，対象文中の数値が表内の値とわずかに異なる場合や，列数の多い大規模な表において一部の数値のみが改変されている場合に発生しやすい．

以上の分析から，TAPEX は表のレイアウト把握や論理構造の特定には極めて長けているものの，大きな桁数を含む微細な数値的不一致を厳密に識別できず，他のラベルを割り当ててしまう性質があることが示唆された．今後は，表構造の理解力を維持しつつ，数値の厳密な比較能力をいかに向上させるかが課題である．

表 1　TAPEX の推論性能に関する混同行列

正解ラベル	TAPEX による予測ラベル			
	True	False	Partly_True	Undeterminable
True	279.0	14.2	6.2	0.5
False	27.8	271.0	0.8	0.5
Partly_True	3.0	1.2	294.0	1.5
Undeterminable	1.2	0.2	0.8	297.8

表 2　Llama の推論性能に関する混同行列

正解ラベル	Llama による予測ラベル			
	True	False	Partly_True	Undeterminable
True	268.8	15.8	8.5	7.0
False	22.8	269.8	2.5	5.0
PPartly_True	10.0	4.0	278.0	7.8
Undeterminable	4.2	6.0	3.8	286.0

5.6.2 定量評価によるモデル比較

表 3 に，TAPEX と Llama を用いた場合のクラス別の適合率，再現率，F1 値を示す． 実験の結果，全てのクラスにおいて TAPEX が Llama を上回る F1 値を達成した．特に「Undeterminable」および「Partly_True」クラスにおいては，TAPEX は 0.97 を超える極めて高い F1 値を記録しており，表構造を理解する能力の高さが示唆された．

一方で，Llama（表 3）は全てのクラスで F1 値が 0.88〜0.95 の範囲に留まった．特に「True」クラスの F1 値は 0.887 と最も低く，事実性が明確なデータであっても，TAPEX と比較して正答率が劣る傾向が見られた． 全体を通して，表形式データに基づく含意関係認識タスクにおいては，事前学習段階で表構造を学習している TAPEX の方が，汎用 LLM である Llama よりも高い適性を持つことが定量的に示された．

表 3　TAPEX と Llama の性能比較

ラベル	適合率		再現率		F1 値	
	TAPEX	Llama	TAPEX	Llama	TAPEX	Llama
True	0.897	0.879	0.930	0.895	0.913	0.887
False	0.945	0.912	0.903	0.899	0.923	0.906
Partly_True	0.974	0.949	0.980	0.927	0.977	0.938
Undeterminable	0.991	0.935	0.992	0.953	0.992	0.944

5.7 考　　察

定量評価（表 3）において，TAPEX が全指標で Llama を上回った主要因は，事前学習における「表形式データとクエリの整合性」の学習量および、表構造を保持したエンコード手法にあると考えられる．特に「Partly_True」および「Undeterminable」における F1 値が 0.97 を超えている点は注目に値する．Llama のような汎用言語モデルは，表を線形なテキストとして処理するため，行・列の対応関係が複雑な場合に情報の欠落が生じやすい．これに対し，TAPEX は表の二次元的な構造を保持したまま処理を行うため，複数のセルを跨ぐ条件参照においても論理的な一貫性を維持できたと推察される．これは，Llama が「True」や「false」の事例を他のラベルへ誤分類し，再現率を低下させている結果（表 2）とも整合する．

混同行列の比較から，Llama 特有の振る舞いとして「Undeterminable への回避」が顕著に見られた．正解が「True」の事例を「Undeterminable」と誤認した件数は，TAPEX が 0.5 件であるのに対し，Llama は 7.0 件に上る．これは，Llama が長大なコンテキストや複雑な数値関係を処理する際，根拠の特定に至らずに判断を放棄する傾向があることを示唆している．

対照的に，TAPEX は判断を回避せず積極的にラベルを割り当てる傾向があるが，一方で数値的な脆弱性も確認された．具体的には，正解が「False」である事例を「True」ラベルへ誤分類するケースが合計 27.8 件発生している．これらの誤分類の多くは，同一行内に存在する複数の数値のうち一部のみが書き換えられた事例や，6 桁以上の大きな桁数を持つ数値において微細な不一致を含む事例で発生している．

これは，TAPEX が表の行と列の対応関係，すなわち「どの行にどの列が紐付いているか」という広域的な構造把握には極めて長けている一方で，同一行内の特定のセルに対する「局所的な数値検証能力」に課題があることを示している．言い換えれば，TAPEX は対象文に関連する「該当行」を正しく特定できたことに引きずられ，その行内の詳細な数値的不一致を見落とし，安易に「含意（True）」等のラベルを選択している可能性が高い．

以上のことから，TAPEX は表構造に特化した事前学習により，汎用 LLM を大きく上回る推論の安定性を獲得しているものの，大きな桁数を含む数値変化を厳密に識別する能力については依然として改善の余地があると言える．

6 ま と め

本研究では，SNS 等における言説の真偽判定を支援することを目的とし，統計データを用いた言説の整合性検証手法を提案した．信頼性の高い公的統計データである NTCIR-15 を基盤とし，大規模言語モデルを活用して SNS 投稿を模した言説データセットを構築した．本データセットは，従来の二値分類では扱いきれなかった曖昧な言説に対応するため，「含意」「矛盾」「部分的真」「判定不能」の 4 クラスで定義されている．

実験では，表構造の理解に特化した事前学習済みモデルである TAPEX と，汎用 LLM である Llama を用いて比較評価を行った．定量評価の結果，TAPEX は全てのクラスにおいて Llama を上回る F1 値を記録し，特に「判定不能」や「部分的真」といった複雑な論理関係の識別において高い優位性を示した．これは，二次元的な表構造を保持したままエンコードを行う TAPEX の構造的理解力が，統計データに基づく事実確認において極めて有効であることを示唆している．

また，誤分類の傾向分析を通じて，モデルごとの推論特性を明らかにした．Llama が複雑な情報に対して判定不能と出力する傾向があるのに対し，TAPEX は安定した推論を行う一方で，同一行内の微細な数値の不一致を見落とすという特有の課題も浮き彫りとなった．今後の課題として数値の厳密な比較能力の向上である．考察で述べた通り，現在のモデルは表構造の特定には長けているものの，数値の微細な差異を許容してしまう性質がある．数値の完全一致を判定するための外部モジュールとの連携や，数値の改変に敏感な学習手法の導入を検討する必要がある．

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] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–13, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Oana Balalau, Simon Ebel, Théo Galizzi, I. Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. Fact-checking multidimensional statistic claims in french. pp. 20–29, 2022.

[3] Tien-Duc Cao, I. Manolescu, and Xavier Tannier. Searching for truth in a database of statistics. *Proceedings of the 21st International Workshop on the Web and Databases*, 2018.

[4] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[5] Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. Factoring fact-checks: Structured information extraction from fact-checking articles. *Proceedings of The Web Conference 2020*, 2020.

[6] 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.

[7] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. Tapex: Table pre-training via learning a neural sql executor, 2022.

[8] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and et.al AJ Ostrow. Gpt-4o system card, 2024.

[9] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, Online, November 2020. Association for Computational Linguistics.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[11] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, *Proceedings of the ACL 2014 Work-*

shop on Language Technologies and Computational Social Science, pp. 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[12] Jianshu Chen Yunkai Zhang Hong Wang Shiyang Li Xiyou Zhou Wenhu Chen, Hongmin Wang and William Yang Wang. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, April 2020.

[13] 中野優, 加藤誠. 被引用統計データのセル特定データセットの構築. 日本データベース学会論文誌 データドリブンスタディーズ, Vol. 1, No. 1, 3 2023.