有価証券報告書における表形式データからの情報抽出

白藤 大幹 [1]　田中 宏治 [1]　斉藤 辰彦 [1]
[1] 三菱電機株式会社
{Shirafuji.Daiki@ay, Tanaka.Koji@bc, Saito.Tatsuhiko@db}
.MitsubishiElectric.co.jp

概要

大規模言語モデルは表に関する質問応答（Table Question Answering; TQA）にも広く用いられているが，表構造を把握できず誤った回答を生成することがある．本研究では，本課題を解決するためセル情報抽出手法を提案する．提案手法は，言語モデルと TF-IDF によるハイブリッド検索で質問と各セルの類似度を計算し表のヘッダを推定し，最も類似する行と列の交点に位置するセルを抽出する．性能向上のため，1,000 件の TQA データを用い言語モデルを対照学習する．評価実験には，有価証券報告書に関する TQA データセットを用いる．評価の結果，提案手法を適用した Ruri-v3-310m の正解率は 74.8% で，GPT-4o mini (63.9%) を上回る性能を示した．

1　はじめに

金融や医療など幅広い分野において，表情報は重要なデータ源である [1,2]．そのため，大規模言語モデル（LLM）などを用いて表データから必要な情報を抽出する技術は必要不可欠である [3,4]．

しかし，表の構造を考慮せずに，表全体を長文コンテキストとして LLM に入力すると，ヘッダの構成を正しく把握できずに誤った出力を生成する可能性がある [5]．例えば，経済分野において，木村ら [6] は，Table Question Answering（TQA）データセットを構築し，複雑なヘッダ構造を有する表や大規模な表では LLM 単体で正しい回答を生成しにくいという示唆を得た [7,8]．本課題は，表データから必要な情報を取得する必要性を示している．

本研究では，表データ向けのセル抽出手法を提案する [6]．提案手法の概要を図 1 に示す．本手法では，(1) データの前処理を実施し，(2) 言語モデルと TF-IDF を組み合わせたハイブリッド検索手法を用いて表ヘッダを推定することで質問に必要な情報を含むセルを特定し，(3) 表中から単位を抽出し，(4) 抽出した単位を用いて値を正規化し質問に回答する．ここで，検索に用いる言語モデルは，質問とそれに紐づく表のペア 1,000 件を用いて対照学習する．

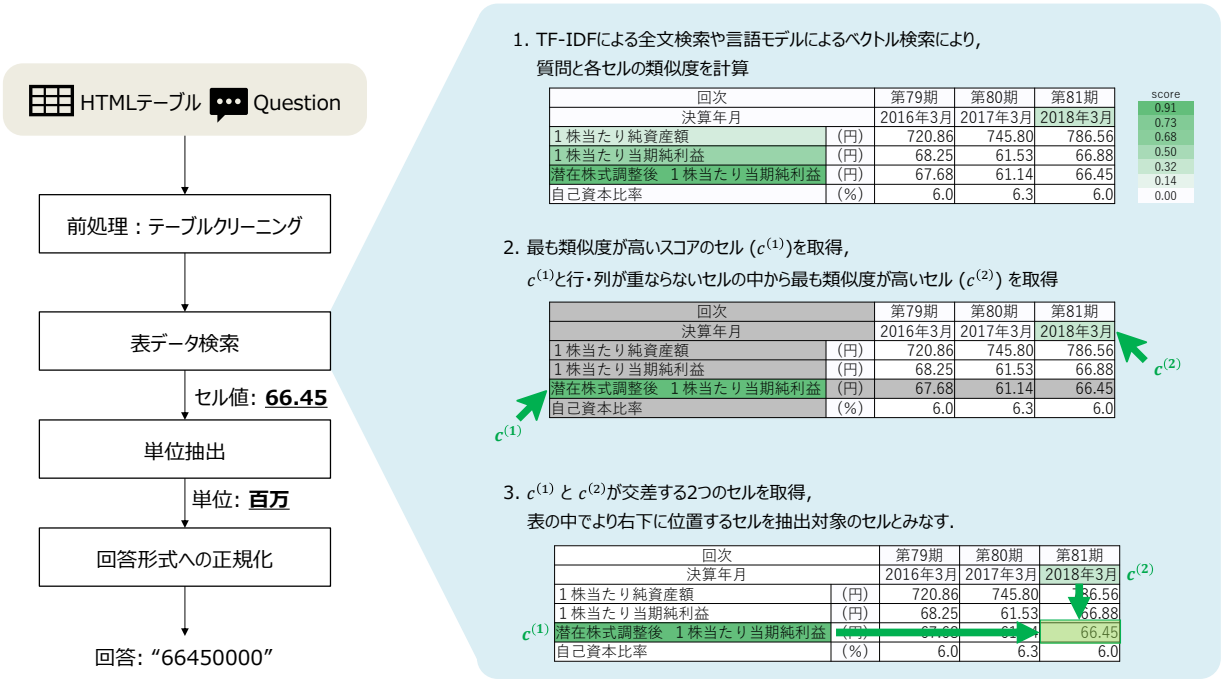
評価実験では，有価証券報告書をベースとした TQA データセット [6] を用い，Ruri-v3 [9] などの日本語対応の言語モデル 6 種類において評価する．また，財務知識を持たない日本語を母語とするアノテータの正解率を評価し，その結果を本手法と比較する．評価の結果，言語モデルに Ruri-v3 を用いた場合，GPT-4o mini の性能（63.9%）を上回る性能（74.8%）を示したものの，人手性能（85.0%）を下回ったことを確認した．

2　関連研究

表データから必要な情報を抽出するため，表内のヘッダを特定する必要がある [10]．ヘッダとは通常，表の列名やカテゴリを示す行・列で，QA タスクにおいてユーザが必要とする情報をテーブルから検索する際の手がかりとなる．このヘッダを推定するために，Fang ら [10] は文字列の書式やセル内容などに基づいて抽出した特徴量を用いて機械学習モデルを学習する手法を提案した．

表データに関する質問応答タスクについても表検索は広く研究されている [11,12]．Sun ら [13] は，質問が与えられた場合に，対応する行ヘッダおよび列ヘッダを特定し，それらが交差するセルを回答とみなす手法を提案した．しかし，既存手法は，ヘッダ内に出現する語と質問中の語が一致し，表中のヘッダがあらかじめ同定されていることを前提としている．

近年では，表データ検索の性能を向上させるために，言語モデルを用いたベクトル検索が頻繁に導入されている [12,14,15]．Jin ら [12] は，クエリ中のキーワードと表の列構造とを対応付けることを目的とした言語モデルを学習し，大規模な表コーパスからの効率的なセル抽出を可能にした．ベクトル検索



回次		第79期	第80期	第81期
決算年月		2016年3月	2017年3月	2018年3月
1株当たり純資産額	(円)	720.86	745.80	786.56
1株当たり当期純利益	(円)	68.25	61.53	66.88
潜在株式調整後 1株当たり当期純利益	(円)	67.68	61.14	66.45
自己資本比率	(%)	6.0	6.3	6.0

図 1　提案手法の概略図

を活用することで，従来のキーワードベース検索では取得できなかった情報を捉え，セル抽出における再現率を向上させている．これにより，用語一致の前提条件という従来の課題の解決を試みている．しかし，既存研究では表内のヘッダの事前特定を前提としている．

3　提案手法

本研究では，複雑なヘッダを含む表から回答に必要な情報を抽出する手法を提案する．本手法は，(1) データの前処理，(2) 表データ検索による関連セル特定，(3) 単位抽出，(4) 回答の正規化の 4 つのステップからなる．

3.1　前処理

本研究の対象データは HTML 形式で構成されており，データには装飾情報や，表でないテキストも含まれている．これらの要素を残すと抽出性能に悪影響を及ぼす可能性があるため，付録 A に沿ってデータのクリーニングする．

3.2　表データ検索

3.1 節にて整形した表データと対応する質問文を入力として，回答となるセルの値を推定する．

3.2.1　検索手法

質問をクエリ q，表データ d を対象に検索する．本研究では，TF-IDF と言語モデルによるベクトル検索を組み合わせたハイブリッド手法を用いる．検索スコア $s_H(q,d)$ は下式で定義される．

$$s_H(q,d) = (1-\alpha)s_v(q,d) + \alpha s_t(q,d),$$

ここで，α は TF-IDF スコア（$s_t(q,d)$）とベクトル検索スコア（$s_v(q,d)$）の比重を決めるハイパーパラメータである．

α を決定するために，日本語 Sentence-BERT を用いて 0 から 1 までを探索し，検証データセット上の正解率を比較し，最も高い性能を示した $\alpha = 0.21$ を採用した．

検索結果より，スコアが最も高い 2 つのセルを $c^{(1)}$ および $c^{(2)}$ として選択する．ここで，両方のセルが同一の行および列に属さないように制約を設ける．最後に，行と列の交点に位置する 2 つのセルのうち，右下に位置するセルを質問に対する回答とみなす．この手法により，表中のヘッダ位置が不明確な場合や，結合セルや多段ヘッダが存在する場合であっても，頑健に抽出できると期待される．

3.2.2　言語モデルの対照学習

言語モデルによる検索精度を向上させるため，モデルを TQA データセット [6] の訓練データから QA

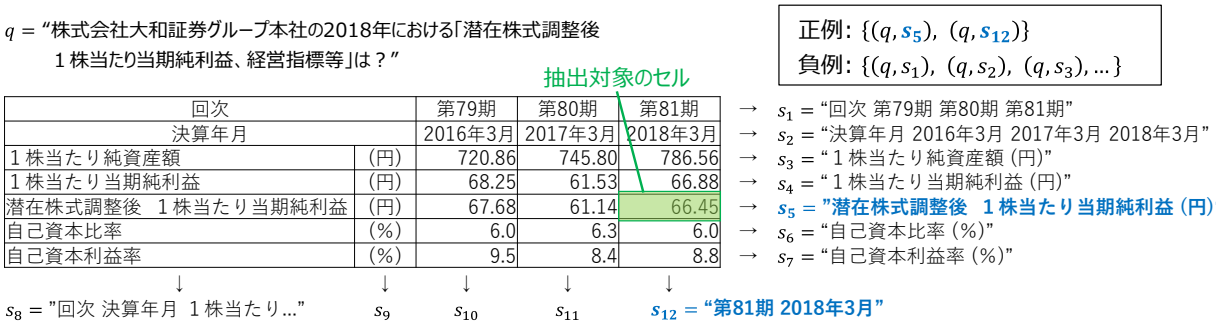


図 2　言語モデル学習用データセットの構築方法

とそれに紐づく表データのペアを 1,000 件取得し学習する．

学習データは，図 2 のように，表データの各行・列に含まれるすべてのセルから数値や記号以外のテキスト情報を抽出・それぞれ連結し，各テキストと質問文をペアとして扱う．テキストに正解セルを含むペアを正例とし，それ以外のすべてのペアを負例として扱い，対照学習を実施する．

3.3　単位抽出

有価証券報告書の表中に含まれる数値は，“千円” や “百株” などの単位が省略されている場合が多い．例えば，企業の売上高が単に “530” と記載されていても，文脈などから，それが “530 千円” や “530 億円” を意味していると推定される場合が多い．

ここでは，質問文とクリーニング後の表を付録 B に詳述したプロンプトで GPT-4o mini に入力し，回答に必要な単位情報を取得する．

3.4　回答形式への正規化

最後に，セルから抽出された数値情報と単位を組み合わせて正規化し，TQA タスクの回答形式に沿う数値表現へと変換する．

単位から正規化するための数値を算出し（例：“千円” → 1,000 を乗算），ルールベースにより単位を反映する．例えば，単位が “千円” と判定された場合，セル値 “530” 千円を “530000” 円に補正する．

4　実験設定

4.1　評価用データセット

本研究では，有価証券報告書をベースとした表形式 QA タスクである TQA データセットを用いて，提案手法の評価を行う [6].

本データセットは，訓練データ（10,300 件），検証データ（1,441 件），テストデータ（2,898 件）から構成され，各データには表を含む文書ファイル，表に紐づく質問，および回答に必要とされる情報が含まれるセルの ID と正答が含まれる．

訓練データセット 10,300 件のうち，900 件のみを言語モデルの学習用データとして，100 件を学習時の検証データとして 3.2.2 節に記載の学習に用いる．オリジナルの検証データは，ハイブリッド検索におけるハイパーパラメータ調整に用いる．

4.2　評価指標

木村ら [6] に倣い，セル検索の正解率（セル検索性能）および出力回答の正解率（回答出力性能）を評価する．いずれの評価においても，正解判定は参照解との完全一致に基づく．本評価では，正規化後の値のうち数値部分のみを判定対象とし，“百万円” のような単位表現は評価対象から除外する．

4.3　評価モデル・人手評価方法

本研究では，日本語に対応する 6 種類の言語モデル（Sentence-BERT [16]，multi-lingual E5-base / large [17]，Ruri-v3-310m [9]，BGE-m3 [18]，および GLuCoSE-base [19]）を評価する．用いた言語モデルの詳細は付録 C に記載する．また，ベースラインとして GPT-4o mini 単体で回答を出力させた場合（プロンプトは付録 B に詳述）と，TF-IDF のみを検索に用いた場合においても評価する．

さらに，人手による検索結果と比較するために，財務知識を持たない日本語を母語とするアノテータにより評価データセットでどの程度の正解率を示すかを測定する．人手評価方法は付録 D に記載する．

5　実験結果・考察

各手法のセル検索性能を図 3 に，回答出力性能の結果を図 4 に示す．

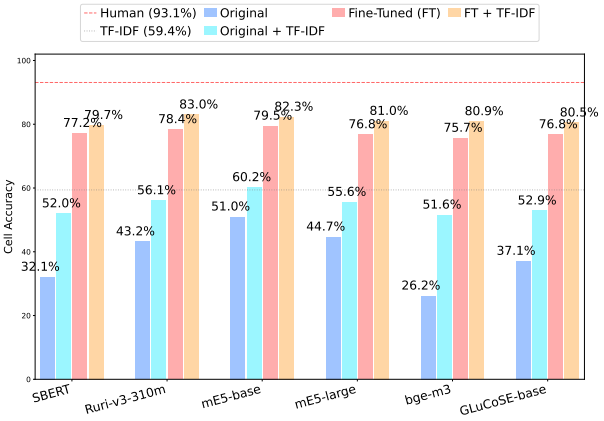


図 3　セル検索性能の評価実験結果

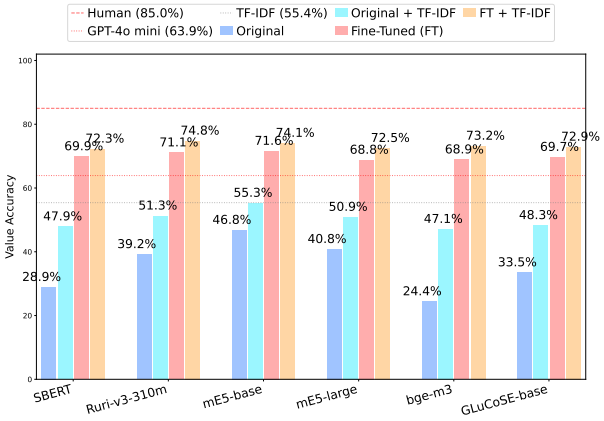


図 4　回答出力性能の評価実験結果

オリジナルの言語モデルはいずれも，TF-IDF ベースのセル検索性能（59.4%）と比較して，低い精度を示した (Sentence-BERT: 32.1%，Ruri-v3-310m: 43.2%，E5-base: 51.0%，E5-large: 44.7%，BGE-m3: 26.2%，GLuCuSE-base: 37.1%)．一般的な検索データセットで高い性能を示す Ruri や E5 などの言語モデルと比べて，単純な表層一致に基づく手法が高い性能を示したことから，本評価データセットにおいては，質問と同一の語句を含むセルを特定する能力が重要であると示唆された．

一方，GPT-4o mini は，回答出力評価において TF-IDF（55.4%）を上回る性能である 63.9%を達成した．しかし，GPT-4o mini は数値や日付の抽出は比較的得意であるが，表中の正しいセル位置の特定に失敗することが多いことが分かった．例えば，会計年度や項目名を取り違えて表の中から誤った行や列を選択しやすいという傾向が確認された．

これに対して，対照学習した言語モデルを用いたベクトル検索手法は，いずれのモデルにおいてもセル検索性能において 75%以上を達成し，特に Ruri-v3-310m では 78.4%，E5-base では 79.5%となった．これは GPT-4o mini の精度（63.9%）を上回る性能であった．本結果は，有価証券報告書の表データから得られた有価証券報告書の表データをセルや行/列ごとにモデルが学習したことで，表層的な表現が類似したセル間でも識別が可能になったためと考えられる．

ハイブリッド検索手法は，いずれのモデルでもセル検索性能において 80%前後，回答出力性能で 72-75%を算出した．最も性能が高かった Ruri-v3-310m では，セル検索性能において 83.0%，回答出力性能で 74.8%であった．この性能向上は，意味的類似度と表層的な類似度を統合することで，安定したセル検索が実現されたことに起因すると考えられる．

人手による評価の結果，財務知識を持たないアノテータでも，表から必要な情報を抽出する正解率は 93.1%であった．この人手正解率と比較すると，提案手法は約 10%低い精度にとどまっている．本結果は，複雑な表データから，人間が構造的に理解できる情報を抽出できない提案手法の課題を示唆した．

6　おわりに

本研究では，有価証券報告書を対象とした表データに関する質問応答タスク (Table Question Answering; TQA) のために，複雑な表構造を考慮した手法を提案し，その有効性を検証した．提案手法は，(1) 表の前処理，(2)，提案手法による表検索，(3)LLM による単位抽出，(4) 回答出力からなる．表検索には，日本語に対応した 6 種類の言語モデルと TF-IDF を組み合わせたハイブリッド検索を採用した．さらに，1,000 件の TQA データから検索用の言語モデルを学習するために訓練データを構築し，モデルを対照学習することで，検索性能の向上を図った．

評価の結果，TQA データセットにおいて，Ruri-v3-310m に本手法を適用することでセル検索性能で 83.0%を達成し，GPT-4o mini を上回り最も高い性能を示した．

一方，財務知識を持たない人手アノテータであっても，セル検索性能において 93%を超える精度であった．この非専門家の精度と比較した結果，提案手法はいずれも 10%以上の精度差が確認された．本結果により，言語モデル単体では，複雑な表データから人間が構造的に理解できる情報を抽出することが難しいケースが多いと示された．

謝辞

本研究の一部は，NTCIR-18 における U4 shared task [6] への参加を通じて得られた成果に基づくものです.

参考文献

[1] Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 8289–8300, 2024.

[2] Zifeng Wang and Jimeng Sun. Promptehr: Conditional electronic healthcare records generation with prompt learning. In **Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing**, Vol. 2022, pp. 2873–2885, 2022.

[3] Weizheng Lu, Jing Zhang, Ju Fan, Zihao Fu, Yueguo Chen, and Xiaoyong Du. Large language model for table processing: a survey. **Frontiers of Computer Science**, Vol. 19, No. 2, 2025.

[4] Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Prediction, generation, and understanding-a survey. **Transactions on Machine Learning Research**.

[5] Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avirup Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. Capturing row and column semantics in transformer based question answering over tables. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1212–1224, 2021.

[6] Yasutomo Kimura, Eisaku Sato, Kazuma Kadowaki, and Hokuto Ototake. Overview of the ntcir-18 u4 task. In **Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies**, Vol. 6, p. 2025, 2025.

[7] Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. Uncovering limitations of large language models in information seeking from tables. In **Findings of the Association for Computational Linguistics ACL 2024**, pp. 1388–1409, 2024.

[8] Qianlong Li, Chen Huang, Shuai Li, Yuanxin Xiang, Deng Xiong, and Wenqiang Lei. Graphotter: Evolving llm-based graph reasoning for complex table question answering. In **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 5486–5506, 2025.

[9] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese General Text Embeddings. **arXiv preprint arXiv:2409.07737**, 2024.

[10] Jing Fang, Prasenjit Mitra, Zhi Tang, and C Lee Giles. Table header detection and classification. In **Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence**, pp. 599–605, 2012.

[11] Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. Tables as semi-structured knowledge for question answering. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 474–483, 2016.

[12] Nengzheng Jin, Dongfang Li, Junying Chen, Joanna Siebert, and Qingcai Chen. Enhancing open-domain table question answering via syntax-and structure-aware dense retrieval. In **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 157–165, 2023.

[13] Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. Table cell search for question answering. In **Proceedings of the 25th International Conference on World Wide Web**, pp. 771–782, 2016.

[14] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8413–8426, 2020.

[15] 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**, pp. 4320–4333, 2020.

[16] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, 2019.

[17] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.

[18] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In **Findings of the Association for Computational Linguistics ACL 2024**, pp. 2318–2335, 2024.

[19] Yotarow Watanabe Akihiko Fukuchi, Yuichiro Hoshino. pkshatech/glucose-base-ja, 2023.

A　提案手法の前処理方法

本章では提案手法の前処理の詳細を説明する．

表の外部に存在するテキスト文書や，表そのものと無関係な注釈情報をすべて除去する．次に，“colspan” および “rowspan” を除くすべての HTML 属性（“style”，“class”，“id” など）を削除する．最後に，“td” タグ内部に存在する追加の HTML タグをすべて削除し，内容をプレーンテキストに変換する．なお，一部のケースでは，単一のセル内に入れ子構造の表が含まれることがある．本来は，このような入れ子構造も扱うことが望ましいが，本研究では実装上の複雑さを回避するため，入れ子の表は 1 つのセルに統合して処理する．

B　入力プロンプト

TQA データセットにおいて，表から質問に対応する正しいセル ID を検出するために，以下のプロンプトを GPT-4o mini に入力する．

プロンプト

以下の表の中で、以下の質問の答えが書かれているセルの “cell_id” を教えてください。“cell_id” の値だけを答えてください。
質問: {question}
表: {table}

また，表から質問に対応する値を抽出するために，以下のシステムプロンプトとプロンプトを GPT-4o mini に入力する．

システムプロンプト

情報を見つけるのに役立つ AI アシスタントです。

プロンプト

以下の表に基づいて、質問に答えてください。答えの値とその単位を
{{ "value": "値", "unit": "単位" }}
という JSON 形式で出力してください。出力のみを答えてください。
質問：{question}
表：{table}

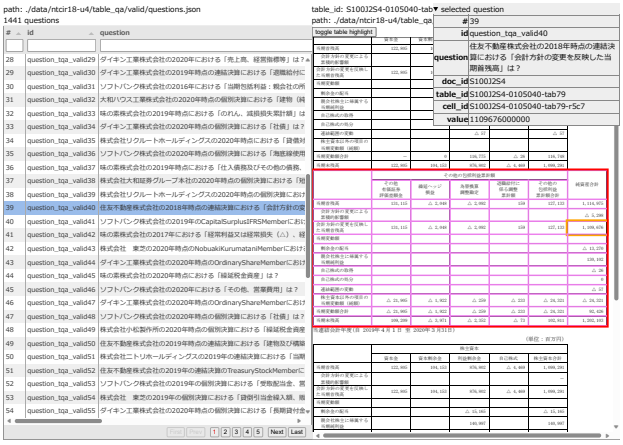


図 5　人手アノテーションの GUI 画面例

C　評価対象モデルの一覧

以下に評価対象モデルを一覧する．

- Sentence-BERT: https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2
- Ruri-v3-310m: https://huggingface.co/cl-nagoya/ruri-v3-310m
- multilingual E5-base: https://huggingface.co/intfloat/multilingual-e5-base
- multilingual E5-large: https://huggingface.co/intfloat/multilingual-e5-large
- bge-m3: https://huggingface.co/BAAI/bge-m3
- GLuCoSE-base: https://huggingface.co/pkshatech/GLuCoSE-base-ja-v2

D　人手評価方法

本研究では，TQA データセットを人手で評価する．アノテータは，財務知識を持たない日本語母語話者 1 人である．

このアノテーション作業を支援するため，与えられた質問に対応する表中のセル ID を作業者が容易に特定可能な GUI を開発した．本インターフェースの例を図 5 に示す．

GUI 左側のリストから質問を選択すると，対応する表が画面右側に表示される．ここで，検証データでは正解の抽出対象のセルはオレンジ色でハイライト表示するが，テストデータでの人手評価においては，この機能はオフにする．また，マウスカーソルをセル上に合わせると，該当セルの ID がカーソル付近に表示される．本機能により，アノテータが目的の情報が含まれるセル ID を容易かつ正確に特定することを可能とする．