農業経営データに対する Multi-table QA ベンチマークの構築

板倉亮真 [1]　中嶋楓花 [1]　坂地泰紀 [1]　小林暁雄 [2]　馬場研太 [2]　大友将宏 [2]
石原潤一 [2]　桂樹哲雄 [2]　高柳剛弘 [3,4]　野田五十樹 [1]
[1] 北海道大学　[2] 農研機構 農業情報研究センター　[3] 東京大学　[4] 株式会社 Simulacra
{itakura.ryoma.x2,nakajima.fuka.h0}@elms.hokudai.ac.jp sakaji@ist.hokudai.ac.jp
{akio.kobayashi,baba.kenta285,masahiro.otomo,ishiharaj612,t.katsuragi}@naro.go.jp
takayanagi@simulacra.co.jp　i.noda@ist.hokudai.ac.jp

概要

日本の農家の標準的な経営モデルである農業経営類型は，農業経営のマニュアルとして全国的に整備されており，その多くは自然言語データを含む表形式データとして提供されている．しかし，農業経営類型は機械可読性を前提として作成されていないため，正規化が不十分で構造が複雑である場合や，暗黙的前提を含む場合が多い．本研究では，農業分野における非正規化複数表データに基づく Multi-table QA により，農業経営の文脈における LLM の情報統合能力および推論性能を評価する．実験の結果，GPT-5 は計算能力において大幅な改善を示したものの，農業の文脈を前提とした情報の解釈においては依然として課題があることを明らかにした．

1　はじめに

近年，GPT-3 や GPT-4 といった大規模言語モデル（LLM）の発展により，多様な知的作業を自動化できる可能性が広がっている [1, 2]．農業経営では，普及指導員が経営指標や技術体系など複数の資料を参照・統合し，個別の農家に対して助言を行う．この過程を LLM で支援できれば，農業従事者の減少 [3] に伴う人手不足の解消に寄与しうる．

農業経営に関する知識や指針は，標準的な経営モデルとして全国的に整備・公開されており，その多くは自然言語を含む表形式データとして提供されている．これらの表は機械可読性を前提として作成されていない場合が多く，正規化が不十分であり，セル結合を含む複雑な構造を持つ．さらに，農業経営の意思決定では，収支，作業体系，労働時間など，異なる観点で整理された複数の表を参照し，それらを統合して判断する必要がある．しかし，これらの表およびクエリには農業経営の文脈に依存した暗黙的前提が多く含まれるため，単純な情報抽出だけでは適切な回答を導くことが難しい．したがって，農業経営の意思決定支援には，複雑な複数表データを横断的に理解・統合し，農業経営の文脈や暗黙的前提を踏まえて回答を生成する能力，すなわち高度な Multi-table QA（Question Answering）が不可欠である．

そこで本研究では，農業経営に関する複数の表形式データを対象に，既存 LLM の Multi-table QA における情報統合能力および推論性能を評価する．本研究の主な貢献は以下の通りである．

- 農業経営類型を基にした暗黙的前提を扱う Multi-table QA データセットを構築し，公開する．
- 既存 LLM を対象に，農業経営の文脈を含む Multi-table QA 性能を評価する．

2　関連研究

農業分野における自然言語処理は，学習用データ資源の不足が指摘されている．杉山ら [4] は農業経営類型資料に含まれる表の構造解析が困難である点を指摘し，課題を整理している．また板倉ら [5] は，農業経営類型の表を人手とツールの併用によってパースし，インストラクション形式のデータセットを構築した．

Table QA の領域では，自然言語の質問から SQL などのクエリを生成し，実行結果に基づいて回答するアプローチが広く用いられている [6, 7]．特に Multi-table QA においては，複数表を対象とし，JOIN や集合演算などの操作を伴う問い合わせに対応する必要性が指摘されている [8]．Wu ら [9] は，Multi-table QA を対象として，RAG（Retrieval-Augmented Generation）の評価フレームワークを提案した．ま

た，表とテキストのハイブリッド情報源を跨ぐ質問応答や，数値推論を対象とするベンチマークも提案されている [10, 11, 12].

既存の Table QA 研究の多くは，行列構造が明確な正規化表を前提としている．一方，本研究はセル結合を含む非正規化表を対象とし，表間文脈の解釈や暗黙的前提の推定を伴う指標算出を評価する点で異なる．

3　農業経営類型データ

本研究で対象とする農業経営類型データについて述べる．農業経営類型とは，地域・品目・経営形態ごとの標準的な農業経営モデルを記述した資料であり，普及指導員等による営農指導の基礎資料として利用され，個別の経営体に対する計画立案や改善提案の参照情報となる．

農業経営類型に含まれる表は，人間による閲覧・運用を前提として設計されている．したがって，セル結合やキー配置の不定性などにより，単純な行列としての解釈が困難である．加えて，多くが PDF で提供されることから，機械処理が困難である．具体例として，Web 上で公開されている長崎県の経営類型資料である長崎県農林業基準技術[1] に含まれる表の一部を図 1 に示す．図 1 より，セル内での複数値の併記，キーと値の不定配置，およびセル結合を含む複雑な表構造が確認できる．

経営類型	家族労働力	品目・栽培型及び規模		経営・技術の特徴
個別経営 I	人 2		a	(1)基盤整備地区における個別経営による営農 (2)作業の一部は委託
		水稲	400	
		小麦(長崎W2号)	500	
		二条大麦	500	
		大豆	600	
		合計	2000	
		経営耕地面積　水田10ha　(自作地4ha,借入地6ha)		
経営目標	1 農業総収入	21,975 千円	4 1日当たり農業所得	44,053 円
	2 農業経営費	14,540 千円	5 1人当たり年間労働時間	675 時間
	3 農業所得	7,435 千円		

図 1　技術体系の表構造の例；長崎県農林業基準技術「個別経営 I」から抜粋

4　評価データセット構築

農業分野における非正規化された複数表データに対する既存 LLM の性能を評価するため，Multi-table QA データセットを構築した．本データセットは JSON 形式で整理され，実験結果とともに CC BY-SA 4.0 ライセンスのもと GitHub[2] 上で公開している．

4.1　タスク設計

本研究では，農業経営類型に含まれる複数表を入力とし，指定された経営指標を算出する Multi-table QA タスクを設計する．入力は単一の経営類型に含まれる全表の集合であり，各表は HTML 形式で提示する．これは，セル結合を含む表構造の複雑さを保持しつつ，PDF のパース困難性を評価対象から分離することを目的とする．モデルには付録の表 3 に示す 12 指標のうち 1 つを指定し，関連する表から必要な値を抽出・統合した上で，指標値を算出して回答させる．指標は，農林水産省「営農類型別経営統計」[3] で用いられる指標のうち，本研究で用いる資料から算出可能なものを採用した．出力は算出結果に加えて，根拠となる抽出値および計算過程を含む形式とする．なお，これらの指標には参照表数が 1 のもの（単一表から数値抽出可能な問い）も含まれる．しかし，本データでは解答対象（経営全体／10a 当たり），単位，項目同定（同義項目の対応付け），按分値と取得価額の区別などの解釈において，他表の記述が前提情報として機能する場合がある．このように複数表からなる資料全体を入力とし，表間文脈の利用を許容・要求する設定を本研究では Multi-table QA と定義する．

4.2　対象データとサンプリング

本研究では，インターネット上で農業経営類型が公開されている，長崎県・鳥取県・山口県の 3 県を対象とし，各県から栽培作物に水稲を含む 1 類型をサンプリングした．本選定は，自治体間で表構造や記述形式が異なる状況におけるモデル挙動の差を検証することを目的とする．そのため，表と暗黙的な農業経営の前提のみから指標が導出できる，かつ同じ作目を含む類型に限定することで，外部知識の補完を必要としない条件での評価を可能にした．

4.3　QA データセット構築

質問生成では，システムプロンプトとして「あなたは農業経営の専門家です.」を設定し，ユーザープロンプトで (i) 指標の定義確認，(ii) 表からの値の抽出，(iii) 計算手順の提示を指示した．その後，HTML 形式の表を提示した．各表のタイトルは<h1>タグで記述し，対応する表の直前に配置した．表データ

1）https://www.pref.nagasaki.jp/bunrui/shigoto-sangyo/nogyo/nouringyoukijyungijyutu/681419.html
2）https://github.com/itakuraryoma/agri-ruikei-mtqa
3）https://www.maff.go.jp/j/tokei/kouhyou/noukei/einou_kobetu/gaiyou/index.html

は，PDF 中の表を複雑な構造を保ったまま XLSX 形式に人手で転記した後，スクリプトにより HTML 形式へ変換して作成した．

解答データは，各経営類型について表 3 に示す各指標の算出に必要な値を人手で特定し，指標定義に従って正解値を算出した．さらに，各問に対して正解値に加え，参照した値と演算内容からなる計算式を付与した．これにより，モデル出力の根拠（抽出値・計算過程）との照合が可能な解答データを整備した．質問データおよび解答データの具体例は，それぞれ付録 B，C に示す．

5 評価実験

5.1 実験設定

本研究では，構築したデータセットを用いて，OpenAI 社の GPT-4o-mini[4)]，GPT-4o[5)]，GPT-5[6)] の 3 つの LLM を評価する．

GPT-4o-mini および GPT-4o は温度=0.0，top-p=1.0 として各問 1 回生成し，この設定を single run（**single**）と呼ぶ．一方，GPT-5 は仕様上，温度が 1.0 に固定されており出力に揺らぎが生じるため，3 つのシード値（0, 42, 777）で各問 3 回生成した．その後，3 回の出力のうち最頻の回答を最終出力とする majority voting（**majority(3)**）を適用し，同率の場合は seed=0 の回答を採用した．また，推論コストを揃えた比較のため，GPT-5 については seed=0 による single run の結果も併記する．

評価は生成回答が正解と一致するかを人手で判定した．数値については単位の表記揺れ（例：円と千円）は実質的に同値であれば正解とし，有効数字の違いは回答の有効桁より一つ小さい位の四捨五入が正しければ正解とした．誤答については以下の 3 種のいずれか一つに分類した．

- **定義誤り**: 指標の定義や計算式，抽出すべき値の解釈が誤っている．
- **計算誤り**: 抽出した値は適切だが，四則演算や単位変換が誤っている．
- **前提解釈誤り**: 表の構造的前提や農業経営の文脈を誤解釈し，不適切な値を用いている．

4） gpt-4o-mini-2024-07-18
5） gpt-4o-2024-08-06
6） gpt-5-2025-08-07

5.2 実験結果

各モデルの正解率（Accuracy）を表 1 に示す．全体の結果として，GPT-5（single）が最も高い正解率（0.611）を示し，次いで GPT-5（majority(3)）（0.583），GPT-4o（0.389），GPT-4o-mini（0.222）の順となった．県別では，GPT-5（majority(3)）が長崎県のデータに対して 1.000 と最も高い正解率を示した一方で，鳥取県および山口県のデータに対しては，すべてのモデルで正解率が低下する傾向が見られた．

表 1 モデルおよび県ごとの正解率（instance-level，全 36 問）

モデル	全体	長崎	鳥取	山口
GPT-4o-mini single	0.222	0.333	0.083	0.250
GPT-4o single	0.389	0.750	0.250	0.167
GPT-5 single	**0.611**	0.917	**0.417**	**0.500**
GPT-5 majority(3)	0.583	**1.000**	**0.417**	0.333

次に，誤答に対するエラータイプの発生割合を表 2 に示す．ここではモデルの出力傾向を詳細に分析する目的から，GPT-5 については single（seed=0，全 36 試行）の設定に加えて，3 回の生成結果の合算（全 108 試行）を対象とした誤り分類も行った（**all seeds**）．GPT-4o-mini および GPT-4o では，前提解釈誤りが最も多く，表の暗黙的なコンテキストを踏まえた解釈に課題があることが示された．一方，GPT-5（all seeds）では定義誤りが 0.596 と最も多く，指標の意図や定義の取り違えが誤答の主要因であることが明らかとなった．計算誤りについては GPT-5（all seeds）においても 3 件発生しており，完全には解消されていない．

6 考察

6.1 誤り傾向の分析

エラー分析（表 2）から，GPT-4o-mini，GPT-4o，GPT-5 の順にモデルの性能向上に伴い計算誤りの割合は減少するものの，完全には解消されていないことがわかる．したがって，農業経営支援のように正確な収支計算や指標算出が求められる場面では，LLM 単体の出力に依存せず，計算の実行には外部ツールを用いることが望ましい．

一方，モデルの性能向上に伴い計算および前提解釈誤りが減少するのに対し，定義誤りの割合は増加している（表 2）．これは，モデルの汎用的な推論性

表 2　モデルごとの誤答数に対するエラータイプ別発生割合（小数第 4 位を四捨五入，括弧内は件数）

モデル	定義誤り	計算誤り	前提解釈誤り	誤答数	全試行数
GPT-4o-mini	0.071 (2)	0.179 (5)	**0.750 (21)**	28	36
GPT-4o	0.273 (6)	0.182 (4)	**0.545 (12)**	22	36
GPT-5 single	**0.571 (8)**	0.000 (0)	0.429 (6)	14	36
GPT-5 all seeds	**0.596 (28)**	0.064 (3)	0.340 (16)	47	108

能が向上したことで表中の暗黙的前提を補完しようとする過程において，指標定義を一般的な経営の文脈で解釈してしまう傾向が強まったためと考えられる．その結果，表面的には整合的な推論が行われているものの，農業経営の固有文脈とは整合しない出力に至る事例が増加した．このことから，現在の LLM は農業経営の文脈に依存する指標定義や暗黙的前提を十分に扱えていないといえる．

県別の傾向として，鳥取県および山口県のデータでは，すべてのモデルで正解率が低下した（表 1）．この要因として，両県の資料では指標の算出において暗黙的前提の推定がより多く求められることが考えられる．

以上より，既存 LLM には，計算の頑健性，農業経営の文脈に従った指標定義の解釈，および暗黙的前提の推定に課題があることが示された．

6.2　具体的な誤りの例

回答と同時に出力された計算手順に基づき，最も性能の良かった GPT-5 の誤りを整理する．鳥取県および山口県の両資料には固定資産額として「今作按分」の値が掲載されているが，本タスクで算出対象とするのは経営全体の指標であるため，参照すべきは取得価格である．しかし，一部の出力では，按分値を固定資産額として用いる誤りが観察された．また，山口県では，「構成員還元額」が掲載されているものの，これは農業所得（利潤）とは異なるため，所得の算出には利潤を用いる必要がある．さらに，同県では，「家族労働」を「基幹労働＋補助労働」とみなす誤解も見られたが，補助労働の労働報酬が別途計上されていることから，本タスクの定義では基幹労働のみを参照するのが適切である．鳥取県では，経営耕地面積が表に明示されないため，10a 当たりの指標を算出してしまう誤りが観察された．しかし，表中に「1200a 当たり」の指標が提示されており，これを根拠として経営全体の耕地面積を 1200a と推定する必要がある．

以上の結果より，モデルの性能向上に伴い誤りの総数は減少傾向にあり，自明な誤りは解消されつつあることが確認された．しかし，上述のような文脈依存性の高い誤りが依然として残存しており，これらは LLM が農業経営支援において実用化されるための重要な課題であると言える．

7　結論

本研究では，農業経営支援への LLM 活用を見据え，非正規化複数表データを基にした Multi-table QA ベンチマークを構築し，公開した．本データセットは，セル結合を含む実世界の複雑な表構造，および農業経営に固有の暗黙的文脈を保持しており，従来のデータセットでは評価しにくかった実践的な課題に対する評価基盤を提供する．

評価実験の結果，GPT-5 は先行モデルと比較して計算能力の改善を示した一方で，計算誤りは依然として残存した．また，農業経営の文脈に依存する暗黙的前提の解釈や，指標定義に沿った値の抽出・統合には課題が残ることを示した．これらの知見は農業分野に限らず，医療，行政，製造業など，正規化されていないレガシーデータを扱う多様な分野における LLM 応用に対しても示唆を与える．

今後の展望として，他の作物や自治体のデータを追加し，データセットの多様性と網羅性を高める．また，実運用を見据えて農業経営に関するテキスト資料を収集・整備し，農業経営の文脈に適応した学習（例：追加学習やドメイン適応）を通じて，暗黙的前提の解釈や定義の取り違えを低減することを検討する．

謝辞

本研究は，内閣府研究開発と Society5.0 との橋渡しプログラム（BRIDGE）「AI 農業社会実装プロジェクト」JP23836805 の補助を受けて行った．

参考文献

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In **Advances in Neural Information Processing Systems (NeurIPS)**, Vol. 33, pp. 1877–1901, 2020.

[2] OpenAI. Gpt-4 technical report. https://arxiv.org/abs/2303.08774, 2023. Accessed: 2026-01-07.

[3] 農林水産省. 令和 3 年度食料・農業・農村白書. https://www.maff.go.jp/j/wpaper/w_maff/r3/r3_h/trend/part1/chap1/c1_1_01.html, 2020. Accessed: 2026-01-07.

[4] 杉山陽菜乃, 阿部瑞稀, 中村彩乃, 前多陸玖, 坂口遥哉, 佐藤栄作, 木村泰知, 小林暁雄, 大友将宏, 石原潤一, 桂樹哲雄, 川村隆浩. 農林業基準技術に含まれる表を対象とした pdf から csv へ変換する際の課題分析. 言語処理学会第 31 回年次大会 発表論文集, 2025.

[5] 板倉亮真, 坂地泰紀, 野田五十樹, 小林暁雄, 大友将宏, 石原潤一, 桂樹哲雄. 生成 ai のための農業データセット構築とモデル評価. 言語処理学会第 31 回年次大会 発表論文集, 2025.

[6] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[7] Tao Yu, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex, cross-domain semantic parsing, text-to-sql task. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3911–3921. Association for Computational Linguistics, 2018.

[8] Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. Multitabqa: Generating tabular answers for multi-table question answering. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 6322–6334, Toronto, Canada, 2023. Association for Computational Linguistics.

[9] Jian Wu, Linyi Yang, Dongyuan Li, Yuliang Ji, Manabu Okumura, and Yue Zhang. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In **The Thirteenth International Conference on Learning Representations**, 2025.

[10] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular, textual data. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1026–1036, Online, November 2020. Association for Computational Linguistics.

[11] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 3697–3711, Online, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular, textual content in finance. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3277–3287, Online, August 2021. Association for Computational Linguistics.

A　QA 構築に用いた経営指標の定義

表 3　評価実験に用いた 12 の経営指標とその定義および計算式

経営指標	定義・式	単位
農業粗収益	農業経営によって得られた総収益額	円
農業経営費	農業経営に要した一切の経費	円
農業所得	農業粗収益 − 農業経営費	円
農業所得率	$\frac{\text{農業所得}}{\text{農業粗収益}} \times 100$	%
家族農業労働 1 時間当たり農業所得	$\frac{\text{農業所得}}{\text{自営（家族）農業労働時間}}$	円/時間
経営耕地面積 10a 当たり農業所得	$\frac{\text{農業所得}}{\text{経営耕地面積 } (a)} \times 10$	円/10a
農業固定資産装備率 (自営農業労働時間あたり)	$\frac{\text{農業固定資産額（土地除く）}}{\text{自営農業労働時間}}$	円/時間
農機具資産比率	$\frac{\text{自動車・農機具の固定資産額}}{\text{農業固定資産額（土地除く）}} \times 100$	%
農業固定資産回転率	$\frac{\text{農業粗収益}}{\text{農業固定資産額（土地除く）}}$	回
経営耕地面積 10a 当たり自営農業労働時間	$\frac{\text{自営農業労働時間}}{\text{経営耕地面積 } (a)} \times 10$	時間/10a
経営耕地面積 10a 当たり農業固定資産額	$\frac{\text{農業固定資産額（土地除く）}}{\text{経営耕地面積 } (a)} \times 10$	円/10a
経営に占める償却費率	$\frac{\text{年間償却額合計}}{\text{農業経営費}} \times 100$	%

B　質問データの例

[System]
あなたは農業経営の専門家です。
[User]
以下の複数表からなる経営類型について、指標「農業粗収益」を求めよ。
まず「農業粗収益」の定義（算出式・必要項目・単位）を確認せよ。
表から必要な値を抽出し、必要に応じて計算せよ。
ただし、計算が必要な指標であっても表中に「農業粗収益」の計算済みの値が記載されている場合は、必ずその値を抽出して回答し、再計算しないこと。
思考の過程は、使用した表の値（どの項目を使ったか）と計算手順を簡潔に示すこと。

<h1>作業別・旬別労働時間（10a 当たり時間）大豆 6ha</h1>
<table>
<tr>
<td rowspan=\"2\">品目・作業/月・旬</td>
（以降 HTML 形式の表データが続く）

C　解答データの例

```
"農業粗収益": {"answer": 21975000, "formula": ""}
"農業経営費": {"answer": 14540000, "formula": ""}
"農業所得": {"answer": 7435000, "formula": ""},
"農業所得率": {"answer": 0.338339022, "formula": "7435000 / 21975000"}
"家族農業労働 1 時間当たり農業所得": {"answer": 5778.796829, "formula": "7435000 / 1286.6"}
```