# 生成 AI のための農業データセット構築とモデル評価

板倉亮真 [1]　坂地泰紀 [1]　野田五十樹 [1]
小林暁雄 [2]　大友将宏 [2]　石原潤一 [2]　桂樹哲雄 [2]
[1] 北海道大学　　[2] 農業・食品産業技術総合研究機構
itakura.ryoma.x2@elms.hokudai.ac.jp　{sakaji,i.noda}@ist.hokudai.ac.jp
{akio.kobayashi,masahiro.otomo,ishiharaj612,t.katsuragi}@naro.go.jp

## 概要

農業の栽培技術や取り組みは、国内の各自治体や公共団体、農協、その他の民間企業などがそれぞれ独自に取りまとめたデータがそれぞれのサイト上などから公開されている。しかしながら、これらの知識は機械可読な形式で提供されていることは稀で、現状の生成 AI などでうまく活用できているとは言い難い。本研究では、そのようなデータから、AI で利用しやすい形のデータセットとして、インストラクション形式でのデータの半自動構築を行う手法を提案する。また、既存の生成 AI がこれらのウェブ上のデータに含まれる知識をどの程度学習しているかについても検証を行った。

## 1　はじめに

農業は、食糧生産を通じて国民の生活基盤を支えるのみならず、経済の安定や国家の持続的成長に寄与する重要な産業である。しかし、近年の日本では、社会全体の高齢化や人口減少が進行する中で、農業従事者の減少とそれに伴う労働力不足が深刻な課題となっている。このような背景を受けて、内閣府が推進する「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」[1)]の一環として、農林水産省は「AI 農業社会実装プロジェクト」[2)]を実施し、農業分野に AI 技術を実装することで労働力不足を補う施策を展開している。

農業を行う際には、地域ごとに異なる土壌や気候といった環境条件や、作物の品種に応じて栽培方法が大きく異なる。本プロジェクトでは、日本における多様な農作物品種や地域特性に適応した農業用言語資源が十分に整備されていない点が重要な課題として指摘されている。また、それに伴って、既存の大規模言語モデル（LLM）が日本の農業に関する知識をどの程度有しているかについては、不明な部分が多いのが現状である。これらの課題を解決するために、本研究では日本の地域の一例として長崎県を取り上げ、以下の 2 つの貢献を行った。

- 長崎県の農業経営類型を用いて、半自動的にインストラクションチューニングデータを生成した。
- 作成したデータセットを基に、既存の生成 AI モデルが長崎県の農業に関する知識をどの程度保持しているかを評価した。

## 2　関連研究

近年の研究では、大規模言語処理モデル（LLM）の農業分野への応用可能性が注目されている。Peng ら [1] は、ドメインに依存しない一般的な事前トレーニング済み LLM を用いて、農業に関する文書から有用なデータを自動あるいは半自動的に抽出する方法を示している。Kästing と Hänig[2] は農業に関する文書から生成 AI を用いてデータセットを作成し、Retrieval-Augmented Generation(RAG) システムのプロトタイプを開発している。この論文の中ではシステムを評価する際に、191 個のデータからなるドイツ語の農業 Q&A データセット (Dataset Card for BVL Q&A Corpus 2024) を用いている。Kuska ら [3] は、文書作成、コンサルティング、病虫害管理といった、農業における LLM の応用可能性について議論している。Balaguer ら [4] は、RAG とファインチューニングを比較しながら、ドメイン固有の知識を LLM に組み込むための手法を提案している。これらの研究では、農業へ LLM を応用することの有用性を示すと同時に、LLM が農業ドメイン固有の知識を得る必要性を認めている。小林ら [5] は、生

1）https://www8.cao.go.jp/cstp/bridge/index.html
2）https://www8.cao.go.jp/cstp/bridge/keikaku/r5-20_bridge_r6.pdf

成 AI に農業ドメインの知識を与えるために、農産物の市場動向に関する数値データとテキストデータに加え、天候データを組み合わせたデータセットを構築する手法を提案している。

本研究では、生成 AI を利用せずに農業関連文書から 2052 個のデータを収集し、データセットを構築した。このアプローチは、生成 AI を利用して構築されたデータが、その利用規約により LLM の学習に使用できない場合があること[3]への対策である。本研究で構築したデータセットは、生成 AI の規約に縛られることなく、自由に学習利用可能なリソースとして設計された。また、本研究では、長崎県の農業という限定的なテーマに焦点を当て、既存の生成 AI モデルがこの分野の知識をどの程度保持しているかを評価した。これにより、地域や品種に大きく依存する農業分野において、生成 AI の応用可能性を明らかにするための基盤を提供する。

# 3 データセット構築

本節では、生成 AI を用いずに農業に関する質問と回答のデータセットを半自動的に構築する方法について述べる。図 1 にデータセット構築の流れを示す。本研究ではデータセットの元となる文書データとして、長崎県の農業経営類型 (PDF 形式) を用いた。この文書中に記載されている「技術体系」の表から、農作業における技術の重要事項をプログラム (3.1 参照) によって自動的に抽出し、抽出された内容を基に質問と回答のペアを作成した。具体例として、表 1 の水稲の「品種の選定」に関する質問と回答のペアを以下に示す。

**質問:** 水稲の品種の選定の重要事項を教えてください。

**回答:** 奨励品種から選定する。毎年種子更新に努める。高温耐性品種を選定**する。**

ここで、上の例のように元の文書内の文章が動詞の終止形と句点で終わる完全な文となっていない場合、例の回答中の太字部分のように、語尾と句点を手作業で補うことで、生成 AI の回答形式に近づけた。また、質問文の多様性を確保するため、1 つの回答に対して 3 種類の質問 Q1、Q2、Q3 を作成した。これらの質問は文末の表現が異なっており、表 1 の「品種の選定」に対する 3 つの質問例を以下に示す。なお、これらの質問に対する回答は共通である。

**Q1:** 水稲の品種の選定の**重要事項を教えてください。**

**Q2:** 水稲の品種の選定**のポイントは？**

**Q3:** 水稲の品種の選定**において何か気をつけることはありますか？**

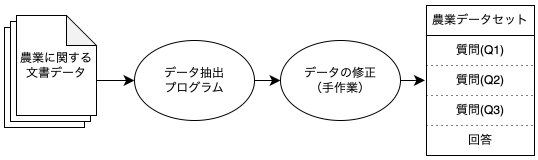


**図 1** データセット構築の流れ

## 3.1 データ抽出プログラム

データの抽出には、Python ライブラリ「PyMuPDF」(バージョン 1.25.1)[4]を使用した。具体的には、文書内のすべての表を対象に、「技術上の重要事項」という文字列を含む表を特定し、その中から「作業の種類」と「技術の重要事項」の列を抽出した。

**表 1** 水稲栽培の技術体系

| 作業の種類 | ... | 技術の重要事項 |
|---|---|---|
| 品種の選定 | ... | 奨励品種から選定する。<br>毎年種子更新に努める。<br>高温耐性品種を選定 |
| ⋮ | ⋮ | ⋮ |

# 4 評価実験

本節では、前節 3 で構築した長崎県の農業経営類型を基にしたデータセットを用い、既存の生成 AI モデルが長崎県の農業に関する知識を十分に有しているかを評価する。具体的には、各モデルに対してデータセット内の質問 Q1 を入力し、生成された回答を候補テキスト（candidate text）、データセットの回答を参照テキスト（reference text）として扱い、BERTScore、ROUGE-1、ROUGE-2、ROUGE-L の 4 つの指標を用いて評価を行った。実験の流れを図 2 に示す。

## 4.1 実験の条件

### 4.1.1 評価対象とする生成 AI モデル

本研究では、以下の 3 つの生成 AI モデルを評価対象とした。

- gpt-3.5-turbo-0125 (以下 gpt-3.5)

3） https://openai.com/ja-JP/policies/terms-of-use/

4） https://pypi.org/project/PyMuPDF/

表 2　OpenAI API の各モデルの特徴

| | |
|---|---|
| GPT-4o | 多用途で高知能なフラッグシップモデル |
| GPT-4o-mini | 集中的なタスクに最適な、高速で手頃な価格の小型モデル |
| GPT-3.5 Turbo | 単純なタスク向けの高速モデル（GPT-4o-mini の下位互換） |

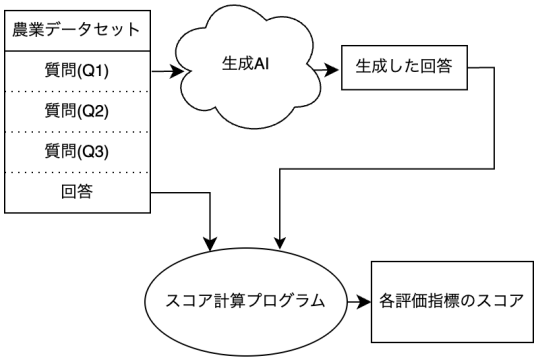


図 2　実験の流れ

- gpt-4o-mini-2024-07-18 (以下 gpt-4o-mini)
- gpt-4o-2024-11-20 (以下 gpt-4o)

OpenAI の Models overview[5] によると、各モデルの特徴は、要約すると表 2 の通りである。ここから、各モデルの性能は gpt-4o、gpt-4o-mini、gpt3.5 の順に高いと予想される。

#### 4.1.2　生成 AI による回答の生成

回答の生成には、OpenAI API を利用した。API リクエスト時の設定として、回答のランダム性を制御する temperature パラメータは 0 に設定した。また、生成 AI が回答する際の立場を決定するために使用する messages パラメータの role:system には、”あなたは農業指導者です。”というプロンプトを指定した。

#### 4.1.3　各評価指標におけるスコアの計算

BERTScore の計算には、Python パッケージ「bert-score」(バージョン 0.3.13)[6] を使用した。評価に使用するモデルには日本語専用のものが存在しないため、「BERT-base-multilingual-cased」を採用した。ROUGE スコアの計算には、Python パッケージ「rouge-score」(バージョン 0.1.2)[7] を使用した。また、形態素解析には MeCab を用い、stemmer を有効にして単語を原形に変換した上で評価を行った。

5）https://platform.openai.com/docs/models
6）https://github.com/Tiiiger/bert_score
7）https://pypi.org/project/rouge-score/

### 4.2　結果

#### 4.2.1　F1 スコア

はじめに、用いた 4 つの評価指標（BERTScore、ROUGE-1、2、L）における、Q1 に対する各生成 AI モデルが生成した回答の F1 スコアを図 3 および表 3 に示す。結果として、BERTScore では 4.1.1 で予想された性能順にスコアが高く、gpt-4o、gpt-4o-mini、gpt-3.5 の順であることが確認された。一方で、ROUGE-1、2、L では予想に反し、gpt-3.5、gpt-4o-mini、gpt-4o の順にスコアが高くなった。

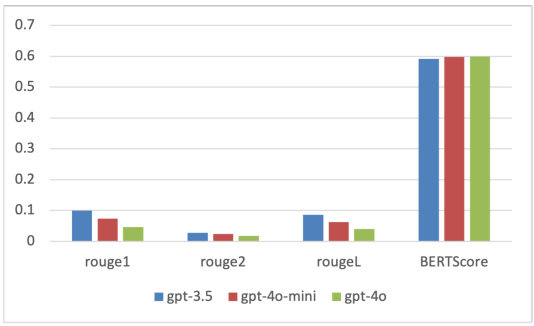


図 3　Q1 での各評価指標における各モデルの F1 スコア

表 3　Q1 での各評価指標における各モデルの F1 スコア

| | rouge1 | rouge2 | rougeL | BERTScore |
|---|---|---|---|---|
| gpt-3.5 | 0.099892 | 0.026790 | 0.086331 | 0.591098 |
| gpt-4o-mini | 0.073146 | 0.023464 | 0.062691 | 0.598158 |
| gpt-4o | 0.046373 | 0.017256 | 0.039284 | 0.599199 |

#### 4.2.2　適合率・再現率

続いて、各評価指標における適合率 (Precision) と再現率 (Recall) について、Q1 に対して各モデルが生成した回答を評価した結果を図 4,、表 4 に示す。この結果から、どの評価指標においても、適合率は 4.1.1 で予想された通り gpt-3.5、gpt-4o-mini、gpt-4o の順にスコアが高くなった一方で、再現率は逆に gpt-4o、gpt-4o-mini、gpt-3.5 の順に高く、予想と逆順になった。

## 5　考察

結論として、本研究において評価対象とした生成 AI モデル（4.1.1 参照）は、長崎県に限定した農業

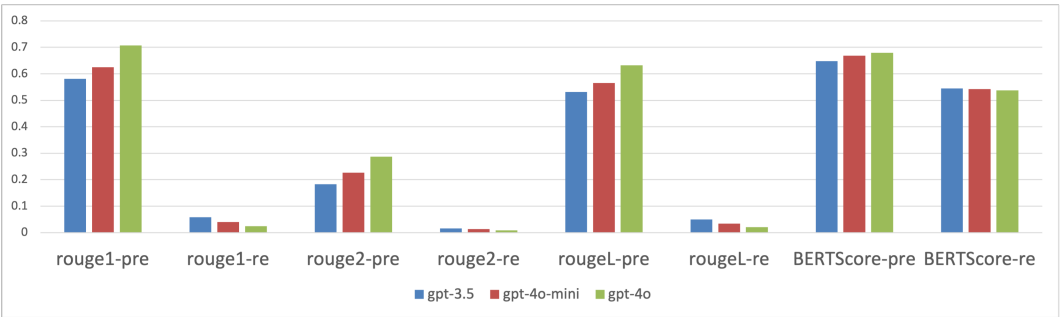


**図 4**　Q1 での ROUGE-1, 2, L における適合率 (pre) 及び再現率 (re)

**表 4**　Q1 での ROUGE-1、2、L における適合率 (pre) 及び再現率 (re)

| | rouge1-pre | rouge1-re | rouge2-pre | rouge2-re | rougeL-pre | rougeL-re | BERTScore-pre | BERTScore-re |
|---|---|---|---|---|---|---|---|---|
| gpt-3.5 | 0.580926 | 0.058056 | 0.182487 | 0.015286 | 0.531161 | 0.049593 | 0.647902 | 0.544429 |
| gpt-4o-mini | 0.624897 | 0.040404 | 0.226089 | 0.012858 | 0.565086 | 0.034396 | 0.668628 | 0.542084 |
| gpt-4o | 0.706664 | 0.024472 | 0.287635 | 0.009080 | 0.631743 | 0.020658 | 0.679025 | 0.537123 |

に関する知識を十分に有していないことが明らかとなった。本節では、この結論を裏付けるために、4.2 で示した結果について考察する。まず、ROUGE-1、2、L の F1 スコアにおいて、4.1.1 で期待された性能順（gpt-4o、gpt-4o-mini、gpt-3.5 の順）とは反比例する結果が得られた。このことから、評価対象とした生成 AI モデルは、長崎県の農業に特化したチューニングが行われていないことが示唆される。次に、具体的なスコアを基に生成 AI が保持する知識量について考察する。作成したデータセットの知識を基準とした場合、ROUGE-1、2、L で測定される単語レベルの知識は約 10 %程度に留まり、BERTScore で測定される文脈を考慮した知識量も約 60 %程度に過ぎないことがわかった。この結果から、生成 AI モデルの回答には、本研究で作成したデータセットに含まれる長崎県に限定した農業の知識が十分に含まれていないと結論付けられる。

以上の結果を踏まえ、生成 AI モデルに日本の地域特性や農業分野の具体的な知識を獲得させ、農業分野への応用を進めるためには、地域に特化したデータセットを基にさらなるチューニングを行う必要がある。本研究では長崎県に着目したが、今後の課題として、長崎県を含む他地域の農業特性を網羅したデータセットを構築し、それを活用して AI モデルの学習を進めることが挙げられる。

## 6　まとめ

本研究では、長崎県の農業経営類型を用いて、半自動的にインストラクションチューニングデータを生成し、作成したデータセットを基に、既存の生成 AI モデルが長崎県の農業に関する知識をどの程度保持しているかを評価した。複数のモデルを用いて実験した結果、既存の生成 AI モデルが長崎県の農業に関する知識を十分に有していないことが明らかになった。今後の課題として、長崎県を含む他地域の農業特性を網羅したデータセットを構築し、それを LLM の学習や RAG に活用して、AI の農業分野への応用を進めていきたい。

## 謝辞

本研究は、内閣府「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」における農林水産省実施施策「AI 農業社会実装プロジェクト」の助成を受けて実施された。

## 参考文献

[1] Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. **ArXiv**, Vol. abs/2308.03107, , 2023.

[2] Marvin Kästing and Christian Hänig. Assessing large language models in the agricultural sector: A comprehensive analysis utilizing a novel synthetic benchmark dataset. In **Jahrestagung der Gesellschaft für Informatik**, 2024.

[3] Matheus Thomas Kuska, Mirwaes Wahabzada, and Stefan Paulus. Ai for crop production - where can large language models (llms) provide substantial value? **Comput. Electron. Agric.**, Vol. 221, p. 108924, 2024.

[4] Maria Angels de Luis Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevao Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo Nunes, Rafael Padilha, Morris Sharp, Bruno Leonardo Barros Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. **ArXiv**, Vol. abs/2401.08406, , 2024.

[5] 小林暁雄, 坂地泰紀, 桂樹哲雄, 森翔太郎, 橋本祥, 鈴木

雅弘, 川村隆浩. 普及指導員の知識を回答可能な生成 ai のための農産物市場価値を表現するデータセットの構築. 人工知能学会全国大会論文集, 2024.

## A　Q2 および Q3 での評価実験結果

図 5、図 6 に、本文に記載できなかった、Q2 と Q3 を用いた評価実験の結果 (適合率と再現率) を示す。加えて、図 7、図 8 に、Q2 と Q3 を用いた評価実験の結果（F1）を示す。

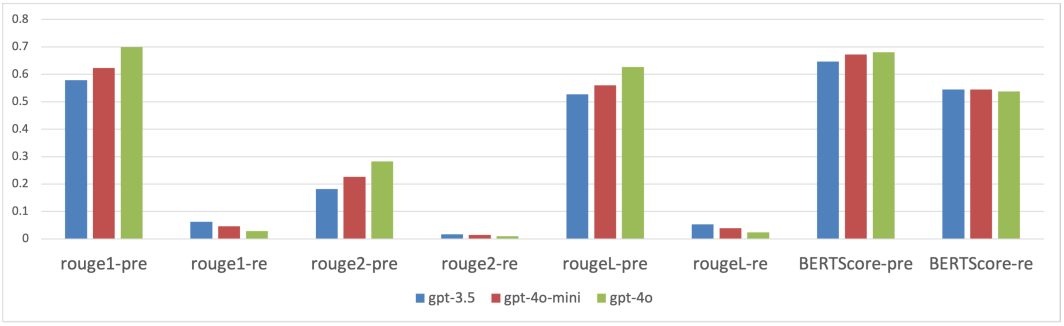


**図 5**　Q2 での ROUGE1、2、L における Recall 及び Precision

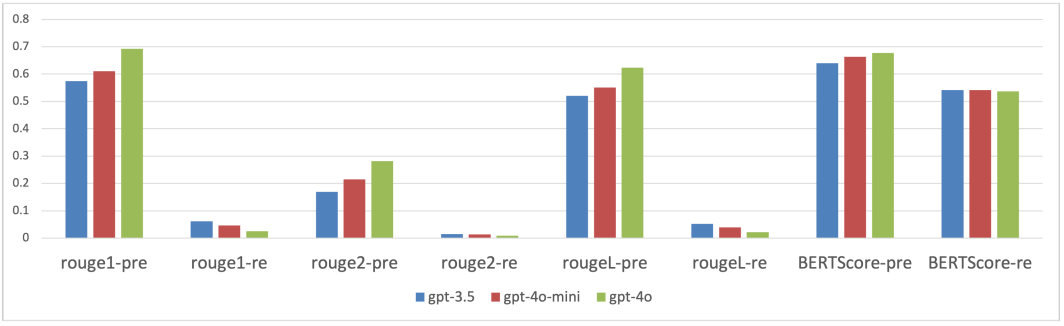


**図 6**　Q3 での ROUGE1, 2, L における Recall 及び Precision

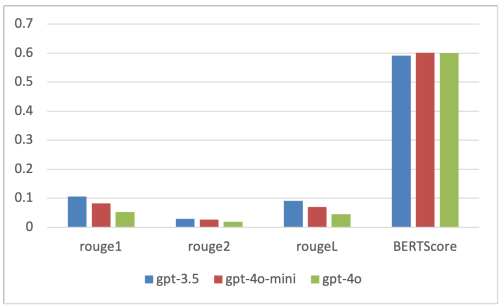


**図 7**　Q2 での各評価指標における各モデルの F1 スコア

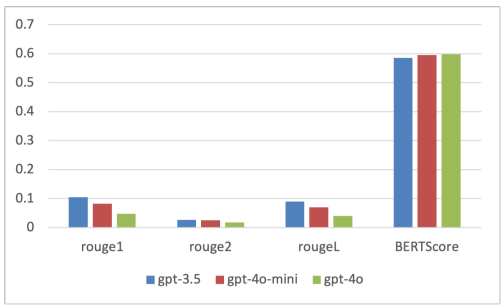


**図 8**　Q3 での各評価指標における各モデルの F1 スコア