DAGRI Subtask 3: 栽培暦を対象とした Figures and Tables Question Answering (FiT-QA) の提案

中川 董 [1]　木村 泰知 [1]　會田 勇斗 [2]　高橋 洸丞 [2]　門脇 一真 [3]
小林 暁雄 [4]　大友 将宏 [4]　石原 潤一 [4]　馬場 研太 [4]　桂樹 哲雄 [4]
[1] 小樽商科大学 [2] ストックマーク株式会社
[3] 株式会社日本総合研究所 [4] 農研機構 農業情報研究センター
kimura@res.otaru-uc.ac.jp

概要

農業分野において使用される作物ごとの作業時期や栽培管理の指針を体系的に示した「栽培暦」は, 写真・図・グラフ・表といった視覚情報に加えて，テキストが混在した複雑な文書である. 本研究では, 複数の表現形式（図, 表, グラフ, テキスト）を含む栽培暦を対象としてマルチモーダル質問応答タスクを提案する. また，既存のマルチモーダル LLM をベースラインとして精度評価を行い，本タスクの難易度を明らかにするとともに，研究課題を整理する.

1　はじめに

農業分野では, 年間を通した作物の栽培・管理スケジュールを示すガイドラインである栽培暦が作成されている. これは, 年間計画をもとに作成され, 営農指導員が個別にパーソナライズした説明をする際に用いられる. 作物・品種ごとに細かく作られており, JA(農業協同組合) や都道府県が PDF 形式で提供している.

栽培暦は, テキスト・図表・画像が混在した文書であり, 年間スケジュールを中心とした複雑なレイアウト構造を有する. 日本特有の縦書きのレイアウトや, 1 枚の栽培暦の情報量の多さから, 従来の機械的な情報抽出手法では正確な情報抽出が困難であるという課題がある.

文書画像解析の分野では, LayoutLM[1] に代表されるエンコーダ型モデルが, レイアウト情報とテキスト情報を統合的に処理し, 分類・情報抽出・構造認識といったタスクで一定の成果を挙げてきた. また, 文書中に含まれる表に対しては, 表をセル単位の文字列や HTML などの構造化表現として扱い, その上で質問に応答するテキストベースの表質問応答手法も存在する [2, 3]. これらの手法に対し，大規模視覚言語モデル（LVLM）の発展に伴い，文書画像を直接入力として End-to-End に質問応答を行う枠組みが注目されている．特に Qwen3-VL[4] は，DocVQA [5] をはじめとする文書質問応答ベンチマークにおいて高い性能を示している.

しかしながら, 農業文書を対象とした質問応答においては, 作業日程などの時系列情報を含む図表が既存のマルチモーダル LLM では適切に解釈できず, 正確な出力が困難であることが指摘されている [6]. 近年では，AgMMU [7]，AgriCoT [8]，AgriGPT-VL [9] など，農業領域に特化したマルチモーダル質問応答データセットや評価フレームワークが提案されているものの, 日本の農業現場で用いられる文書ベースの情報や，日本語特有の時間表現・レイアウト構造を含む資料を対象としたマルチモーダルでの性能は，依然として十分に評価されていない.

そこで，本研究では，テキスト・図表・画像が混在する栽培暦を対象として，新たなマルチモーダル質問応答タスクを定義する．本タスクは, 栽培暦から質問に対応する情報を適切に参照し, 回答を生成する能力を問うものである. これにより, 農業従事者が栽培暦から必要な情報へより容易に到達できるようにするための技術的基盤の構築を目指す.

本研究の貢献は下記の 3 点である.

- 栽培暦を対象とした新しいマルチモーダル QA タスクを定義した.
- ベースライン手法として, 既存のマルチモーダル LLM による実験と評価を行った.
- 実験結果の分析を通して, 本タスク固有の課題を整理した.

2 関連研究

2.1 文書画像に関する質問応答

DocVQA[5] は, 12000 件以上の画像と 50000 件の質問からなる VQA ベンチマークである. 抽出型のタスクであり, 画像内のテキストから回答を抜き出すことができる.

また, 同系列のベンチマークに InfograficVQA[10] が存在する. これは, レイアウト・図表・テキストを含む画像を対象としており, テキスト抽出に加え, 数値計算などの操作が必要な質問も含まれている.

さらに, 文書におけるチャートを対象としたベンチマークに CharXiv[11] や ChartQA-pro[12] などが存在する. これらのベンチマーク結果は, 既存の LVLM がチャート理解能力において不完全であることを示している.

視覚言語モデルのアナログ時計とカレンダーの読み取り能力を調査した研究 [13] もある. この研究では, カレンダーの読み取りにおいて, OpenAI o1 が 80 %程度の精度であると報告されたが, 栽培暦は一般的なカレンダーとは異なる構造を持ち, より複雑である.

2.2 農業分野への応用

日本の農業知識に関する研究としては, 板倉ら [14] が長崎県における農業関連文書を対象にデータセットを構築し, 大規模言語モデルが有する知識について評価している. この研究では, LLM が地域固有の農業知識を十分に保持していない点が指摘されている.

農業分野におけるマルチモーダル評価基盤の一例として，AgroBench [15] が提案されている. AgroBench は，実際の農業現場を想定した 7 つのテーマに基づき，Vision-Language Model（VLM）の性能を評価するために構築されたベンチマークであり，作物や病害虫の画像に対する質問応答タスクを中心に構成されている.

3 提案タスク

本研究で提案するタスクは，テキスト・図表・画像から構成される栽培暦の PDF 文書を画像として入力し，与えられた質問に対して適切な回答を生成するマルチモーダル質問応答タスクである.

3.1 栽培暦の概要

栽培暦の文書は, 主に以下の 5 つで構成される.

年間カレンダー　栽培暦の中心となる大きな図である. 月ごとの明確な区切り線がない大きな表形式であることが多い. 画像や図, あるいは, ○△などの記号で作業を示す.

図　作物や水管理に関する図が存在し, グラフのような読み取りを要するものも存在する.

表　年間カレンダーの内部や, 文書中に含まれる. 文書中に複数存在する.

画像　病害虫や作物, 農業機械等の画像が含まれる.

テキスト　栽培の要点や注意点の文が含まれる.

3.2 データセットの特徴

本タスクでは, 宮脇ら [16] により LLM の QA 自動生成と人手の修正で作成されたデータセットを用いる. このデータセットは, 回答が簡潔で短い単語・フレーズ・句になるように QA の作成が調整されているのが特徴である. そのため, 単純な文章抽出などの質問回答で多く構成されており, 文書中の表を読み取るだけの質問も存在する. このような簡潔な QA は, 情報量が多い栽培暦の画像に対して付与されていることが多い. また, 情報量の少ない栽培暦に対しては, 時系列を問う質問や, わずかながら△などの記号で作業を表しているような視覚特徴を利用した質問が存在している.

3.3 入出力と評価

本タスクの入力, 出力, 評価は以下の通りである.

入力	1. 質問
	2. 画像
出力	回答
評価	BLEU, LLM-as-a-Judge, 人手評価
	(詳細は第 3.5 節を参照されたい)

次に, 入力, 出力の例を示す.

入力	1. 幼穂形成期は何月の何旬ですか？
	2. `48-145-003_page_1.png`[1)]
出力	7 月中旬

1)　具体的な画像ファイルは図 1 の左側に示されている.

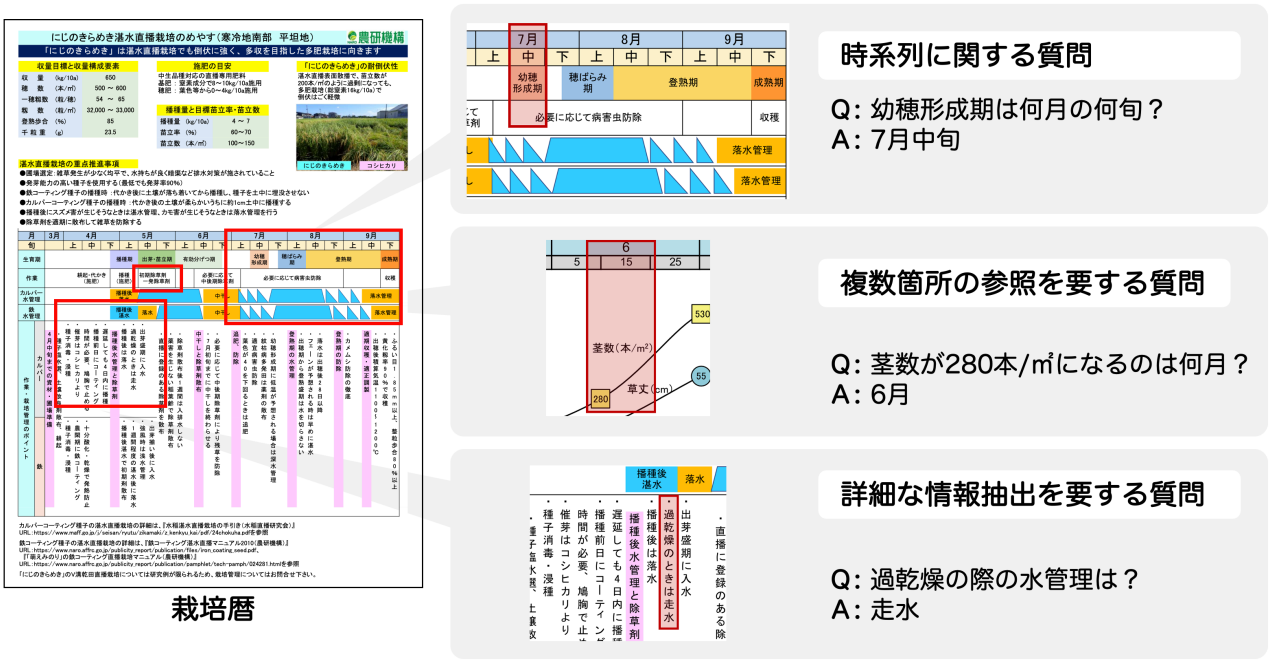


図 1　栽培暦と, 本タスクで対象とする 3 種類の質問タイプの例

3.4　質問の種類

本タスクでは, 以下の 3 種類の質問タイプを対象とする. 質問の例を図 1[1)] に示す.

時系列に関する質問　栽培暦に含まれる年間カレンダーに示されている作業の期間について問う.

詳細な情報抽出を要する質問　栽培暦に含まれる細かな図表について問う.

複数箇所の参照を要する質問　栽培暦の複数箇所を参照することで分かる情報について問う.

3.5　評価方法の検討

評価では, 自動評価指標に加え LLM-as-a-Judge や人手評価の併用を検討している. 本稿では参考値として BLEU を用いたが, BLEU のみでは回答の妥当性を十分に反映できない例が確認されたためである.

4　ベースライン手法

4.1　実験の概要

データセットの QA データ 687 件を使用し, 既存モデルの性能を評価するための実験を行った. 最新かつ高性能な商用モデルとして, **GPT-5.2** (gpt-5.2-2025-12-11) を採用した. また, オープンソースで提供される高性能かつ小型なモデルとして, **Qwen3-VL-8B-Instruct** を採用した.

入力の前処理として, 各 PDF 文書について各ページを 150 dpi で画像に変換した. また, 実験時には, QA ペアに対応するページ 1 枚の画像を入力に利用した.

4.2　実験設定

実験においては, 簡潔な回答を得るため, 以下のシステムプロンプトを用いた.

システムプロンプト

Answer the question using only the information shown in the image.
Follow these rules:
- The answer must be a short phrase or single word. Do not use full sentences.
- Use the minimum number of words necessary.
- If multiple pieces of information are required, list them using the Japanese comma “、”.

なお, temperature = 0.7, top_p = 0.9 に設定している.

1)　図 1 は, 農研機構から刊行されている「にじのきらめき」栽培暦 改訂版」(https://www.naro.go.jp/publicity_report/publication/pamphlet/tech-pamph/134301.html) を利用した作例である.

5　実験結果

表 1 に結果を示す．ここで, BLEU スコアは, 正解とモデルの出力それぞれについて, Unicode 正規化と, "〜", "-"の文字コードの正規化を行ったのち, fugashi と unidic-lite を用いてトークナイズしたものを対象として算出した．

表 1　実験結果

Model	BLEU
gpt-5.2-2025-12-11	0.5766
Qwen3-VL-8B-Instruct	**0.6517**

BLEU スコアでは，Qwen3-VL-8B-Instruct が GPT-5.2 と比較して約 7 ポイント高い値を示した．一方で，BLEU スコアは語彙の一致度に基づく指標であるため，Qwen3-VL-8B-Instruct が正解に近い表現を用いる傾向が結果に影響している可能性がある．

6　考察

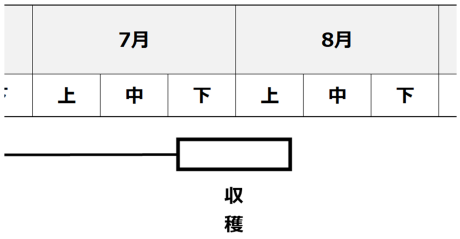
6.1　誤り分析

GPT-5.2, Qwen3-VL-8B-Instruct の両モデルが不正解であった問題の多くは時系列に関する質問である. 正解に比べて幅広い期間を回答するケースや, 上旬・中旬といった配置の微妙な差を間違うケースが見られた (図 2, 図 3). また, 単純な抽出型の質問においても, 画像の複雑さ, 情報の密集などから, 抽出箇所を誤るケースが見られた.

6.2　既存モデルにおける課題

これらのエラーは，既存のモデルが栽培暦において本質的に重要となる時系列構造を十分に捉えられていないことを示唆している．特に，上旬・中旬・下旬といった細かな区切りは，横方向のテキスト間の対応とは異なり，縦方向の相対的な位置関係の理解を必要とするが，既存モデルではこのような構造的情報がうまく捉えられていないと考えられる．

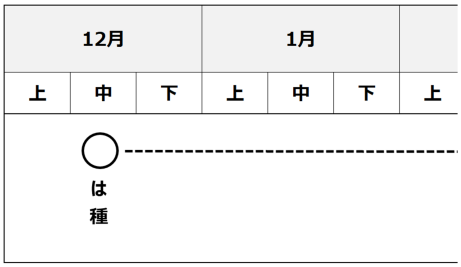
また，単純な文章抽出であっても，情報量の多い栽培暦の場合には, 同一語句が文書中に現れる数が多いことや, 視覚的に密集していることにより誤認が生じやすい．さらに，栽培暦のレイアウトは自治体ごとに異なる場合があるため，特定のレイアウトに依存した学習では十分とは言えず，レイアウトの多様性に対して汎用的に対応できる能力が求められる．



Q：収穫が行われるのは何月何旬からいつまでですか？
A：7月下旬〜8月上旬

誤り例：「7月中旬〜8月上旬」「7月下旬〜8月中旬」

図 2　より幅広い期間を回答するケース (作例)



Q：は種はどの月のどの旬に記載されていますか？
A：12月中旬

誤り例：「12月上旬」「12月下旬」

図 3　上・中旬の微妙な差で誤答が生じるケース（作例）

7　おわりに

本研究では，栽培暦を対象としたマルチモーダル質問応答タスクを定義し，データセットを用いてベースライン実験を行った．分析の結果，既存の LVLM は栽培暦において重要な時系列構造を十分に捉えられていないことが明らかとなり，提案タスクの有用性を示した．

今後の課題として，評価指標のさらなる検討が挙げられる．本稿では BLEU スコアを用いたが，表層一致に基づく評価に限られるため，LLM-as-a-Judge の導入や最終的な人手による確認による適切な評価の枠組みを検討する. また，実験結果から既存モデルは 6 割以上の質問に正答できていると考えられる．今後は，数値計算や条件付き選択など, 複雑な推論を要する QA データを拡充し，農業分野での実運用を想定した難易度の高い QA タスクの実現を目指す.

謝辞

本研究は，JST，RISTEX，JPMJRS25L2 の支援，および，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] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD ’20, p. 1192–1200. ACM, August 2020.

[2] 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, January 2025.

[3] Phuc Nguyen, Nam Tuan Ly, Hideaki Takeda, and Atsuhiro Takasu. TabIQA: Table questions answering on business document images, 2023.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025.

[5] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for vqa on document images, 2021.

[6] 杉山陽菜乃, 木村泰知. 農林業関連の文書に含まれる図表を対象とした質問応答に向けた分析. 第 41 回ファジィシステムシンポジウム, 9 2025.

[7] Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S. Adve, and Yu-Xiong Wang. AgMMU: A comprehensive agricultural multimodal understanding benchmark, 2025.

[8] Yibin Wen, Qingmei Li, Zi Ye, Jiarui Zhang, Jing Wu, Zurong Mai, Shuohong Lou, Yuhang Chen, Henglian Huang, Xiaoya Fan, Yang Zhang, Lingyuan Zhao, Haohuan Fu, Huang Jianxi, and Juepeng Zheng. AgriCoT: A chain-of-thought benchmark for evaluating reasoning in vision-language models for agriculture, 2025.

[9] Bo Yang, Yunkui Chen, Lanfei Feng, Yu Zhang, Xiao Xu, Jianyu Zhang, Nueraili Aierken, Runhe Huang, Hongjian Lin, Yibin Ying, and Shijian Li. AgriGPT-VL: Agricultural vision-language understanding suite, 2025.

[10] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. InfographicVQA, 2021.

[11] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal llms, 2024.

[12] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering, 2025.

[13] Rohit Saxena, Aryo Pradipta Gema, and Pasquale Minervini. Lost in time: Clock and calendar understanding challenges in multimodal LLMs, 2025.

[14] 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.

[15] Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. AgroBench: Vision-language model benchmark in agriculture, 2025.

[16] 宮脇一輝, 會田勇斗, 高橋洸丞, 中川菫, 木村泰知, 門脇一真, 小林暁雄, 大友将宏, 石原潤一, 馬場研太, 桂樹哲雄. 農業分野の栽培暦 VQA データセットの構築 —LLM のマルチモーダル質問自動生成能力の評価—. 言語処理学会 第 32 回年次大会 発表論文集, 2026.

A　付録

A.1　画像ファイルの命名規則

ファイル名は, XX-XXX-XXX_page_X.png の形で 8 つの数字とページ番号からなる. 冒頭 2 文字は都道府県コード[2)]を表し, 次の 3 文字は我々が各自治体に対して独自に割り当てた固有の番号である. 末尾の 3 文字は, 自治体ごとのファイルの中で 1 から順番に番号を付与している.

表 2　都道府県コード対応表

No.	都道府県	No.	都道府県
01	北海道	25	滋賀県
02	青森県	26	京都府
03	岩手県	27	大阪府
04	宮城県	28	兵庫県
05	秋田県	29	奈良県
06	山形県	30	和歌山県
07	福島県	31	鳥取県
08	茨城県	32	島根県
09	栃木県	33	岡山県
10	群馬県	34	広島県
11	埼玉県	35	山口県
12	千葉県	36	徳島県
13	東京都	37	香川県
14	神奈川県	38	愛媛県
15	新潟県	39	高知県
16	富山県	40	福岡県
17	石川県	41	佐賀県
18	福井県	42	長崎県
19	山梨県	43	熊本県
20	長野県	44	大分県
21	岐阜県	45	宮崎県
22	静岡県	46	鹿児島県
23	愛知県	47	沖縄県
24	三重県		

A.2　Train/Test データのフォーマット

提案タスクでは, 以下の形式の **JSONL ファイル**を想定する. Train データでは, COCO 形式 ([x, y, w, h]) の bbox 座標や, その座標に対応する画像サイズの情報, PDF における該当ページの番号の情報等を提供する. `synthetic` 属性は, LLM による自動合成 QA の出力を採用したかどうかである. また, 本稿で利用したデータセットについて, 難易度を easy と定義し, 今後より難易度の高い QA を提供する際は difficult として, JSONL ファイルに記載する予定である.

Listing 1　入力データ (Train.jsonl) の例

```
{"question_id": "train_1",
 "file_name": "01-020-001_page_1.png",
 "question": " ... ",
 "page": 1,
 "image_width": 1754,
 "image_height": 1241,
 "bbox": [1098.68, 444.87, 68.85, 197.05],
 "answer": " ... ",
 "synthetic": true,
 "difficulty": "easy"}

{"question_id": "train_2",
 "file_name": "01-020-001_page_1.png",
 "question": " ... ",
 "page": 1,
 "image_width": 1754,
 "image_height": 1241,
 "bbox": [1095.9, 723.2, 71.0, 52.8],
 "answer": " ... ",
 "synthetic": true,
 "difficulty": "easy"}
  ...
```

Listing 2　入力データ (Test.jsonl) の例

```
{"question_id": "test_1",
 "file_name": "01-020-001_page1.png",
 "page": 1,
 "question": " ... "}

{"question_id": "test_2",
 "file_name": "01-020-001_page1.png",
 "page": 1,
 "question": " ... "}
  ...
```

Listing 3　回答データ (Output.jsonl) の例

```
{"question_id": "test_1", "answer": " ... "}

{"question_id": "test_2", "answer": " ... "}
  ...
```

2)　https://www.soumu.go.jp/denshijiti/code.html