表・グラフに対する高難度の日本語 QA データセットの構築

阿部 晃弥 [1]　新納 浩幸 [2]
[1] 茨城大学大学院 理工学研究科 情報工学専攻
[2] 茨城大学大学院 理工学研究科 情報科学領域
{24nm701t,hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

概要

VLM は，図表の読み取りやそれに基づく質問応答 (QA) への応用が期待される．本研究では，既存の日本語グラフ QA ベンチマークでは捉えにくい図表に対する計算・比較を伴う推論能力に着目した，高難度の日本語 QA データセットを提案する．本データセットは，棒グラフ，線グラフ，円グラフ，および表を含む画像に対する質問・正答のペアから構成される．また，各質問に対して，図表タイプ，推論ステップの種類，ステップ数，および計算・比較の必要性の有無を示すタグ情報を付与することで，推論構造に基づく詳細な分析を可能とする．実験の結果，本データセットが VLM に対して高度な図表推論能力を要求することが確認された．

1　はじめに

近年，自然言語処理の分野では，大規模言語モデル (Large-Language Model，LLM) が研究の中心的役割を担っており，マルチモーダルな情報を処理する Vision-Language Model(VLM) への関心も高まっている．VLM は，画像と自然言語を統合的に処理できることから，統計資料や業務レポートなどに含まれる図表の読み取りやそれに基づく質問応答 (QA) への応用が期待される．

これまで，英語を中心として図表に対する QA を対象としたデータセットや評価ベンチマークが数多く提案されてきた．DVQA[1] や FigureQA[2] は，合成的な図表を用いた基礎的な視覚推論を対象としており，PlotQA[3] や ChartQA[4] は，実世界のグラフに対する数値推論や論理推論を扱う．また，InfographicVQA[5] は，視覚情報とテキストを組み合わせた複合的な推論能力と算術能力を評価する．

一方，日本語を対象とした図表に対する QA に関する研究は依然として限定的である．Japanese Heron-Bench[6] や JDocQA[7] は，日本語のテキストおよび画像情報を対象としており，JGraphQA[8, 9] は，日本語の IR 資料に含まれる図表への QA を対象としたベンチマークである．JGraphQA を用いた評価の結果，比較的小規模な VLM であっても高い正答率が得られる一方，数値の比較や計算を伴う質問では正答率が低下する傾向が確認された．この傾向は，視覚的文脈における数学的推論の難しさを指摘した MathVista[10] の分析とも整合的であり，日本語での図表に対する QA においても同様の課題が存在することを示唆している．

そこで本研究では，図表に対する計算および比較を含む，多段的な推論を必要とする高難度日本語 QA データセットを構築する．本データセットは 200 問から構成され，各質問に対して，図表タイプ，推論ステップの種類，ステップ数，および計算・比較の有無を示すタグ情報を付与することで，推論構造に基づく多角的な分析を可能とする．実験では，JGraphQA において高い性能を示した VLM を用いて評価をおこない，本データセットが従来の日本語図表 QA ベンチマークと比較して，より高度な図表および算術推論能力を要求することを示す．

また，日本語マルチモーダルモデルの高度な推論能力を測定するベンチマークとして，MMMU[11] を日本語に特化させた JMMMU[12] が提案されているが，本研究のデータセットは，図表に対する算術的な推論を中心とした多段的推論能力の評価に特化している点で，既存ベンチマークとは異なる位置づけを持つ．

2　JGraphQA ベンチマークの分析

本章では，1 章で紹介した日本語のベンチマークのうち，表およびグラフを対象とする JGraphQA[1)] を用いて，複数の VLM による評価実験をおこなう．本実験の目的は，現在の VLM における日本語の図表に対する質問応答能力を定量的に評価し，その傾

1) https://huggingface.co/datasets/r-g2-2024/JGraphQA

向を分析することである．

2.1 ベンチマークの概要

JGraphQA は，投資家向け広報 (IR) 資料内に含まれる図表を対象とした日本語 QA ベンチマークである．ベンチマークで扱う画像は，棒グラフ，線グラフ，円グラフ，および表の 4 種類であり，各種類につき 50 問，合計 200 問の質問から構成されている．各質問は，図表画像，質問文，および正答から構成されており，図表画像と質問文をプロンプトに組み込むことで，VLM に入力される．プロンプトはすべてのモデルで共通とし，特定のモデルに依存した調整は行っていない．

2.2 実験に用いた VLM

本実験では，Gemma，Qwen2.5，および Qwen3 の 3 系列の VLM を採用した．各系列について複数のモデルサイズを用意し，合計 9 種類のモデルを対象として評価をおこなった．具体的なモデルおよびサイズを以下に示す．

- google/gemma-3-4b/27b-it[2)]
- Qwen/Qwen2.5-VL-3B/7B/32B-Instruct[3)]
- Qwen/Qwen3-VL-2B/4B/8B/32B-Instruct[4)]

2.3 評価結果

各 VLM における JGraphQA 全体での正答率を算出した結果を表 1 に示す．全体として，多くのモデルにおいて高い正答率が得られ，日本語の図表 QA に対する基本的な理解能力が一定程度達成されていることが確認された．

表 1 各 VLM における JGraphQA の正答率

VLM	calc.
gemma-3-2b-it	0.653
gemma-3-27b-it	0.832
Qwen2.5-vl-3B-inst.	0.796
Qwen2.5-vl-7B-inst.	0.878
Qwen2.5-vl-32B-inst.	0.923
Qwen3-vl-2b-inst.	0.852
Qwen3-vl-4b-inst.	0.923
Qwen3-vl-8b-inst.	0.939
Qwen3-vl-32b-inst.	0.980

2) https://huggingface.co/collections/google/gemma-3-release
3) https://huggingface.co/collections/Qwen/qwen25-vl
4) https://huggingface.co/collections/Qwen/qwen3-vl

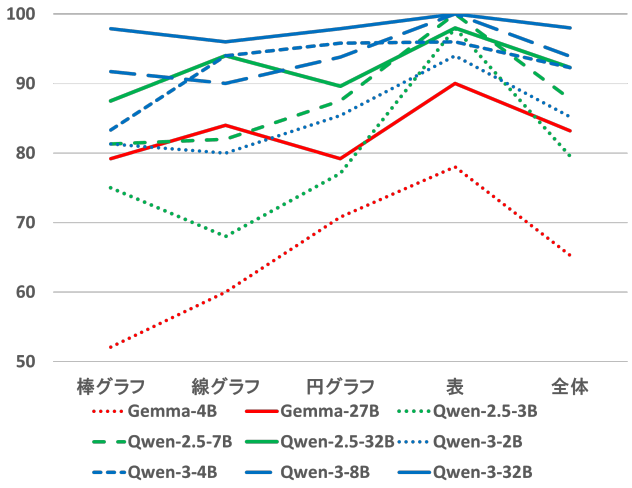

図 1 グラフ・表ごとの各 VLM の正答率

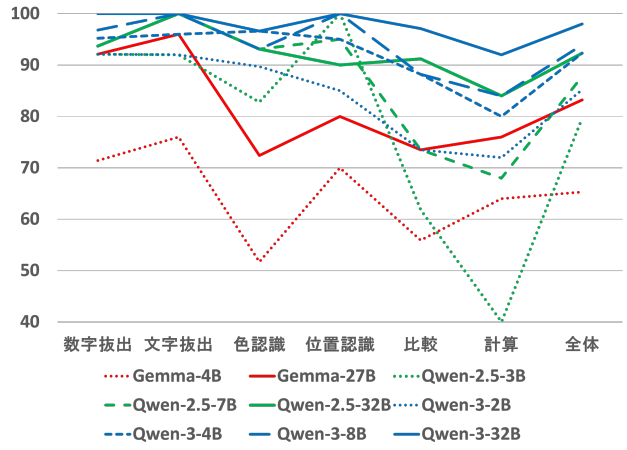

図 2 質問カテゴリごとの各 VLM の正答率

2.4 質問カテゴリおよび図表タイプ別分析

より詳細な分析をおこなうため，質問の性質に基づいて各質問を 6 つのカテゴリに分類し，カテゴリ別に分析を行った．各カテゴリの定義および具体例は付録 A に示す．

図表の種類ごとに質問カテゴリの質問数分布を整理し，併せて，各 VLM に対して図表タイプ別および質問カテゴリ別の正答率を算出した．これら結果を図 1 および図 2 に示す．

これらの分析結果から，JGraphQA 全体としては非常に高い正答率が得られる一方で，数値の比較や計算を伴う質問においては，正答率が相対的に低下する傾向が確認された．この結果は，現在の VLM が基本的な図表の読み取りには強い一方，算術・比較といった推論に課題を残していることを示唆している．

3　高難度データセットの構築

3.1　タスク概要

本研究では，図表に基づく QA タスクを対象とする．入力として，図表を含む画像と，それに対応する日本語の質問文を与え，出力として自然言語または数値による回答を生成させる．対象とする図表は，棒グラフ，線グラフ，円グラフ，および表の 4 種類であり，実社会で用いられる統計資料や報告書に含まれる形式を想定する．

使用する画像は，総務省が公開している「令和 7 年版情報通信白書」[5)]に含まれる各ページを分割して PDF 化し，さらに PNG 画像に変換したものである．

本データセットは，図表の視覚的な読み取りに加えて，数値の計算や比較，条件付き推論といった高度な推論能力を評価することを目的として設計している．

3.2　質問設計方針

第 2 章で述べた通り，既存の日本語の図表に対する QA ベンチマークには，単純な値の読み取りや限定的な算術的推論にとどまる質問が多く含まれており，近年の高性能な VLM に対しては容易に解かれてしまう傾向が確認された．そこで本研究では，質問の難易度を「必要とされる推論構造」によって制御する方針を採用した．

具体的には，以下の点を重視して質問を設計した．

- 図表中の複数要素を対象とした比較推論
- 数値の加減算や割合計算などの算術的推論
- 複数の推論を段階的に組み合わせる多段的推論
- 年度や条件を限定した非隣接要素の参照

これにより，VLM にとって高度な推論を要する質問となるよう配慮した．

3.3　推論タグの設計

本データセットでは，各質問に対して，推論構造を明示的にあらわすタグ情報を付与する．これにより，単一の正答率評価にとどまらず，質問の性質に基づいた詳細な分析を可能とする．

付与するタグは以降の節で紹介する 4 種類である．

3.3.1　図表タイプ

質問の回答に使用する図表のタイプをあらわす．棒グラフ (bar)，折れ線グラフ (line)，円グラフ (circle)，および表 (table) に対応しており，複数の図表を参照する質問については，該当するすべてのタイプを列挙する．

3.3.2　推論ステップ種別

質問の回答に必要な推論の種類をあらわすタグである．各推論種別は，あらかじめ定義した以下の分類表 2 に基づいて付与する．また，段階的な推論を必要とする質問については，`"steps":["calc_multi","compare","rank","count"]` このように要求される推論の種類の順番通りに列挙する．

表 2　推論ステップ種別の表示方法

付与するタグ	図表の種別
calc_single	単一式による計算
calc_multi	複数回の同一種類の計算
compare	値の比較
rank	項目の順位付け
position	位置情報の認識
logic	複合条件・積集合の推論
filter	排他的条件の適用
unit	単位変換，誤差変換
count	条件付き計数

3.3.3　ステップ数

質問に回答するために必要な推論ステップの段階数をあらわす．

3.3.4　計算・比較の有無

数値計算および比較推論を含むか否かを明示し，要求される推論能力を区別する．

3.4　データ形式

本データセットは，ID，質問文，正答，および画像 ID を記述した CSV ファイルと，ID および各種タグ情報を記述した JSONL ファイルによって構成される．いずれも機械可読な形式であり，推論構造に基づく定量的な分析やサブセットごとの評価を容易におこなうことができる．

5)　https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/pdf/index.html

4　本データセットの実験

本章では，第 3 章で構築したグラフ・表に対する高難度の日本語 QA データセットを用いて実験をおこなう．実験条件は第 2 章と同様とし，使用する VLM は，高難度データセットにおける性能差を明確にする目的から，JGraphQA に対して優秀な成績を残した Qwen3 系列のモデルに限定した．

4.1　評価結果

Qwen3 の各パラメータサイズにおける出力結果について，著者による人手評価をおこなった．その結果を以下の表 3 に示す．

表 3　本データセットおよび JGraphQA の Qwen3 における正答率

VLM	JGraphQA calc.	本稿 calc.
Qwen3-vl-2b-inst.	0.852	0.090
Qwen3-vl-4b-inst.	0.923	0.195
Qwen3-vl-8b-inst.	0.939	0.340
Qwen3-vl-32b-inst.	0.980	0.545

JGraphQA に対する評価結果と比較すると，本データセットでは全体的に正答率が大幅に低下しており，パラメータ数の違いによる性能差がより顕著に現れていることが確認された．この結果は，本データセットが JGraphQA と比較して，より高度な推論能力を要求することを示している．

4.2　結果の分析

次に，本データセットに付与したタグ情報を用いて，実験結果の詳細な分析を行う．

まず，図表タイプごとに正答率を算出した結果を図 3 に示す．その結果，折れ線グラフにおいて正答率の若干の低下が見られたものの，全体として図表タイプ間で顕著な性能差は確認されなかった．

続いて，質問カテゴリごとに正答率を算出した結果を図 4 に示す．その結果，「filter」および「unit」のカテゴリにおいて正答率が低下する傾向が確認された．これは，排他的条件を含む推論や，単位変換を伴う算術的推論が，VLM にとって特に困難であることを示唆している．

5　おわりに

本研究では，図表に対する高度な推論能力の評価を目的として，計算および比較推論に特化した高難度の日本語 QA データセットを構築した．本データセットは，推論構造に着目した質問設計およびタグ付与をおこない，図表に対する推論能力を評価可能とした．実験結果から，本データセットが JGraphQA と比較して，より高度な図表理解および推論能力を要求することが示された．

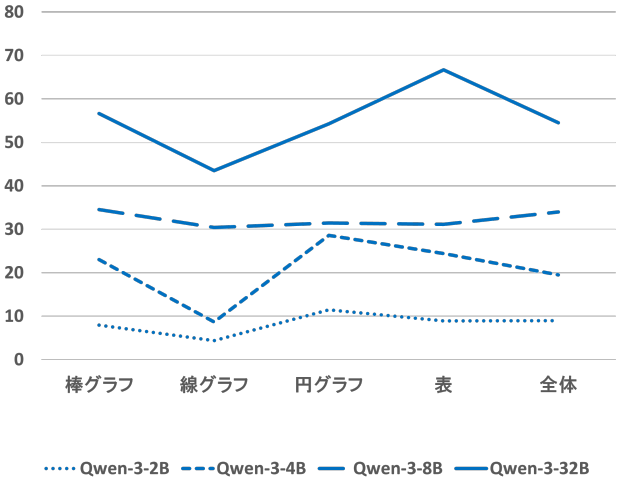

図 3　グラフタイプごとの Qwen3 の正答率

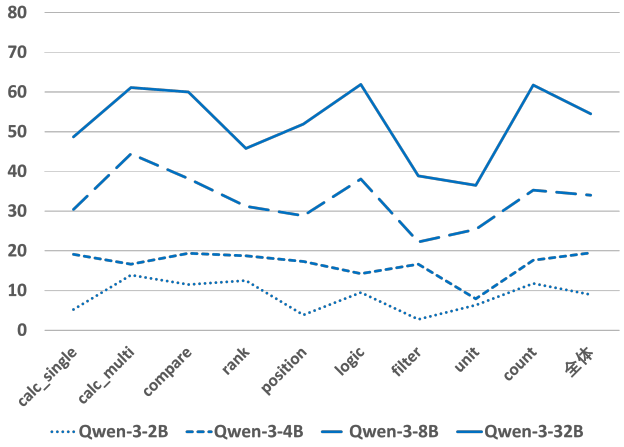

図 4　質問カテゴリごとの各 VLM の正答率

今後の課題として，質問数の拡充に加え，より多様な VLM を用いた評価や，人手による正答率との比較を通じた難度設定の妥当性検証が挙げられる．また，本データセットを用いた分析を通じて，VLM が苦手とする推論構造を明らかにすることで，今後のモデル設計や評価手法の改善に資する知見が得られると期待される．

謝辞

本研究は JSPS 科研費 23K11212 の助成を受けています．

参考文献

[1] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering, 2018.

[2] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018.

[3] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots, 2020.

[4] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022.

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

[6] Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. Heron-bench: A benchmark for evaluating vision language models in japanese, 2024.

[7] 大南英理, 栗田修平, 宮西大樹, 渡辺太郎. Jdocqa: 図表を含む日本語文書質問応答データセットによる大規模言語モデルチューニング, 2024.

[8] 経済産業省. 生成 ai 基盤モデル開発 第 2 期 成果物 (モデル・データセット) 公開, 2024.

[9] 株式会社リコー. Jgraphqa, 2024.

[10] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.

[11] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024.

[12] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation, 2025.

A　JGraphQA の質問カテゴリおよび図表タイプの質問数の分布

JGraphQA に対する質問カテゴリによる分類に使用した各カテゴリの定義および質問の具体例を以下の表 4 に示す.

表 4　JGraphQA の各カテゴリの定義および質問の具体例

カテゴリ	説明	具体例
計算	計算が必要な質問	2021 年から 2022 年の配当性向は何%上がったか？
比較	値の比較が必要な質問	ROE(%) が最も低い年は？
位置認識	画像内の位置関係の認識が必要な質問	預金利息の値は右軸か左軸かどちらになるか？
色認識	画像内の色の認識が必要な質問	赤の折れ線グラフは何を表しているか
数字認識	上記に分類不能かつ数字が正答の質問	2022 年度の地銀平均は何%か？
文字認識	上記に分類不能かつ文字が正答の質問	このグラフのタイトルはなにか

JGraphQA ベンチマークに対する質問カテゴリによる分類の詳細な質問数の分布を表 5 に示す.

表 5　JGraphQA における図表の種類ごとの質問カテゴリの分布

	計算	比較	位置認識	色認識	数字認識	文字認識
棒グラフ	7	9	1	11	15	5
線グラフ	11	15	3	10	8	3
円グラフ	7	9	0	7	23	2
表	0	1	16	1	17	15

B　本データセットの例

質問・正答ペアの CSV ファイルの内容から一部を抜粋して以下に示す.

```
id, question, answer, image
161,2011 年から 2024 年にかけて、70 代のスマートフォンの利用率は何パーセント変化したか,52.3 %,16
201, 事業者・消費者間電子商取引市場規模が連続して増加した最長区間の長さを求めよ,6 区間,20
221,e-Tax の利用状況において、利用率が 4 区間連続で 4 %以上増加したことがある手続きは何か, 所得税申告,22
361,2012 年以降、Apple の売上高が連続して増加した最長区間の長さを求めよ,4 区間,36
831, 重大な事故発生件数の推移において、前年度比で変化の割合が最も大きいのは何年度か,2023 年度,83
1071,2024 年から 2025 年にかけて、Apple の時価総額はいくら上昇したか,7230 億ドル,107
```

同様に，タグ情報を付与する JSONL ファイルの内容から一部を抜粋して以下に示す.

```
{"id":161,"chart_type":["line"],"steps":["position","calc_single"],"n_steps":2,"has_calc":true,"has_compare":false}
{"id":201,"chart_type":["bar"],"steps":["compare","count"],"n_steps":2,"has_calc":false,"has_compare":true}
{"id":221,"chart_type":["line"],"steps":["calc_multi","compare","count"],"n_steps":3,"has_calc":true,"has_compare":true}
{"id":361,"chart_type":["line"],"steps":["position","compare","count"],"n_steps":3,"has_calc":false,"has_compare":true}
{"id":831,"chart_type":["bar"],"steps":["calc_multi","compare"],"n_steps":2,"has_calc":true,"has_compare":true}
{"id":1071,"chart_type":["table","table"],"steps":["calc_single"],"n_steps":1,"has_calc":true,"has_compare":false}
```

C　データセットの図表とカテゴリの内訳

実際のデータセットから，図表タイプと，各カテゴリの質問の数の内訳を以下の表 6 に示す.

表 6　図表タイプと各カテゴリの質問数の内訳

	calc_single	calc_multi	compare	rank	position	logic	filter	unit	count
bar	58	47	114	25	23	19	14	38	14
line	23	15	38	6	23	7	7	13	11
circle	24	5	25	11	6	1	10	12	8
table	27	23	34	10	11	6	10	12	8