E-069

MECHA-Ja を用いた視覚言語モデルの日本の文化・常識理解度の評価

長谷川 騎平[1,4]　徳久 良子[1,5]　前田 航希[2,4]　小田 悠介[4]　栗田 修平[3,4]　岡崎 直観[2,4]

1. はじめに

近年，視覚と言語を統合的に扱える視覚言語モデル (Vision-Language Model; VLM) の性能が飛躍的に向上している．VLM の性能を評価する代表的なタスクには，画像質問応答 (Visual Question Answering; VQA) が広く用いられているが，最近では視覚認知能力の評価対象として，地域特有の文化の理解に注目が集まっている [1], [2]．我々はこれまでに，日本の文化や日常生活の知識を問う VLM 用のベンチマーク MECHA-Ja(Multimodal Everyday-life and Cultural Habits Assessment for Japanese) を構築し[*1]，GPT-4o [3]，Qwen [4]，LLaVA [5] などの主要な VLM が日本の文化や地域特有の生活知識をどの程度理解しているか，評価してきた [6]．図 1 に MECHA-Ja の設問例を示す．MECHA-Ja は「画像」「画像に対する質問」「4 択の選択肢」「正解」から構成されている．本稿では，MECHA-Ja に新たに「日本らしさ」と「日常らしさ」のスコアを人手で付与し，主要な VLM の性能をより詳細に分析する．これは，日本文化や生活に関する知識を Web のデータから獲得するのは難しく，既存の VLM はこれらに関する理解が不十分であろう，という仮説に基づく．

2. 関連研究

2.1 VLM の多様な文化・常識理解

近年，さまざまな VLM が提案され，画像を対象とする質問応答などの高度なタスクを高い精度で実現できるようになってきた [7], [8]．また最近では，地域特有の文化理解が注目されており，CulturalVQA [2] や GD-VCR [9] など，文化や地域特有の常識理解を評価するためのベンチマークの整備も進められている．例えば，MaRVL [10] は画像記述の真偽を判定する多言語・多文化対応のベンチマークで，インドネシア語や中国語など複数言語で構成されている．また，JMMMU [11] は日本の文化を扱うデータセットで，アートや文化遺産などの学術的なドメイン知識を必要とする設問を中心に構成されている．

我々は，日本の文化や習慣に関する視覚的な理解度を測るベンチマークとして MECHA-Ja を構築している [6]．MECHA-Ja は，画像に映っている物体やその属性などについての直接的な質問（例：「映っている動物は何か」「画像に映っている人の服は何色か」）ではなく，画像に映った物の背景にある日本特有の生活様式や文化的慣習に基づく知識を必要とする質問によって構成されている．

質問

この写真の料理の中で、勝負運をアップさせる縁起の良い食べ物とされているのは次のうちどれですか。

選択肢

A.数の子　B. 筍　C. 黒豆　**D. 栗きんとん**

日本らしさ ↑ (0.0~1.0)	日常らしさ ↑ (0.0~1.0)
0.96	0.89

A. 数の子　C. 栗きんとん

図 1: MECHA-Ja[*1] に対して「日本らしさ」と「日常らしさ」を付与した例．選択肢の中の赤字は正解を表す．

2.2 人手によるアノテーションの質の向上

クラウドソーシングで収集されるデータには，アノテータ間の判断基準の違いや，タグ付与基準の理解不足などにより，付与されるラベルにばらつきが生じることが知られている．こうしたラベルのばらつきは，アノテーション結果の分析の正確性を低下させる要因となる．この課題を解決するため，Hovy らは MACE（Multi-Annotator Competence Estimation）を提案した [12]．MACE は複数のアノテータが付与したノイズを含むラベルから，より信頼性の高い「真のラベル」を推定する手法で，アノテータ間のラベルの一致率に基づいて各アノテータの信頼度を数値化する仕組みも持つ．例えば，10 名のうち 9 名が「1」というラベルを付与した設問に対して，アノテータ A だけが「0」を付与した場合には，アノテータ A の信頼度は低く見積もられる．本稿では MACE を活用し，信頼性の高いアノテータのラベルのみを用いて分析を行う．

3. MECHA-Ja へのスコア付与

3.1 ラベルの定義

本稿では「日本らしさ」と「日常らしさ」を以下のよう

1 愛知工業大学
2 東京科学大学
3 国立情報学研究所
4 国立情報学研究所 大規模言語モデル研究開発センター
5 理化学研究所

*1 https://huggingface.co/datasets/llm-jp/MECHA-ja

に定義する．

日本らしさ（4 段階）：日本の独自性の度合いを表す

0. 世界中どこにでもある
1. 日本を含めた世界の一部にある
2. 日本を含めた世界の一部にあるが，日本を思い出されることが多い
3. 日本にしかない

日常らしさ（2 段階）：教育や経験から身につく知識かどうか

0. 日本の教育や日常生活の経験からは，身につけられない知識
1. 日本で生活していれば教育や日常生活の経験から身につけられる知識

3.2 スコアリングの方法

図 2 を用いて，「日本らしさ」と「日常らしさ」のスコアの付与方法を説明する．

(手順 1) アノテーション：図 2 の (1)

まず，MECHA-Ja の全設問 1819 問のうち 1800 問 [*2]に対して，Yahoo クラウドソーシング [*3] で「日本らしさ」と「日常らしさ」のラベルを付与する．ラベルの定義は 3.1 節に示す．各設問には 10 名のアノテータが独立にラベルを付与し，アノテータには MECHA-Ja の「画像，質問，背景テキスト，選択肢，正解」を提示し，これらすべての情報から総合的に「日本らしさ」と「日常らしさ」を判断するよう指示した．なお，アノテーションの質を担保するため，事前タスクを実施して 521 名をホワイトリストに登録し，該当者のみが参加可能な形で「日本らしさ」と「日常らしさ」を付与するタスクを実施した．その結果，521 名中 308 名が本タスクに参加した．アノテータあたり 10 問のラベルづけを依頼し，150 円の報酬を支払った．

(手順 2)MACE：図 2 の (2)

次に，MACE [12] を用いて各アノテータの信頼度を算出し，下位 15%に該当するアノテータのラベルを除外した．その結果，各設問あたり最大 10 名，最小 6 名，合計 261 名のアノテータが付与したラベルが得られた．これらに対して Krippendorff の α [13] で信頼度を評価した結果，$\alpha = 0.756$ となった．この値は「十分とは言えないが探索的研究など限定された目的には使用可能」を意味する [*4]．このことから，分析に支障のない品質で「日本らしさ」と「日常らしさ」のタグ付けが実施できたと考えられる [13]．

[*2] MECHA-ja は，画像に写っている物体やその属性を直接的に問う設問を含まないように設計されていたが，それを満たさない 19 問をタグづけの対象から除いた．

[*3] https://crowdsourcing.yahoo.co.jp/

[*4] Krippendorff の α は ≥ 0.800 はデータの信頼性は高く結論を出すのに十分信頼できる，0.667〜0.800 は十分とは言えないが探索的研究など限定された目的には使用可能，< 0.667 は信頼性が低くデータの使用は推奨されない，と定義されている．

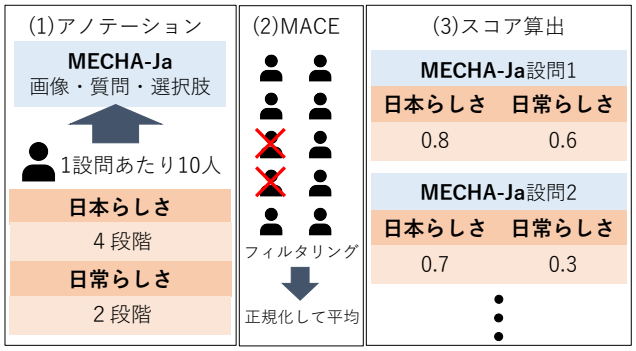

図 2: 「日本らしさ」と「日常らしさ」のスコアの付与方法

(手順 3) スコア算出：図 2 の (3)

最後に，MECHA-Ja の各設問 q_i に対して「日本らしさ」「日常らしさ」「日本らしさと日常らしさの統合」のスコアを算出した．具体的には MACE を通過した N_i 名のアノテータが付与した「日本らしさ」と「日常らしさ」のラベルを 0〜1 に正規化し，その平均を設問 q_i の「日本らしさ」「日常らしさ」とし，それらの和を「日本らしさと日常らしさの統合スコア」とした．数式を以下に示す．

$$\mathrm{Jscore}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\frac{\mathrm{Jraw}_{i,n}}{3} \tag{1}$$

$$\mathrm{Dscore}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\mathrm{Draw}_{i,n} \tag{2}$$

$$\mathrm{JDscore}_i = \mathrm{Jscore}_i + \mathrm{Dscore}_i \tag{3}$$

ここで，N_i は設問 q_i に対して MACE を通過したアノテータ数を，$\mathrm{Jraw}_{i,n}$ および $\mathrm{Draw}_{i,n}$ は設問 q_i にアノテータ n が付与した「日本らしさ $(0,1,2,3)$」と「日常らしさ $(0,1)$」のラベルを表す．

4. 実験

4.1 MECHA-Ja の性質の評価

3.2 節で求めた「日本らしさ」と「日常らしさ」のスコアごとの MECHA-Ja の設問数を図 3 に示す．「日本らしさ」の平均値は 0.77，中央値は 0.83，「日常らしさ」の平均値は 0.72，中央値は 0.75 であった．これらの結果から MECHA-Ja は「日本らしさ」や「日常らしさ」の高い設問が中心ではあるものの，日本特有とは言えない設問や，日常らしさが低い設問も一定数含んでいることが分かった．

4.2 「日本らしさ」と「日常らしさ」ごとの LLM の評価

VLM の評価基盤である llm-jp-eval-mm [14] を用いて，MECHA-Ja における VLM の精度を評価した [*5]．図 4 お

[*5] llm-jp-eval-mm は複数の日本語マルチモーダルベンチマークを統合した VLM のための統合評価基盤である．llm-jp-eval-mm は右記から利用できる：https://github.com/llm-jp/llm-jp-eval-mm

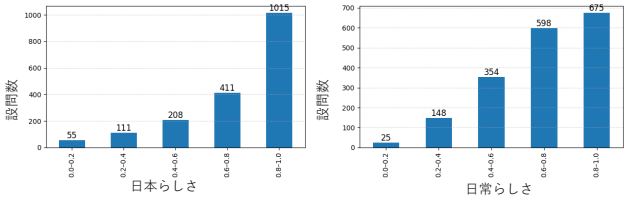

図 3: 「日本らしさ」と「日常らしさ」のスコアごとの MECHA-Ja の設問数.

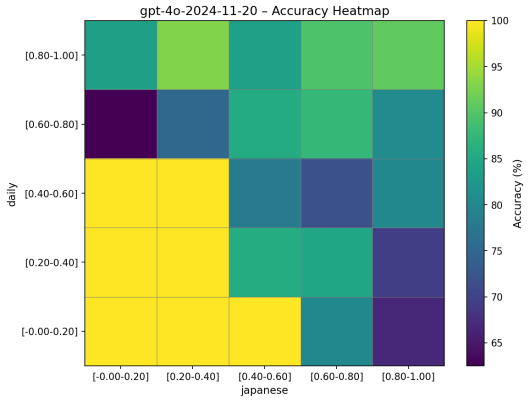

図 4: GPT-4o-2024-11-20 の精度.

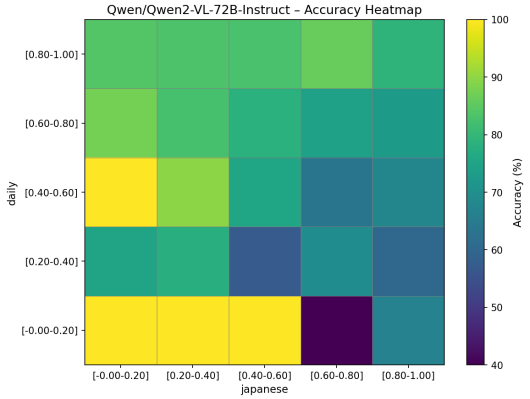

図 5: Qwen2-VL-72B の精度.

よび図 5 に，GPT-4o [3] と Qwen2-VL-72B [4] の評価結果を示す．図 4 と図 5 の横軸は「日本らしさ」のスコアを，縦軸は「日常らしさ」のスコアを表し，各領域の色（黄色〜紫）は当該領域の設問に対する VLM の精度を表している．まず，図 4 と図 5 の左下の黄色の領域が示す通り，「日本らしさ」と「日常らしさ」の両方が低い設問に対しては，GPT-4o と Qwen2-VL-72B の精度は極めて高かった．設問例を図 6 に示す．この例の場合，河豚（ふぐ）に毒が含まれるという知識は日本特有の知識ではない．このように，日本特有の知識を必要とせず，また日常生活の経験から自然と習得されるような知識ではない内容については，大規模な汎用データにより十分学習されているため，VLM は高い精度を示したと考えられる．

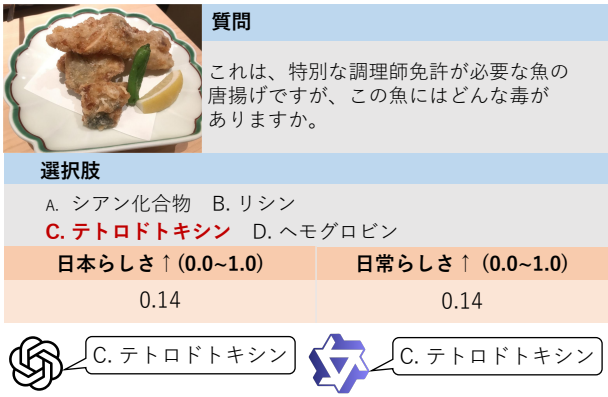

図 6: 「日本らしさ」と「日常らしさ」が低い事例.

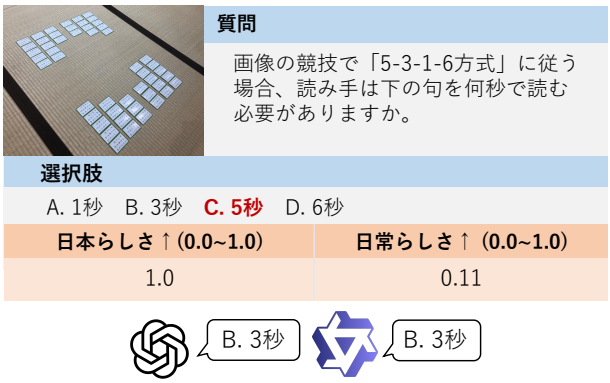

図 7: 「日本らしさ」が高く，「日常らしさ」が低い事例.

一方，図 4 と図 5 の右下の紫や濃い青の領域が示す通り，「日本らしさ」が高く「日常らしさ」が低い事例については，GPT-4o や Qwen2-VL-72B の精度は低かった．設問例を図 7 に示す．この例は「競技かるた」における下の句を詠む時間についての設問である．このような日本特有の知識を必要とする設問については，Web 上の情報量が限られているため，最先端の VLM であっても十分な性能が出なかったと考えられる．

4.3 統合的なスコアごとの LLM の評価

3.2 節の式 3 で求めた「日本らしさ」と「日常らしさ」を統合したスコアに対する各 VLM の精度を，図 8 に示す．図 8 の横軸は統合スコアの値を，縦軸の各行は VLM の精度を表している．例えば，図 8 の一行目は GPT-4o-2024-11-20，2 行目は gemma-3-4b-it の精度を示す．図全体を見ると，VLM ごとに精度の絶対値には差があるものの，概して「日本らしさ」と「日常らしさ」の統合スコアが低い領域（図の左側）では黄色（高精度）が目立ち，統合スコアが高くなるにつれて（右側へ進むほど）濃い青色（低精度）が増加する傾向が確認できた．この結果は，「日本らしさ」や「日常らしさ」が高い事例ほど VLM にとっては難易度が高く，精度が低下する傾向があること

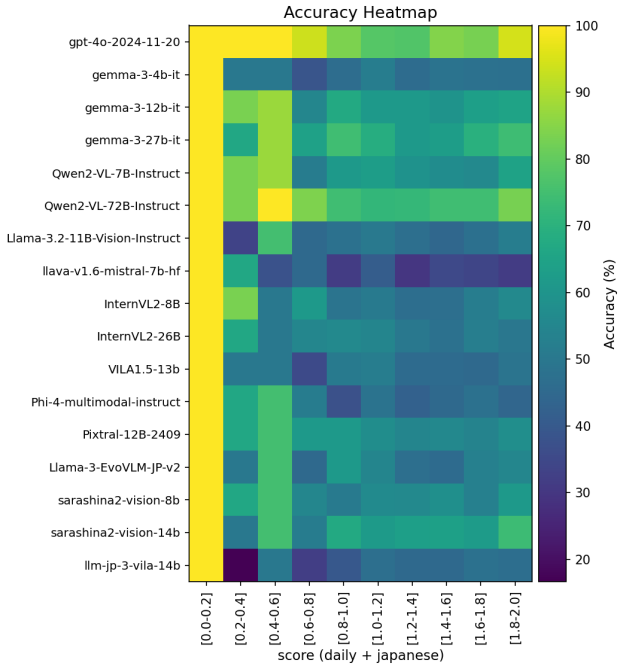

図 8: 統合的なスコアごとの主要な LLM の精度.

を示唆している．特に，小規模なモデルや日本語への特化度が低いモデルでは，この傾向が顕著に見られた．

5. 結論

本研究では，日本の文化や日常生活に関する知識を問う VLM 用のベンチマーク MECHA-Ja に対して，「日本らしさ」と「日常らしさ」のスコアを付与し，既存の VLM の性能を分析した．その結果，「日本らしさ」や「日常らしさ」が低い設問に対しては，GPT-4o や Qwen2-VL-72B をはじめとする汎用 VLM は比較的高い精度を示すことが分かった．一方，「日本らしさ」が高く「日常らしさ」が低い設問では，性能が低下する傾向が観察された．また，「日本らしさ」と「日常らしさ」を統合したスコアに基づく分析では，統合スコアが高い設問ほど VLM の精度が低下する傾向が確認された．これらの結果から，日本独自の文化や地域特有の日常的な知識など，Web 上の情報が限られている内容に関しては，既存の VLM は十分対応しきれていないことが示唆された．

今後は，日本の文化や地域特有の日常的な知識に関するデータ整備と，日本語 VLM における日本特有の知識の学習の強化を進めていく．

謝辞 京都大学の杉浦氏には，llm-jp-eval-mm を用いた評価に関してご助言とご協力を賜った．また，株式会社サイバーエージェントの佐藤氏および東北大学の守屋氏には，MACE の評価について有益なご助言をいただいた．さらに，LLM-jp マルチモーダルチームのメンバには示唆に富む議論をいただいた．ここに深く感謝する．

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです．

参考文献

[1] Romero, D. et al.: CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark, arXiv:2406.05967 (2024).

[2] Nayak, S., Jain, K., Awal, R., Reddy, S., Steenkiste, S. V., Hendricks, L. A., Stanczak, K. and Agrawal, A.: Benchmarking Vision Language Models for Cultural Understanding (2024).

[3] OpenAI: GPT-4 Technical Report, arXiv:2303.08774 (2023).

[4] Wang, P. et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, arXiv:2409.12191 (2024).

[5] Liu, H. et al.: LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge, `https://llava-vl.github.io/blog/2024-01-30-llava-next/` (2024).

[6] 前田航希，長谷川騎平，栗田修平，小田悠介，徳久良子，岡崎直観：日本の文化常識・日常生活知識理解のための視覚言語ベンチマーク MECHA-Ja の構築，情報処理学会自然言語処理研究会 研究報告 (2024-NL-263) (2025).

[7] Liu, H., Li, C., Wu, Q. and Lee, Y. J.: Visual instruction tuning, *Advances in neural information processing systems*, Vol. 36, pp. 34892–34916 (2023).

[8] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. and Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

[9] Yin, D., Li, L. H., Hu, Z., Peng, N. and Chang, K.-W.: Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 2115–2129 (2021).

[10] Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N. and Elliott, D.: Visually grounded reasoning across languages and cultures, *arXiv preprint arXiv:2109.13238* (2021).

[11] Onohara, S., Miyai, A., Imajuku, Y., Egashira, K., Baek, J., Yue, X., Neubig, G. and Aizawa, K.: JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark for Culture-aware Evaluation, *arXiv preprint arXiv:2410.17250* (2024).

[12] Hovy, D., Berg-Kirkpatrick, T., Vaswani, A. and Hovy, E.: Learning whom to trust with MACE, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130 (2013).

[13] Hayes, A. F. and and, K. K.: Answering the Call for a Standard Reliability Measure for Coding Data, *Communication Methods and Measures*, Vol. 1, No. 1, pp. 77–89 (2007).

[14] 前田航希，杉浦一瑳，小田悠介，栗田修平，岡崎直観：llm-jp-eval-mm: 日本語視覚言語モデルの自動評価基盤，言語処理学会第 31 回年次大会 (NLP) (2025).