

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

大南英理^{1,2} 栗田修平² 宮西大樹³ 渡辺太郎¹

¹ 奈良先端科学技術大学院大学 ² 理化学研究所 ³ 国際電気通信基礎技術研究所
{onami.eri.ob6,taro}@is.naist.jp shuhei.kurita@riken.jp miyanishi@atr.jp

概要

実用的な文書にはテキストだけでなく様々な図表等が含まれる。このような文書への高精度な質問応答を実現するためには、視覚情報とテキスト情報の両方の理解が必要となる。本研究では図表を含む日本語文書をもとにして、視覚情報とテキスト情報の両方を参照する質問応答データセット JDocQA を提案し、複数の日本語大規模言語モデルや画像を入力とするモデルを使用して、データセットのベンチマーキングを行った。その結果、JDocQA を用いて大規模言語モデルのチューニングを行うことで、図表を含む日本語文書の質問応答課題の性能を向上することができた。また、文書中の情報から解答できない質問を学習することにより、モデルのハルシネーション抑制に有効であることを確認した。

1 はじめに

視覚的なコンテンツとテキストとの双方を含む文書への質問に解答する課題は、文書画像質問応答 (Document VQA) として注目を集めている [1, 2, 3, 4]。また、学校の理科教材として使われるダイアグラムにアノテーションや質問応答を付与した AI2D [5] や、中学校理科テキストブックへの質問応答 TQA [6] などを筆頭に、グラフ読取、幾何学、科学などの分野で、図表やテキストを入力とした複雑な推論を志向した研究とも深く関連する [7, 8, 9, 10, 11]。図表を含む文書は社会のあらゆる場面で使用されており、実応用としても非常に重要である。昨今は、OpenAI GPT [12] に代表される大規模言語モデル (LLM) やマルチモーダルな入力可能なモデル [13, 14] が急速に発展し、社会実装が進むに伴い、様々なチューニング用や評価用のテキストデータセットが提案されてきている [15, 16, 17]。しかしながら、テキスト情報だけで

なく視覚的な情報を必要とする質問応答のチューニング用データセットはまだ十分でなく、特に英語以外の言語での視覚的な公開データセットは、大学入試分野 [18] などを除けば非常に稀有な存在である。

本研究では、グラフ、表、地図、日本語の縦書き横書き混在等の視覚的な情報とテキスト情報の両方を参照しなければ解答できない質問応答データセット **JDocQA** を提案する。JDocQA は、オープンアクセス可能な PDF 形式の文書で構成され、経済、教育、農業、歴史、文化など多岐にわたる分野を網羅し、レポート、スライド、パンフレット、Web サイト、縦書き等の様々な形式の文書を提供するデータセットである。提案する JDocQA は、日本語文書に特化した質問応答のチューニング用データセットとしても使用することができ、高性能な日本語質問応答モデルの作成が期待できる。また、本研究では、LLM が事実に基づかない情報を生成するハルシネーションに対処するため、文書中の情報だけでは解答できない質問を全データセットの 1 割程度作成し、このデータを使って追加で学習する方法を提案する。複数の日本語 LLM や画像を入力とするマルチモーダル LLM を使用して、JDocQA のベンチマーキングと提案手法の有効性を検証する。

本論文の貢献は以下のようまとめられる。(i) テキストと図・表・縦書きなどの視覚情報の両方を理解しなければ解答できない質問応答データセット JDocQA を提案した、(ii) 日本語対応 LLM に JDocQA を学習させることで、ゼロショットの OpenAI GPT と比べ高い性能を示すことを確認した、(iii) LLM のハルシネーションを防ぐため、文書中の情報では解答できない質問 (解答不能問題) を作問した、(iv) 解答不能問題を含む全データを学習させた場合と、解答不能問題を含まないデータで学習させた場合とを比較し、前者がハルシネーションの抑制に有効であることを確認した。

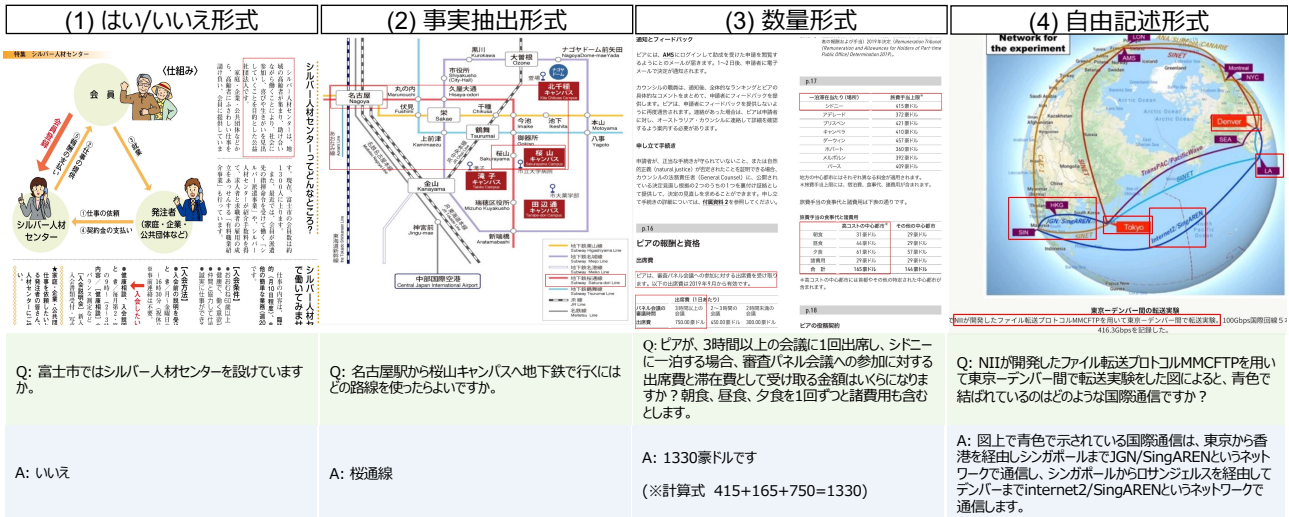


図1 質問形式毎の質問、解答および解答根拠範囲（赤色矩形領域）アノテーションサンプル。

2 JDocQA データセット

2.1 タスク

(1) 質問に対して「はいいいえ」で解答する**はいいいえ形式**、(2)本文中から事実を抜き出して解答する**事実抽出形式**、(3)簡単な四則演算や単位を含めて数字で解答する**数量形式**、(4)質問に対してある程度の長さの文章を作成して解答する**自由記述形式**の4つの形式に沿った質問を作成した。また、4つの質問形式とは独立に**複数ページを参照しなければ解答できない質問**および**文書中の情報では解答できない質問(解答不能問題)**を作成した。

2.2 データ収集

ドキュメント収集 日本の官公庁が公開しているスライドやレポート、地方自治体のパンフレット、国立国会図書館デジタルライブラリに保存されているウェブサイト、からPDFファイルを収集した。収集した文書は、レポート、スライド、パンフレット、ウェブサイトの4つに分類される。これらの文書のうち、表、グラフ、図、絵などの視覚的な情報理解が重要なものを人手で選択した。

テキスト標準化 モデルに入力するテキストを取得するため、PDFに埋め込まれた文字の抽出およびOCR¹⁾によるテキストの取得を行った。モデルの入力には、原則的に埋め込み文字の抽出を使用し、埋め込み文字が取得できない場合、例外的にOCRで取得したテキストを用いた。

1) <https://github.com/tesseract-ocr/tesseract>

アノテーション 計43名のアノテーターが、テキスト情報と視覚情報の両方を使用しなければ解答できない質問とその解答、解答根拠の矩形領域を作成した。また、解答不能問題を作成し、正解解答を「本文中に記載がありません」とした。

質問・解答データ標準化 句読点と句点の有無や全角と半角、語尾の「です」「ます」などの差異が存在したため、ルールベースで標準化を行った。

画像入力データ モデルの入力として、同じピクセルで質問の該当ページ全体の画像データ、解答の根拠を囲う矩形領域内の画像データ、白画像データの3種類を準備した。

2.3 統計

合計5,508個のPDFファイルに対して11,600件の質問応答ペアを作成した。図1は、4つの質問形式で作成された質問解答ペアの例であり、表1は、作成したデータセットの統計である。質問解答ペアは必ず参照する視覚情報を含み、図2は、視覚情報別の質問応答件数を表す。学習、検定、テストセットにそれぞれ9,290件、1,134件、1,176件の質問応答が含まれるようにデータセット全体を分割した。同一PDFファイルは必ず同一の分割に出現する。

3 実験

3.1 モデル

テキストのみの入力モデル 提案データセットの有用性を評価するため、2023年9月までに公開されていた日本語に対応しかつ13Bまでのパ

文書カテゴリ	PDF	QA	(1) はい/いいえ	(2) 事実抽出	(3) 数量	(4) 自由記述	複数ページ問題	解答不能問題
パンフレット	1,715	4,025	605	748	660	2,012	46	671
スライド	1,640	3,276	545	593	507	1,631	448	449
レポート	2,086	4,167	703	687	693	2,084	506	668
Web サイト	67	132	2	24	6	100	0	0
合計	5,508	11,600	1,855	2,052	1,866	5,827	1,000	1,788

表1 ドキュメントカテゴリ別の統計量, 質問解答種別数, 複数ページにわたる質問応答件数, 解答不能問題件数.

ラメータサイズの代表的な LLM を選定し, 学習データによるチューニングおよび評価を行った. 具体的には, rinna japanese-gpt-4B-8k, rinna japanese-gpt-1B, OpenCALM-7B, weblab-10B, PLaMo-13B, StabilityAI Japanese-StableLM-Base-Alpha-7B, StabilityAI Japanese-StableLM-Instruct-Alpha-7B および Llama-2-7B を使用した. 入力トークンサイズは, 公平性および簡単のため基本的に 1024 トークンとし, トークン数を変更した際の挙動を調べるため, rinna japanese-gpt-4B-8k を 2048, 4096, および 8192 トークンで学習させた.

マルチモーダル入力モデル 日本語に対応しかつテキストと画像の両方を入力できるマルチモーダルな LLM を選定し, 評価を行った. 具体的には Stability AI Japanese-StableLM-Instruct-Alpha-7B を使用した. 入力画像には, PDF の 1 ページ全体の画像, 矩形の解答根拠箇所を結合した画像, 白画像の 3 種類を用いた. 入力画像サイズは 800x600 ピクセルに統一した. モデルの入力制限のため, 入力トークン数は 512 トークンとした.

OpenAI GPT ベースライン OpenAI GPT の性能と比較するため, gpt-3.5-turbo-16k と gpt-4 を使用した²⁾. GPT モデルについては学習セットを利用したモデルのチューニングは行わず, なるべく精度のよいプロンプトテンプレートを人手で作成した上で, ゼロショットでの評価を行った.

3.2 評価方法

4 つの質問形式のうち, はい/いいえ, 事実抽出, 数量の質問は, ルールベースで句読点・句点の違いやですますなどの語尾の違いを除去し, 正解解答との一致を調査した. 正解解答との一致だけでなく正解解答に含まれる要素があれば加点する方式でも評価を行ったが, スコアに大きな差がなかったため正解との一致による評価とした. 自由記述形式の質問は, 解答の平均長が 65.97 文字と長いため, MeCab でトークン化された BLEU スコアを使用した.

2) 2023 年 10 月 9 日時点での最新モデルを使用した.

3.3 実験設定

データセットには解答不能問題が含まれるため, LLM がハルシネーションを起こすことが予想される. そこで, 学習段階で解答不能問題のデータを含める場合と含めない場合の 2 通りの学習を行い, 評価段階でもテストセットに解答不能問題を含める場合と除外する場合の 2 通りのテストを行った.

4 結果

テキストのみを入力するモデルの実験結果は表 2, テキストおよび画像を入力するモデルの実験結果は表 3 の通りである.

4.1 全ての質問応答で学習した場合

解答不能問題を含めた全ての学習データで学習し, テストセットに解答不能問題を含む場合と含まない場合の 2 通りでテストを行った. 表 2 にて, 解答不能問題を含む全てのデータで学習をした場合, 特にパラメータ数の大きいモデルにおいて, OpenAI の gpt-3.5, gpt-4 モデル以上の精度が出ることがわかる. 解答不能問題ありでテストした場合 (表 2 左) となしでテストした場合 (表 2 右) では, 精度に同じ傾向が見られるため, 全データで学習した場合の振る舞いが似ていることがわかる. また, Llama-2-7B は StableLM の 7B モデルと比較した場合, 解答不能問題での精度に伸び悩む傾向があった.

4.2 解答不能問題を除いて学習した場合

学習時に解答不能問題を除いたデータで学習し, 解答不能問題を含む場合と含まない場合の 2 通りでテストを行った. まず, ほぼ全てのモデルで, 全データで学習した場合の方が, 解答不能問題を除くデータで学習した場合よりも精度が良いことがわかる. これは, モデルがハルシネーションを起こすためであり, 学習データに正解が解答できない質問を入れることは, ハルシネーションの抑制に寄与すると言える.

Model	テストセット全体					解答不能問題を除いたテストセット				
	Avg.	Y/N	Fact	Num	Free	Avg.	Y/N	Fact	Num	Free
学習なしモデル (ゼロショット)										
gpt-3.5-turbo-16k	20.62	50.29	7.44	11.11	13.64	23.07	58.21	8.08	12.5	13.49
gpt-4	19.47	43.19	6.51	11.11	17.07	22.03	50.00	7.07	12.5	18.57
学習セット全体で学習したモデル										
rinna gpt-1B	20.46	59.76	5.58	8.77	18.13	20.18	64.38	4.04	8.55	17.40
rinna bi-4B-8k (8192 tok.)	23.02	62.13	8.83	11.11	20.57	21.17	62.32	6.56	7.89	19.11
OpenCALM-7B	18.33	43.78	11.62	9.94	16.03	17.53	46.57	9.59	8.55	15.10
weblab-10B	16.94	47.92	10.23	8.18	13.24	16.20	50.68	9.09	6.57	12.17
PLaMo-13B	20.33	53.84	10.69	7.01	18.21	18.35	50.68	8.58	5.26	16.87
StableLM Base-AI.-7B	29.71	70.41	15.81	22.22	25.51	28.33	71.91	12.62	21.05	24.32
StableLM Inst.-AI.-7B	29.56	72.78	16.27	21.05	24.75	28.66	76.02	12.62	19.07	24.40
Llama2-7B	27.01	61.53	17.20	18.71	23.29	25.70	65.06	12.62	18.42	21.87
解答不能問題を除いた学習セットで学習したモデル										
rinna gpt-1B	17.59	51.47	6.04	5.26	15.76	20.02	59.58	6.56	5.92	18.20
rinna bi-4B-8k (8192 tok.)	23.01	69.23	7.90	9.94	19.27	25.95	80.13	8.58	11.18	21.79
OpenCALM-7B	16.95	42.01	7.44	6.43	16.33	17.97	48.63	8.08	7.23	16.32
weblab-10B	17.96	52.07	6.51	8.18	15.34	19.06	60.27	7.07	9.21	15.06
PLaMo-13B	21.11	64.49	8.83	11.11	16.31	24.29	74.65	9.59	12.50	19.34
StableLM Base-AI.-7B	25.68	68.63	12.09	17.54	20.94	29.21	79.45	13.13	19.73	24.14
StableLM Inst.-AI.-7B	26.25	70.41	15.34	16.37	20.74	29.75	81.50	16.66	18.42	23.69
Llama2-7B	28.25	73.96	11.62	16.95	24.68	30.57	85.61	12.62	19.07	25.47

表 2 各チューニング済みモデルおよび OpenAI GPT の、はい/いいえ (Y/N), 事実形式 (Fact), 数量形式 (Num), 自由記述形式 (Free), およびそれらの重み付き平均 (Avg.) での結果を示す。

Model	テストセット全体				
	Avg.	Y/N	Fact	Num	Free
学習セット全体で学習したモデル					
InstBLIP (blank)	26.92	65.68	16.27	19.88	22.00
InstBLIP (img)	27.44	68.63	15.34	19.88	22.50
InstBLIP (bbox)	27.87	72.78	18.13	19.29	21.37
解答不能問題を除いたテストセット					
Model	Avg.	Y/N	Fact	Num	Free
解答不能問題を除いて学習したモデル					
InstBLIP (blank)	26.45	76.71	13.63	13.15	21.26
InstBLIP (img)	28.52	82.87	13.13	18.42	22.25
InstBLIP (bbox)	27.79	80.13	11.61	16.44	22.71

表 3 マルチモーダル入力を行ったモデルの結果。

4.3 マルチモーダルモデル

表 3 に詳細な結果を示す。元となったモデルの入力トークン数は 512 が上限のため、この制限が性能を低下させた恐れがある。白画像の精度と他の 2 種類のモデルの精度がまだあまり差がなく、この対策として、事前学習も含めた日本語のマルチモーダル基盤モデルの拡充が有効であると考えられる。

4.4 トークンの長さによる結果への影響

JDocQA にはコンテキスト長が長い文書が含まれているが、入力トークンの最大長を超える場合はその冒頭部分のみを入力している。この影響を調べるため、rinna bi-4B-8k モデルを用いて 2,048, 4,096,

Token length	テストセット全体				
	Avg.	Y/N	Fact	Num	Free
学習セット全体で学習したモデル					
2048 tokens	20.97	57.39	10.69	9.94	17.66
4096 tokens	21.96	56.21	9.30	13.45	19.38
8192 tokens	23.02	62.13	8.83	11.11	20.57
解答不能問題を除いて学習したモデル					
Token length	Avg.	Y/N	Fact	Num	Free
解答不能問題を除いて学習したモデル					
2048 tokens	24.57	72.60	10.10	9.21	21.18
4096 tokens	24.26	67.12	9.09	11.18	21.90
8192 tokens	25.95	80.13	8.58	11.18	21.79

表 4 3段階の異なるトークン入力最大長と結果への影響

8,192 の 3 つの入力トークン最大長で学習および評価し性能を比較した結果を表 4 に示す。入力トークン最大長が増加した場合にモデルの性能がやや改善する傾向が見られた。

5 結論

日本語ドキュメントに対する質問応答タスクである JDocQA を提案し、テキスト情報だけでなく、視覚情報を用いなければ解答できない質問応答タスクを提案した。このタスクは、文書内の情報からは解答することができない質問を含んでおり、本文中に解答のない質問を含めてモデルの学習を行うことで LLM の信頼性向上に寄与する可能性が示された。

謝辞

本研究は JST さきがけ JPMJPR20C2, および JSPS 科研費 JPMJPR22P8 の助成を受けたものです。JDocQA データセットのアノテーションはその一部を株式会社バオバブに実施いただきました。

参考文献

- [1] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In **2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 2582–2591, 2022.
- [2] Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. WebSRC: A dataset for web-based structural reading comprehension. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 4173–4185, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [3] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **AAAI**, 2021.
- [4] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In **AAAI**, 2023.
- [5] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In **Computer Vision – ECCV 2016**, pp. 235–251, Cham, 2016. Springer International Publishing.
- [6] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, July 2017.
- [7] Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. Semantic parsing to probabilistic programs for situated question answering. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 160–170, Austin, Texas, November 2016. Association for Computational Linguistics.
- [8] Mrinmaya Sachan, Kumar Dubey, and Eric Xing. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 773–784, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [9] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In **CVPR**, 2018.
- [10] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-linguistic question answering (VLQA) challenge. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 4606–4616, Online, November 2020. Association for Computational Linguistics.
- [11] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In **The 36th Conference on Neural Information Processing Systems (NeurIPS)**, 2022.
- [12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- [14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.
- [15] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [16] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In **International Conference on Learning Representations**, 2022.
- [17] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会, 2022.
- [18] 東ロボプロジェクト. ロボットは東大に入れるか. In <https://21robot.org>.

本研究の制約

本研究では、代表的な LLM を JDocQA で学習させた場合の精度を報告した。しかし、これらの結果はあくまで JDocQA による学習との相性による可能性もあり、必ずしも学習前の LLM の性能を示すものではないと考える。また、ゼロショットでの質問応答性能はプロンプトに大きく依存するため、本研究では OpenAI GPT のゼロショット性能を提示するに留めた。

付録

データセットの追加の統計量 視覚情報カテゴリ別の質問応答件数を図 2 に、質問応答カテゴリ別のテキスト統計量を表 5 に示す。表 5 から、JDocQA の入力データは比較的長めのテキストが多い。

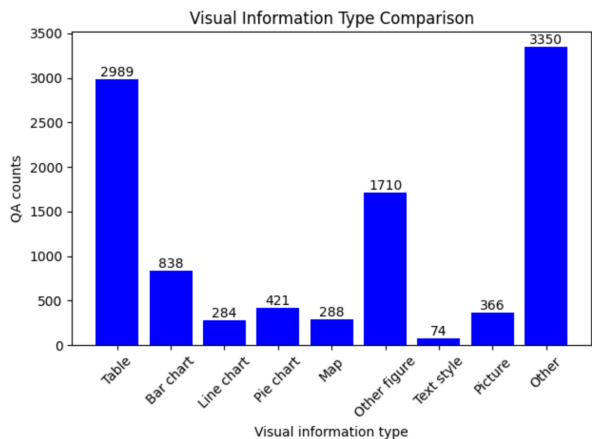


図 2 視覚情報カテゴリ別の質問応答件数

Model comparison

市税などの滞納により差し押さえをした自動車(下表)を公売します。
入札期間: 9月8日(午後1時~10日(午後11時) **代金納付期限:** 9月18日(午後2時30分) **その他:** 8月17日(前)~9月1日(以下)に見がけができます(事前申し込み必要)。滞納市税完納などにより公売が中止になる場合があります。詳細はホームページ (<http://www.citytoyohashi.lg.jp/4374.htm>) 参照 **申し込み:** 8月17日(午後1時~9月1日(午後11時)に Yahoo!JAPAN 官公庁オークションホームページ (http://koubai.auctions.yahoo.co.jp/aic_toyohashi_city/) から入力 **問い合わせ:** 納税課(課51・2241)

■差押財産の公売

売却区分番号	名称	初年度登録	走行距離(平成27年6月30日時点)	最低入札価額	公売保証金	車検有効期限
15J001	トヨタ プロボックス	平成14年	124,344km	10,000円	1,000円	平成27年10月13日
15J002	日産 リビティ	平成11年	98,629km	5,000円	500円	平成28年5月20日

Train dataset comparison

ここで、IMF は GDP 成長率が 2018 年に-1%、2019 年には-4%に落ち込み、インフレ率は上昇すると予測している。このように米国の制裁はイランに甚大な影響を及ぼすことが見込まれている。後述する通り、2020 年にトランプ大統領が再選され、対イラン制裁が継続されれば、イラン経済・社会は耐え切れず米国の協賛再開を余儀なくされると思われる。

図表 1-3 イランの実質 GDP と消費者物価指数の推移

図 6 Stable-LM モデルにおける自由記述形式出力結果の実験条件による比較

Category	(1) Y/N	(2) Fact.	(3) Num.	(4) Free
Context	963.81	1036.63	1020.04	1017.25
Question	67.75	61.26	60.36	65.44
Answer	3.77	16.01	8.22	65.97

表 5 質問応答カテゴリ別の本文、質問、解答の平均文字数。

人手評価 表 6 は、人手評価の結果を示す。テストケースの中の 100 件の質問応答をサンプルとして人手評価を行った。テキストを入力するモデルから PLaMo-13B と StableLM Instruct-Alpha-7B、テキストと画像を入力するモデルから StableLM InstructBLIP-Alpha の白画像を入力した場合を評価した。アノテーターは、アノテーションされた答えが生成結果に含まれているかどうか、間違った答えが生成結果に含まれていないかどうかの 2つの観点に対して、それぞれ 0-2 点をつけて採点を行った。

Model	Human Evaluation ↑
テキスト入力モデル	
PLaMo-13B	1.24
StableLM Instruct-Alpha-7B	1.49
マルチモーダル入力モデル	
StableLM InstructBLIP-Alpha (blank)	1.04
StableLM InstructBLIP-Alpha (img)	1.25

表 6 サンプルセットにおける人手評価結果。学習セット全体で学習したモデルを使用し、サンプルセットには解答不能問題も含まれる。

Q: 市税などの滞納により差し押さえをした自動車を公売しますが、売却区分番号15J001はどのような自動車ですか? / The vehicle seized due to tax delinquent will be sold at a public auction. Please explain the feature of ID 15J001.

A: 売却区分番号15J001は、トヨタ「プロボックス」です。初年度登録は平成14年で、走行距離(平成27年6月30日時点)が124,344kmで、車検有効期限が平成27年10月13日の自動車です。/ ID 15J001 is TOYOTA PROBOX. Register date was 2002, total mileage was 124,344km, and inspection expires 2015 Oct 13th.

StableLM InstBLIP-AI (img): トヨタ・プロボックスです。平成14年124,344kmの走行距離です。It is TOYOTA PROBOX. Mileage in 2002 was 124,344km.

StableLM Inst.-AI.-7B: 本文中に記載がありません
Not mentioned in the text

StableLM Base-AI.-7B: トヨタプロボックス平成14年124,344kmです。TOYOTA PROBOX 2002 124,344km.

Q: 2016年のイランの実質GDPと消費者物価指数の推移に影響した主要因は何か? / What is the main factor affected to the Iran's actual GDP and consumer price index in 2016?

A: 本文中に記載がありません/Not mentioned in the text

StableLM Inst.-AI.-7B (trained with all instances):
本文中に記載がありません/ Not mentioned in the text

StableLM Inst.-AI.-7B (trained without unanswerable questions):
2016年はイランのGDPは18%もの成長を記録し、インフレ率も低下した。しかし、2018年の米国の制裁復活が影響することで、IMEIはGDP成長率が2018年に-1%、2019年には-4%に落ち込み、インフレ率は上昇する。/ In 2016, Iran's GDP recorded an impressive growth of 18%, and the inflation rate decreased. However, the re-imposition of U.S. sanctions in 2018 had a significant impact, causing Iran's GDP growth rate to plummet to -1% in 2018 and further to -4% in 2019, while inflation rates rose.