# JDocQA: Japanese Document Question Answering Dataset for Generative Language Models

**Eri Onami**[1,2], **Shuhei Kurita**[2], **Taiki Miyanishi**[3], **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology, [2]RIKEN, [3]ATR
{onami.eri.ob6, taro}@is.naist.jp, shuhei.kurita@riken.jp, miyanishi@atr.jp

## Abstract

Document question answering is a task of question answering on given documents such as reports, slides, pamphlets, and websites, and it is a truly demanding task as paper and electronic forms of documents are so common in our society. This is known as a quite challenging task because it requires not only text understanding but also understanding of figures and tables, and hence visual question answering (VQA) methods are often examined in addition to textual approaches. We introduce Japanese Document Question Answering (JDocQA), a large-scale document-based QA dataset, essentially requiring both visual and textual information to answer questions, which comprises 5,504 documents in PDF format and annotated 11,600 question-and-answer instances in Japanese. Each QA instance includes references to the document pages and bounding boxes for the answer clues. We incorporate multiple categories of questions and *unanswerable* questions from the document for realistic question-answering applications. We empirically evaluate the effectiveness of our dataset with text-based large language models (LLMs) and multimodal models. Incorporating *unanswerable* questions in finetuning may contribute to harnessing the so-called hallucination generation.

**Keywords:** Multimodal Document Processing, Question Answering, Natural Language Generation

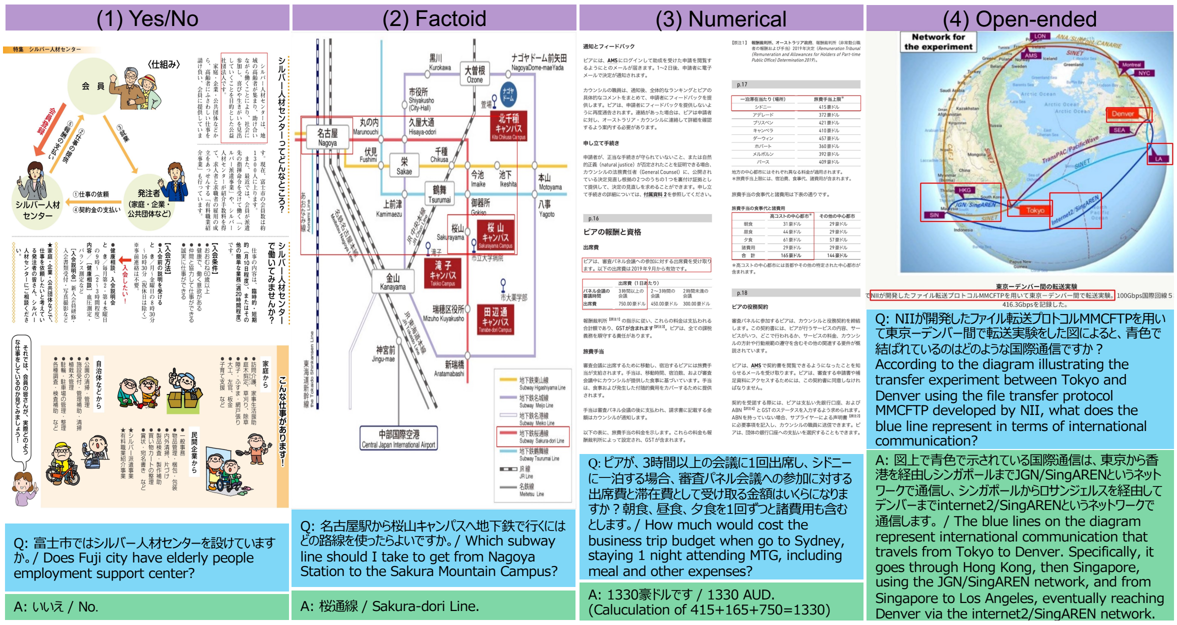
## 1. Introduction

A thorough understanding of documents that are composed of both texts and graphical elements such as slides, reports, webpages, and pamphlets is essential for intelligent agents that process multimedia documents and answer some questions on such documents. Document visual understandings have been studied to achieve joint understandings of textual and visual elements in such documents or images, including bookcovers (Mathew et al., 2021a), scene images with characters (Singh et al., 2019), webpages (Tanaka et al., 2021), tables (Smock et al., 2022) and slides (Tanaka et al., 2023). These datasets have received significant attention as documents are a common form in various industrial, public, and private sectors in the English domain. It is also notable that the document visual question answering tasks are still quite difficult despite its significance in industries because they heavily rely on both textual and visual modalities as the documents often include complex visual alignments of texts on figures, charts and illustrations. Especially in document question answering, models are required to connect multiple modalities to figure out answers. There are quite limited datasets in which both visual and textual information is required to answer questions on documents. It is also a problem that despite the significance of these tasks, the primary focus of these datasets is limited to the English domain and dataset constructions on other languages are still limited. As a document question-answering task, Japanese documents have several characteristics compared to English documents. One of the major difficulties in Japanese document processing lies in the two official writing styles in Japanese: one is a left-to-right horizontal style and the other is an upside-to-bottom vertical style, which requires both writing style comprehension in the dataset.

There has been significant progress in the generative large language models (LLMs) and multimodal models these days. GPT-4 (OpenAI, 2023) allowing zero-shot applications in both language related and even in multimodal tasks. InstructBLIP (Dai et al., 2023) takes both textual and image inputs and generates texts such as image captions or visual question answering following textual prompts. The success of LLMs also triggers the competitive development of several publicly-available LLMs in Japanese. Instruction tuning of LLM can improve its ability to adhere to certain domains or usage, rendering them more suitable for particular applications rather than maintaining its ability in a general understanding of language and limited expertise (Mishra et al., 2022; Sanh et al., 2022). While there have been numerous attempts to fine-tune instruction for highly technical and professional adaptation (Ouyang et al., 2022; Wei et al., 2022), there is still a lack of adequately prepared high-quality visual question and answering datasets that can be used for generative language model-based question answering and particularly developed outside of the English domain.

To address the demand for a large-scale and fully annotated Japanese document question answering dataset, we introduce a JDocQA dataset by collecting Japanese documents in PDF styles from

*LREC-COLING 2024*, pages 9503–9514
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Q: 富士市ではシルバー人材センターを設けていますか。/ Does Fuji city have elderly people employment support center?

A: いいえ / No.

Q: 名古屋駅から桜山キャンパスへ地下鉄で行くにはどの路線を使ったらよいですか。/ Which subway line should I take to get from Nagoya Station to the Sakura Mountain Campus?

A: 桜通線 / Sakura-dori Line.

Q: ピアが、3時間以上の会議に1回出席し、シドニーに一泊する場合、審査パネル会議への参加に対する出席費と滞在費として受け取る金額はいくらになりますか？朝食、昼食、夕食を1回ずつと諸費用も含むとします。/ How much would cost the business trip budget when go to Sydney, staying 1 night attending MTG, including meal and other expenses?

A: 1330豪ドルです / 1330 AUD.
(Caluculation of 415+165+750=1330)

Q: NIIが開発したファイル転送プロトコルMMCFTPを用いて東京ーデンバー間で転送実験をした図によると、青色で結ばれているのはどのような国際通信ですか？ According to the diagram illustrating the transfer experiment between Tokyo and Denver using the file transfer protocol MMCFTP developed by NII, what does the blue line represent in terms of international communication?

A: 図上で青色で示されている国際通信は、東京から香港を経由しシンガポールまでJGN/SingARENというネットワークで通信し、シンガポールからロサンジェルスを経由してデンバーまでinternet2/SingARENというネットワークで通信します。 / The blue lines on the diagram represent international communication that travels from Tokyo to Denver. Specifically, it goes through Hong Kong, then Singapore, using the JGN/SingAREN network, and from Singapore to Los Angeles, eventually reaching Denver via the internet2/SingAREN network.

Figure 1: JDocQA sample question answering instances in four question categories with the annotated bounding boxes of the supporting facts in red color.

open-access sources including multiple formats of documents: slides, reports, websites and pamphlets and manually annotating question-answer pairs on them. JDocQA consists of 11,600 question and answer pairs on the collected 5,504 documents as references for answering the question, four different question categories and 1,000 multi-page questions. Each question is designed to refer to both textual and visual components such as tables or figures by annotators. We also introduce *unanswerable* questions: questions that have no answer clues in the referenced documents. In experiments, we first present the effectiveness of finetuning LLM with our dataset. We also suggest that incorporating these *unanswerable* questions can contribute to mitigating *hallucination*, which is often observed during the generation by LLMs.

## 2. Related Work

**Multimodal question answering datasets.** Visual question answering is a task of question-answering given visual contexts such as images following textual queries (Malinowski and Fritz, 2014; Antol et al., 2015). Earlier VQA studies have been not limited to images alone but cover various forms of media such as textbook (Kembhavi et al., 2017), recipe (Yagcioglu et al., 2018), comic book (Iyyer et al., 2016), movie (Tapaswi et al., 2016). Among them, a document VQA, which is a task designed for text embedded in real-world images, has attracted a lot of attention toward a comprehensive understanding of documents from both the visual and textual sides. Currently, some useful document VQA datasets have been published, such as OCR-VQA (Mishra et al., 2019), TextVQA (Singh et al., 2019), and DocVQA (Mathew et al., 2021a), VisualMRC (Tanaka et al., 2021), WebSRC (Chen et al., 2021), and InfographicVQA (Mathew et al., 2022). Most of these studies concentrate on the single-image VQA where each question-answer pair has a single relevant image that always include sufficient information for question-answering. Unlike the single-image VQA, the ability to comprehend multiple pages or charts to answer questions is more practical for understanding the slides and documents people read in the daily work. To tackle such multi-image VQA, MultiModalQA (Talmor et al., 2021), MP-DocVQA (Tito et al., 2022) and SlideVQA (Tanaka et al., 2023) concentrate on the multi-hop and numerical reasoning while considering multimodal context similar to previous works (Rajpurkar et al., 2016; Yang et al., 2018; Dua et al., 2019). It is also notable that in document question answering Tang et al. (2023) proposed Universal Document Processing (UDOP), unifying vision, text, and layout of the input document through vision-text-layout Transformer.

**Text-based question answering in Japanese.** Some related tasks of text-based question answering for Japanese have been studied (Miyazaki and Shimizu, 2016; Yanaka and Mineshima, 2022; Takahashi et al., 2019; Kurihara et al., 2022). Miyazaki and Shimizu (2016) created a Japanese image captioning dataset, which is the Japanese version of the MS-COCO captions dataset, and demonstrated that using both bilingual datasets outperforms us-

| Category | Documents | QA | (1) Yes/No | (2) Factoid | (3) Numerical | (4) Open-ended | Multi-page | Unanswerable |
|---|---|---|---|---|---|---|---|---|
| Pamphlet | 1,715 | 4,025 | 605 | 748 | 660 | 2,012 | 46 | 671 |
| Slide | 1,640 | 3,276 | 545 | 593 | 507 | 1,631 | 448 | 449 |
| Report | 2,086 | 4,167 | 703 | 687 | 693 | 2,084 | 506 | 668 |
| Website | 67 | 132 | 2 | 24 | 6 | 100 | 0 | 0 |
| Total | 5,504 | 11,600 | 1,855 | 2,052 | 1,866 | 5,827 | 1,000 | 1,788 |

Table 1: Number of document styles and question-answer pairs by the four question categories, multi-page and unanswerable questions.

| Category | (1) Y/N | (2) Fact. | (3) Num. | (4) Open. |
|---|---|---|---|---|
| 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 |

Table 2: Average character length.

ing monolingual ones. Yanaka and Mineshima (2022) introduced a Japanese textual entailment dataset and highlighted that many existing models that have focused on English do not adequately account for Japanese language characteristics. Takahashi et al. (2019) proposed a QA dataset based on Japanese blogs related to driving, with the aim of creating a model that can understand the meaning of sentences or texts in Japanese. It is also notable that Miyao and Kawazoe (2013) and relevant Todai Robot Project[1] arranged the dataset of Japanese University entrance-exams. In math and physics subjects, their dataset includes limited multimodal contents in DTD file format, although it doesn't cover general domains as of JDocQA. Among the datasets related to the Japanese language, JGLUE (Kurihara et al., 2022) is similar to our work in terms of the aim of the datasets. JGLUE is a large-scale natural language understanding (NLU) benchmark purposed for the evaluation of LLM. It includes various tasks, such as text classification, sentence pair classification, and QA to assess Japanese comprehension. In contrast to JGLUE, our dataset offers a diverse range of question types which can be useful for instruction tuning and contains both Japanese text and image data, which can be used for multimodal models. Another key difference with JGLUE is that JDocQA incorporates unanswerable questions to help suppress hallucinations.

## 3. Dataset

### 3.1. Task Overview and Formulation

We consider generative question answering where a model generates a textual answer following the document context and textual question. For realistic applications of a wide range of user questions for documents, we prepare four categories of questions: (1) **yes/no**, (2) **factoid**, (3) **numerical**, and (4) **open-ended**. In yes/no questions, answers are "yes" or "no." In factoid questions, answers are some facts, such as named entities, that typically appear in the given documents. In numerical questions, answers are numeric values, often including some numerals (some units, e.g., *km* or Japanese numerals such as "個 (*objects*)" and "人 (*persons*)"). These numeric values are written in the documents or are calculated from other numbers in the documents. In open-ended questions, free-form responses are required. For such questions, we aim to assess complex comprehension abilities, such as the ability to form opinions or brief explanations based on the provided contexts and questions. Figure 1 presents samples of these four categories of questions. All examples include diverse images and question types related to some Japanese documents collected. We also include *unanswerable* questions for each question category. In the realistic applications of the question answering, no answers can be found in the referenced document. Therefore, it is expected that the correct responses for such questions are "not mentioned in the text." The prediction of the *unanswerable* questions is not addressed in previous Japanese question answering datasets such as Kurihara et al. (2022).

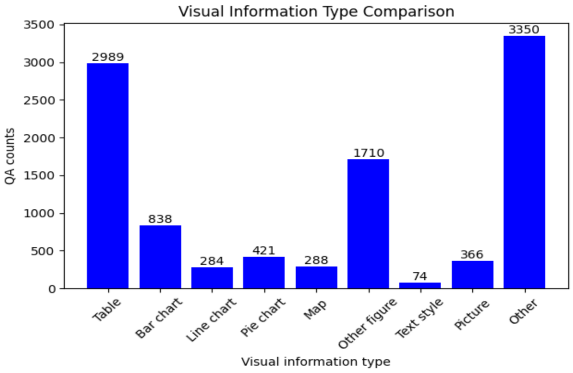


Figure 2: The number of visual information categories.

### 3.2. Dataset Statistics

JDocQA dataset comprises 5,504 files and 11,600 question-and-answer pairs in Japanese. The statistics of categorized question types are as follows: (1) yes/no questions: 1,855, (2) factoid questions: 2,052, (3) numerical questions: 1,866, (4) open-

---

[1] https://21robot.org/index-e.html

| Dataset | #Questions | #Images | #BBoxes | Language | Multihop |
|---|---|---|---|---|---|
| OCR-VQA (Mishra et al., 2019) | 1002k | 207k | - | English | - |
| DocVQA (Mathew et al., 2021a) | 50k | 12k | - | English | - |
| InfographicVQA (Mathew et al., 2022) | 5.9k | 30k | - | English | - |
| MP-DocVQA (Tito et al., 2022) | 46k | 48k | - | English | ✓ |
| SlideVQA (Tanaka et al., 2023) | 14.5k | 52k | 890k | English | ✓ |
| JDocQA (Ours) | 11.6k | 268k | 11k | Japanese | ✓ |

Table 3: The comparison of the document question answering datasets.

ended questions: 5,827. Additionally, 1,788 questions require referencing multiple pages to answer, and in 1,000 questions the correct answer is not mentioned in the text, as shown in Table 1. Some PDF documents contain both slide and report formats within the same documents. For such documents, we count them in both categories of the slide and report formats when we calculate the total number of the documents[2]. Table 2 represents the average length of the context, question, and answer in our dataset, and Figure 2 shows the category of the visual information referenced by question or answer in our dataset. The comparison of document question answering datasets are shown in Table 3.

### 3.3. Dataset Creation

The overall dataset creation and annotation process is presented in Figure 3.

**PDF collection.** We gather public documents, such as, municipality pamphlets and websites, that are created by Japanese governmental agencies or local governments. We manually collected PDF documents from open-access resources such as Japanese National Diet Library (NDL)'s digital collection, web archive projects (WARP)[3] and websites of Japanese government ministries. We manually gathered documents such as reports, pamphlets or websites that are published by public or quasi-public sectors, such as local governments or public universities through WARP. We also gather Japanese ministry documents such as slides and reports from their websites following the government agencies' policies. Those documents cover a wide range of topics, for instance, economic policies, education policies, labor issues, health and hygiene, agriculture, forestry, fisheries, culture and arts, history, related to governmental policy or policy guidelines, as well as the everyday affairs of local governments. These documents also include visual elements such as figures, tables, charts, pictures, or mandala charts, complex figures with a combination of texts and objects typically seen in the Japanese public administrative sector's official document. We classify these documents into four categories, namely, **pamphlet**, **slide**, **report**, and **website** considering the form of the documents.

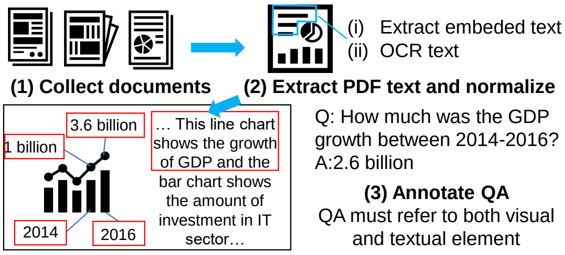

(i) Extract embeded text
(ii) OCR text
(1) Collect documents
(2) Extract PDF text and normalize
3.6 billion
1 billion
2014
2016
… This line chart shows the growth of GDP and the bar chart shows the amount of investment in IT sector…
Q: How much was the GDP growth between 2014-2016?
A:2.6 billion
(3) Annotate QA
QA must refer to both visual and textual element

Figure 3: Annotation process.

**Text Extraction & Normalization.** We extracted texts from PDF documents with PyPDF2[4]. We also notice that some PDF documents are probably created from paper scans, and we cannot extract embedded texts from such documents. Therefore, we extracted texts from the document page images by OCR (Optical Character Recognition) as an alternative source. After the text extraction or OCR, we removed mistakenly recognized symbols and emojis, or duplicated characters from texts when the same character continuously and repeatedly appeared more than five times.

**Annotation Procedure.** We ask 43 annotators in total for the question-answering pairs annotation on documents. As documents include rich textual and visual elements (e.g., graphs, charts, maps, illustrations, and a mix of vertical and horizontal written text), we made question answer pairs that are related to both textual and visual information. We ask annotators to write up two to four question-answer annotations in each document. We also ask not to use any AI-tools such as OpenAI ChatGPT during the annotation process. Each question is accompanied with the supporting facts as marked in red in Figure 1 and Figure 3. We classify a subset of questions that have multiple supporting facts in multiple pages as *multi-page* questions. Multi-page questions are considerably difficult from their single-page counterparts. For unanswerable questions, we ask annotators to write questions that lack supporting facts in the documents, making them impossible to answer based on the given

---

[2] The number of total documents is not the sum of the number of the subcategories in Table 1 due to this counting.

[3] `https://warp.ndl.go.jp/`

[4] We also examined PyMuPDF. However, the quality of the extracted texts was not changed greatly.

documents.

**Visual inputs and bounding boxes.** We prepared three types of images for visual inputs for multimodal models. The first type of images are those of the whole page of the documents including the annotated question answering pairs. The second type of images are those cropped by bounding boxes on which annotators based their answers such as tables or figures of the pages. When multiple bounding boxes are annotated to a single question-answer pair, multiple cropped images are combined together into a single image here. The third type of images are blank (white) images that are used for ablation studies.

# 4. Experiments

## 4.1. Question-Answering Task

Our dataset aims to evaluate the question answering ability following the document contexts, including textual and visual information, and questions via open-ended text generation. As discussed in Sec. 3.1, our dataset consists of four forms of questions: yes/no[5], factoid, numerical, and open-ended. All of these four category questions include *unanswerable* questions that are not answerable solely from the given document file. Models are expected to generate answers in open-ended text generation in any of these question types. Textual model inputs, or simply *prompts*, consist of the embedded texts or OCR results of documents as described in Sec. 3.3 and the questions. We also include the answer-format guidelines such as "please answer in Yes/No form", "please answer the fact that is referred to in the document", "please answer by numerical information from the document" and "please write the answer in open-ended format" into the prompts depending on the four question categories. For *unanswerable* questions, we prepare a special answer for all question categories: "本文中に記載がありません (*not mentioned in the text*)."

## 4.2. Models

We conduct experiments with both text-input models and multimodal models of text and vision inputs. For model training, we use supervised finetuning. The best hyperparameters are searched with train and validation sets, then the model performance is evaluated with the best hyperparameters.

**Models with text input.** We adapted up to 13 billion (13B) model parameter scale Japanese large language models for experiments. We experimented with the following representative Japanese models that take only textual inputs: rinna japanese-gpt2-medium[6], japanese-gpt-4B-8k[7], rinna japanese-gpt-1B[8], Cyberagent OpenCALM-7B[9], Matsuo-Lab weblab-10b[10], PFNet PLaMo-13B[11], Stability AI Japanese-StableLM-Base-Alpha-7B[12], and Stability AI Japanese-StableLM-Instruct-Alpha-7B[13]. We also include multilingual large language model of Llama-2-7B [14]. We trained and evaluated models with 1024 token length for fair comparisons and computational efficiency except rinna japanese-gpt-4B-8k which is trained with 8192 tokens at most. For analyses with longer token lengths, we train rinna japanese-gpt-4B-8k model with 2048, 4096, and 8192 tokens.

**Models with multimodal input.** The purpose of JDocQA is to analyze documents with textual and visual perceptions. To assess the impact of using both images and text on the JDocQA dataset, we applied multimodal models that take inputs from both images and texts. We used Stability AI Japanese-StableLM-Instruct-Alpha-7B[15], a Japanese version of InstructBLIP (Dai et al., 2023; Li et al., 2023a,c) for this purpose, as they are applicable to Japanese text and image inputs. We trained and evaluated this model with 512 token lengths following its max capacity. We develop three different models for with three different visual inputs as explained in Sec. 3.3. The first model takes visual inputs of a blank image that is always the same white image of the 800x600 pixel size as ablation study. The second model takes an image of a whole document page that are related to the question-answering in the annotation. These images are also scaled to 800-pixel width. The third model take inputs of the images following the annotated supporting facts. Following the annotated

---

[5] The chance rate of yes/no questions including *unanswerable* is 61.57 when the model always marks "yes."

[6] `https://huggingface.co/rinna/japanese-gpt2-medium`

[7] `https://huggingface.co/rinna/bilingual-gpt-neox-4b-8k`

[8] `https://huggingface.co/rinna/japanese-gpt-1b`

[9] `https://huggingface.co/cyberagent/open-calm-7b`

[10] `https://huggingface.co/matsuo-lab/weblab-10b`

[11] `https://huggingface.co/pfnet/plamo-13b`

[12] `https://huggingface.co/stabilityai/japanese-stablelm-base-alpha-7b`

[13] `https://huggingface.co/stabilityai/japanese-stablelm-instruct-alpha-7b`

[14] `https://huggingface.co/meta-llama/Llama-2-7b-hf`

[15] `https://huggingface.co/stabilityai/japanese-stablelm-instruct-alpha-7b`

|  | Validation set |  |  |  |  | Test set |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg.** | **(1) Y/N** | **(2) Fact.** | **(3) Num.** | **(4) Open.** | **Avg.** | **(1) Y/N** | **(2) Fact.** | **(3) Num.** | **(4) Open.** |
| *Evaluated with all instances.* |  |  |  |  |  |  |  |  |  |  |
| gpt-3.5-turbo-16k | 19.86 | 47.89 | 7.85 | 7.97 | 15.75 | 20.62 | 50.29 | 7.44 | 11.11 | 13.64 |
| gpt-4 | 17.96 | 34.73 | 9.42 | 8.51 | 19.17 | 19.47 | 43.19 | 6.51 | 11.11 | 17.07 |
| *Evaluated without "unanswerable."* |  |  |  |  |  |  |  |  |  |  |
| gpt-3.5-turbo-16k | 22.72 | 57.23 | 8.82 | 9.20 | 15.63 | 23.07 | 58.21 | 8.08 | 12.5 | 13.49 |
| gpt-4 | 20.90 | 41.50 | 10.58 | 9.81 | 21.72 | 22.03 | 50.00 | 7.07 | 12.5 | 18.57 |
| *Models trained with all training instances and evaluated all instances.* |  |  |  |  |  |  |  |  |  |  |
| rinna gpt2-medium-336M | 21.33 | 63.15 | 7.32 | 4.78 | 17.51 | 19.41 | 62.13 | 4.65 | 8.18 | 15.99 |
| rinna gpt-1B | 23.79 | 58.42 | 10.99 | 6.38 | 22.27 | 20.46 | 59.76 | 5.58 | 8.77 | 18.13 |
| rinna bi-4B-8k (8192 tok.) | 26.35 | 55.26 | 14.65 | 13.29 | 24.93 | 23.02 | 62.13 | 8.83 | 11.11 | 20.57 |
| OpenCALM-7B | 21.65 | 47.36 | 14.65 | 5.31 | 20.81 | 18.33 | 43.78 | 11.62 | 9.94 | 16.03 |
| weblab-10B | 19.20 | 46.31 | 9.42 | 7.97 | 17.13 | 16.94 | 47.92 | 10.23 | 8.18 | 13.24 |
| PLaMo-13B | 25.79 | 55.26 | 15.18 | 15.42 | 22.92 | 20.33 | 53.84 | 10.69 | 7.01 | 18.21 |
| StableLM Base-Al.-7B | 32.92 | 67.89 | 21.98 | 18.61 | 29.62 | 29.71 | 70.41 | 15.81 | 22.22 | 25.51 |
| StableLM Inst.-Al.-7B | 33.80 | 67.36 | 20.94 | 20.21 | 31.39 | 29.56 | 72.78 | 16.27 | 21.05 | 24.75 |
| Llama2-7B | 30.29 | 60.00 | 20.41 | 15.42 | 28.59 | 27.01 | 61.53 | 17.20 | 18.71 | 23.29 |
| *Models trained with all training instances while evaluated without "unanswerable."* |  |  |  |  |  |  |  |  |  |  |
| rinna gpt2-medium-336M | 22.22 | 69.81 | 4.70 | 3.68 | 18.71 | 19.61 | 65.75 | 4.54 | 6.57 | 16.31 |
| rinna gpt-1B | 22.84 | 62.26 | 7.05 | 5.52 | 21.14 | 20.18 | 64.38 | 4.04 | 8.55 | 17.40 |
| rinna bi-4B-8k (8192 tok.) | 24.09 | 52.83 | 12.94 | 10.42 | 23.10 | 21.17 | 62.32 | 6.56 | 7.89 | 19.11 |
| OpenCALM-7B | 20.75 | 50.31 | 12.35 | 4.90 | 19.20 | 17.53 | 46.57 | 9.59 | 8.55 | 15.10 |
| weblab-10B | 17.74 | 45.91 | 8.82 | 6.13 | 15.34 | 16.20 | 50.68 | 9.09 | 6.57 | 12.17 |
| PLaMo-13B | 22.66 | 54.08 | 10.58 | 9.81 | 20.76 | 18.35 | 50.68 | 8.58 | 5.26 | 16.87 |
| StableLM Base-Al.-7B | 31.55 | 70.44 | 18.82 | 15.95 | 28.24 | 28.33 | 71.91 | 12.62 | 21.05 | 24.32 |
| StableLM Inst.-Al.-7B | 31.74 | 69.81 | 17.05 | 15.95 | 29.55 | 28.66 | 76.02 | 12.62 | 19.07 | 24.40 |
| Llama2-7B | 28.78 | 62.26 | 17.05 | 14.72 | 26.44 | 25.70 | 65.06 | 12.62 | 18.42 | 21.87 |
| *Models trained without "unanswerable" while evaluated with all instances.* |  |  |  |  |  |  |  |  |  |  |
| rinna gpt2-medium-336M | 15.91 | 52.10 | 3.14 | 3.72 | 12.11 | 17.81 | 63.31 | 4.18 | 5.26 | 13.60 |
| rinna gpt-1B | 17.66 | 45.26 | 6.80 | 2.65 | 17.04 | 17.59 | 51.47 | 6.04 | 5.26 | 15.76 |
| rinna bi-4B-8k (8192 tok.) | 23.44 | 57.89 | 12.04 | 9.57 | 20.33 | 23.01 | 69.23 | 7.90 | 9.94 | 19.27 |
| OpenCALM-7B | 18.94 | 42.63 | 8.90 | 3.72 | 19.44 | 16.95 | 42.01 | 7.44 | 6.43 | 16.33 |
| weblab-10B | 20.43 | 50.00 | 9.42 | 6.91 | 18.70 | 17.96 | 52.07 | 6.51 | 8.18 | 15.34 |
| PLaMo-13B | 22.04 | 60.00 | 7.85 | 9.57 | 18.23 | 21.11 | 64.49 | 8.83 | 11.11 | 16.31 |
| StableLM Base-Al.-7B | 27.02 | 63.68 | 14.65 | 12.76 | 23.61 | 25.68 | 68.63 | 12.09 | 17.54 | 20.94 |
| StableLM Inst.-Al.-7B | 27.22 | 61.57 | 15.70 | 14.36 | 23.84 | 26.25 | 70.41 | 15.34 | 16.37 | 20.74 |
| Llama2-7B | 30.15 | 63.68 | 18.32 | 13.82 | 28.30 | 28.25 | 73.96 | 11.62 | 16.95 | 24.68 |
| *Models trained without "unanswerable" and evaluated without "unanswerable" instances.* |  |  |  |  |  |  |  |  |  |  |
| rinna gpt2-medium-336M | 18.75 | 62.26 | 3.52 | 4.29 | 14.33 | 20.12 | 73.28 | 4.54 | 5.92 | 15.41 |
| rinna gpt-1B | 21.15 | 54.08 | 7.64 | 3.06 | 21.17 | 20.02 | 59.58 | 6.56 | 5.92 | 18.20 |
| rinna bi-4B-8k (8192 tok.) | 27.63 | 69.18 | 13.52 | 11.04 | 24.26 | 25.95 | 80.13 | 8.58 | 11.18 | 21.79 |
| OpenCALM-7B | 20.55 | 50.94 | 10.00 | 4.29 | 19.67 | 17.97 | 48.63 | 8.08 | 7.23 | 16.32 |
| weblab-10B | 22.25 | 59.74 | 10.58 | 7.97 | 18.55 | 19.06 | 60.27 | 7.07 | 9.21 | 15.06 |
| PLaMo-13B | 26.48 | 71.69 | 8.82 | 11.04 | 22.74 | 24.29 | 74.65 | 9.59 | 12.50 | 19.34 |
| StableLM Base-Al.-7B | 32.30 | 76.10 | 16.47 | 14.72 | 29.15 | 29.21 | 79.45 | 13.13 | 19.73 | 24.14 |
| StableLM Inst.-Al.-7B | 32.41 | 73.58 | 17.64 | 16.56 | 29.16 | 29.75 | 81.50 | 16.66 | 18.42 | 23.69 |
| Llama2-7B | 33.11 | 76.10 | 20.58 | 15.95 | 28.84 | 30.57 | 85.61 | 12.62 | 19.07 | 25.47 |

Table 4: Results of all finetuned models and OpenAI GPT zeroshot. **Avg.** is weighted average of scores.

bounding boxes, we crop the referenced regions of the page images, combine the bounding boxes and scale them for the model visual input. As some questions have several annotated supporting facts to their answers, the combined image may contain more than one region of the annotated bounding boxes. All of these multimodal models also take textual prompts that are similar to the text-input models.

**OpenAI GPT baselines.** We also present OpenAI GPT performances as baselines. Here we use gpt-3.5-turbo-16k and gpt-4 models[16]. They take similar prompts to those of text-input models. However, as they are the zero-shot models for our task, we observed they are quite sensitive to the prompts. To improve their performance, we manually tune the prompts for OpenAI GPT models. We avoid finetuning OpenAI GPT models although finetuning them may greatly improve performances due to the following reasons. First, our purpose is to develop local models that work on limited computational resources, Second, the details of finetuning are unavailable for these models, and finally due to the API cost issues.

## 4.3. Evaluation Methods

For Yes/No, factoid, and numerical questions, we used the exact match metric after trimming trivial differences such as the presence or absence of punctuation marks or Japanese suffix phrases such as "です (*is*)" by simple rules. We have also examined the variations of exact match, such as the ratio of whether model prediction phrases are included in the correct answer phrases or not. However, we realize that this metric performs quite similarly to the exact match evaluation and the difference is typically less than 10 question-answer pairs in the validation set. For open-ended questions, answers are typically long, e.g., the average length of answers is 65.97 characters in Japanese, and hence the exact match does not work for evaluation. Therefore,

---

[16] Latest model at Oct. 9 2023.

we used BLEU score[17] tokenized by MeCab[18] for automatic evaluation of open-ended questions.

## 4.4. Experimental Settings

Our dataset includes *unanswerable* questions in all question categories. While it is expected that fine-tuning models with unanswerable instances may harness models to surpass illusion answers known as hallucinations, detecting unanswerable questions is also notoriously difficult as of Rajpurkar et al. (2018). Therefore we prepare two types of models for all base models experimented: models finetuned with all question answering pairs including unanswerable instances and models finetuned without unanswerable instances. In the evaluation, we similarly prepare two separate validation and test sets: the standard validation and test sets that consists of all question-answering pairs, and the smaller validation and test sets where unanswerable questions are removed.

## 4.5. Results

Table 4 presents the performance of text-input models for all question types on valid and test splits.

**Models trained with all instances.** We compare the first and third blocks in Table 4. They are the results of all JDocQA instances by zero-shot and finetuned models that are trained with all JDocQA training instances including “unanswerable questions”. We realize that fine-tuned models outperformed gpt-3.5 and gpt-4 results especially when the model size is larger. Next, we compare the second and fourth blocks in Table 4. They are the same models with previous block evaluations on the JDocQA without unanswerable questions. We observe a similar tendency to all instances evaluations, suggesting that models finetuned with unanswerable instances perform similar performances in both answerable and unanswerable questions. Among them, StableLM models perform best despite their parameter size of 7B. The rinna bi-4B-8k model also performs well despite its parameter size. We attribute this to its token length size of 8192 and will discuss this later. We notice that (1) yes/no questions are relatively easy although they do not have much effect on the averaged score (Avg.) that mostly follows the most common question category of (4) open-ended.

**Models trained without unanswerable.** As explained in Sec. 4.4, we also prepared models fine-tuned without *unanswerable* questions in the training instances. We present the evaluation results including and excluding *unanswerable* questions in the fifth and sixth blocks in Table 4. It is quite interesting when we compare the third and fifth block results in Table 4 as they share the same evaluation set including *unanswerable* questions while the models are finetuned with and without the *unanswerable* instances. Comparing the third and fifth block results, we notice almost all models finetuned with *unanswerable* questions perform better than their *answerable*-only finetuned counterparts in the averaged scores. Exceptions are the OpenCALM-7B, weblab-10B and Llama2-7B models, which we will discuss in the next paragraph. We attribute this is due to the concept of *hallucination,* where models generate answers that do not appear in context texts. We will present an example of this in the qualitative analysis paragraph. Doping *unanswerable* instances may contribute to harnessing *hallucination* in this experimental comparison.

| | Test set | | | | |
|---|---|---|---|---|---|
| **Model** | **Avg.** | **(1) Y/N** | **(2) Fact.** | **(3) Num.** | **(4) Open.** |
| *Trained all and evaluated all.* | | | | | |
| 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 |
| *Trained all while evaluated w/o “unanswerable.”* | | | | | |
| InstBLIP (blank) | 25.12 | 65.75 | 10.60 | 17.10 | 21.68 |
| InstBLIP (img) | 25.74 | 69.17 | 11.61 | 15.13 | 22.12 |
| InstBLIP (bbox) | 27.99 | 78.76 | 14.14 | 17.76 | 22.16 |
| *Trained w/o “unanswerable” while evaluated all.* | | | | | |
| InstBLIP (blank) | 23.13 | 66.27 | 12.55 | 11.69 | 18.21 |
| InstBLIP (img) | 25.01 | 71.59 | 12.09 | 16.37 | 19.19 |
| InstBLIP (bbox) | 29.00 | 78.10 | 14.88 | 19.29 | 23.19 |
| *Trained w/o “unanswerable” and evaluated w/o “unanswerable.”* | | | | | |
| 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 |

Table 5: Results of multimodal input models. **Avg.** is weighted average.

| | Test set | | | | |
|---|---|---|---|---|---|
| **Token length** | **Avg.** | **(1) Y/N** | **(2) Fact.** | **(3) Num.** | **(4) Open.** |
| *Trained all and evaluated all.* | | | | | |
| 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 |
| *Trained w/o “unanswerable” and evaluated w/o “unanswerable.”* | | | | | |
| 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 |

Table 6: Results of rinna bi-4B-8k models with different token length. **Avg.** is weighted average.

**OpenCALM-7B and weblab-10B do not predict questions as *unanswerable* so much.** In the third block of Table 4, we notice interesting phenomena: OpenCALM-7B and weblab-10B do not perform well despite their parameter size. When we closely check these models’ outputs, we realize these models, finetuned with all instances, predict “本文中に記載がありません (*not mentioned in the text*)” much less than other models. OpenCALM-

---

[17] https://github.com/mjpost/sacrebleu

[18] https://pypi.org/project/mecab-python3/

| | Test set | | | |
|---|---|---|---|---|
| **Model** | **Pamphlet** | **Slide** | **Report** | **Website** |
| *Trained all and evaluated all.* | | | | |
| rinna gpt2-med-336M | 18.62 | 16.32 | 14.57 | 3.09 |
| rinna gpt-1B | 16.66 | 15.10 | 15.08 | 4.72 |
| rinna bi-4B-8k (8192) | 21.81 | 16.73 | 18.41 | 3.53 |
| OpenCALM-7B | 15.19 | 15.10 | 13.04 | 2.56 |
| weblab-10B | 14.70 | 17.55 | 13.04 | 2.69 |
| PLaMo-13B | 21.56 | 13.06 | 15.85 | 2.90 |
| Base-Al.-7B | 26.96 | 20.40 | 24.04 | 4.64 |
| Inst-Al.-7B | 27.69 | 23.26 | 23.01 | 5.71 |
| InstBLIP-Al (blank) | 25.49 | 22.04 | 21.48 | 3.79 |
| InstBLIP-Al (img) | 25.00 | 21.63 | 23.78 | 3.75 |
| InstBLIP-Al (bbox) | 25.00 | 22.04 | 22.50 | 4.20 |

Table 7: Detailed file-type result. "**Website**" is included only in test set as an out-of-domain set.

7B and weblab-10B predicts 11.9% and 13.7% of all instances are *unanswerable* while other models in the third block of Table 4 predict around or more than 20%. As they are trained with the same dataset, we suspect this is due to their pretraining. It is also notable that predicting questions as *unanswerable* is a difficult task for models and often affects the overall performance.

**Multimodal model results.** We present the results of models with multimodal inputs models of StableLM-InstructBLIP-Alpha in Table 5. The model performances are enhanced especially when we use cropped images of referenced tables or figures (bbox). We also notice that the model with black image inputs performs close to visual models to some extent, suggesting the effectiveness of textual inputs in our task. We also carefully note that the max token length of StableLM-InstructBLIP-Alpha is 512, which can limit the textual understanding abilities of current multimodal models.

**Token length dependency.** We survey the token length effects on the performance. For this purpose, we finetuned three models of different token lengths of rinna bi-4B-8k, e.g., 2048, 4096, and 8192 respectively for both all training instances and without *unanswerable* conditions. We present the performances in Table 6. We notice the finetuning token length surely affects the final result, although we also notice finetuning models with long token length are much more computationally costly than those of short token length models, which can be the reason for the good performance of the rinna bi-4B-8k (8192 tokens) model in Table 4.

**Detailed Analyses of document-types.** Table 7 presents the performance comparisons between each file type. File types are classified into Japanese pamphlets such as public relations booklets or magazines, slides such as presentation materials, and report documents including figures and tables. We also prepare the out-of-domain test set of website scans where models still perform worse.

| **Model** | **Human Evaluation ↑** |
|---|---|
| *Trained all and evaluated all.* | |
| PLaMo-13B | 1.24 |
| StableLM Instruct-Alpha-7B | 1.49 |
| StableLM InstructBLIP-Alpha (blank) | 1.04 |
| StableLM InstructBLIP-Alpha (img) | 1.25 |

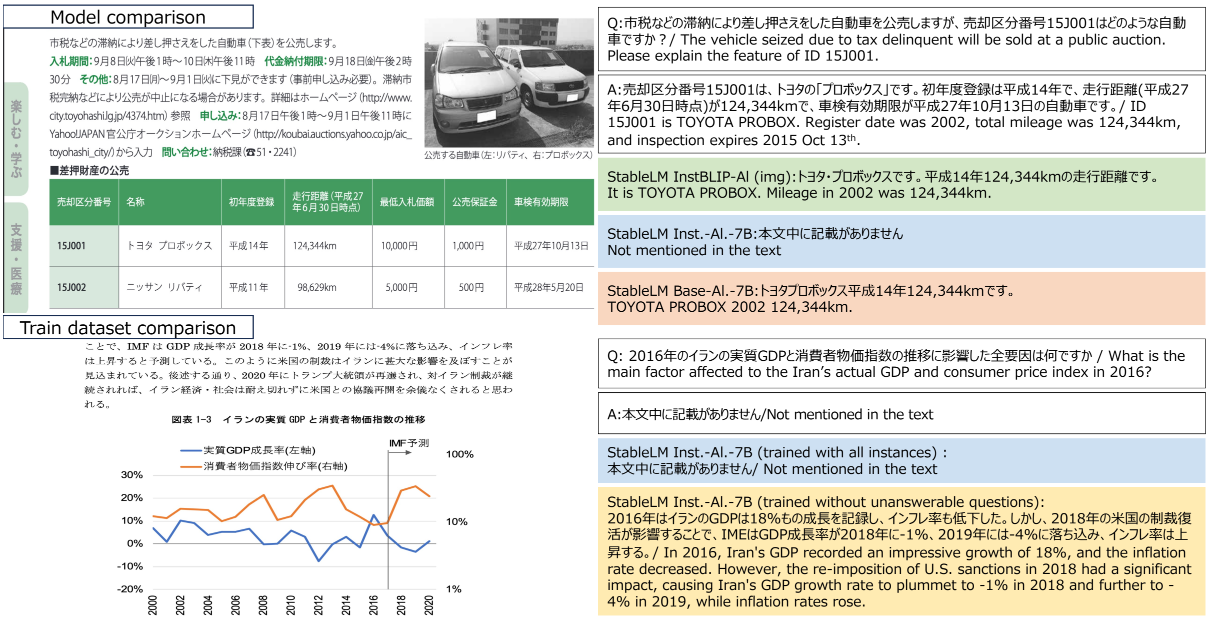
Table 8: Human evaluation on the sampled set.

**Qualitative Analysis.** Figure 4 presents two example instances with the models' generations. In the top example, we present the question, annotated answer, and generation from three models trained with questions including *unanswerable*: StableLM InstructBLIP-Alpha (img), StableLM Instruct-Alpha-7B and StableLM Base-Alpha-7B respectively. StableLM InstructBLIP-Alpha (img) can *see* the alignments of the table via visual inputs and generate reasonable descriptions for the car. StableLM Base-Alpha also generates a similar answer while it cannot attribute the running mileage of 123,334 presented in the table. At the bottom of Figure 4, we presented the comparison of two finetuned models that originate from the same pretrained model and are trained with and without *unanswerable* questions. The all instances model accurately predicts that there are no answers written in the text while the model without *unanswerable* questions falsely generates the open-ended answers, causing the phenomena known as hallucination.

**Human evaluation.** Finally, we performed the human evaluation on the sampled results for some representative models. We sampled 100 open-ended questions from the test set for this purpose. We choose two text-input models of PLaMo-13B and StableLM Instruct-Alpha-7B. We also choose multimodal input models of StableLM InstructBLIP-Alpha with back and image inputs. We ask an annotator to attach scores from 0 to 2 for two criteria: whether the generated answers include the annotated answer and whether the generated answers do not include wrong statements as questions' answers. The results are presented in Table 8. We notice that StableLM Instruct-Alpha-7B outperforms PLaMo-13B and image model outperforms its blank image counterpart.

## 5. Conclusion

We introduced the JDocQA dataset concentrating the integration of both visual and linguistic cues in question answering in Japanese. We incorporated *unanswerable* questions from given documents, which we confirmed is effective for harnessing the hallucinated generation to some extent in our experiments. Our detailed evaluations revealed the effectiveness of our dataset in a wide range of question categories from yes/no to open-ended and that the prediction of *unanswerable* questions can be



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㈪～9月1日㈫に下見ができます（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.city.toyohashi.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（☎51・2241）

公売する自動車（左：リバティ、右：プロボックス）

**■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** | ニッサン　リバティ | 平成11年 | 98,629km | 5,000円 | 500円 | 平成28年5月20日 |

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、トヨタの「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-Al (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.-Al.-7B:本文中に記載がありません
Not mentioned in the text

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

Train dataset comparison

ことで、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・社会は耐え切れずに米国との協議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**

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全要因は何ですか / 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.-Al.-7B (trained with all instances) :
本文中に記載がありません/ Not mentioned in the text

StableLM Inst.-Al.-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、IME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.

Figure 4: Qualitative analysis of open-ended question answering.

a clue to improve model performances, illustrating the effectiveness of the JDocQA dataset in realistic applications where multiple categories of questions are feasible and some extent of questions do not have explicit written answers in the documents.

## 6. Acknowledgments

This work was supported by JSPS Grant-in-Aid for Young Scientists (#22K17983), JSPS Fostering Joint International Research (A) (#22KK0184) and by JST PRESTO (#JPMJPR20C2).

## Ethical Statement & Limitation

In data collection, we gather document PDF files and webpages from Japanese National Diet Library (NDL)’s digital collection, web archive projects (WARP), and websites of Japanese government ministries. Administrative PDFs, documents, pamphlets, or websites published by local governments or universities are gathered through WARP. We carefully avoid private documents and choose considerably public documents published by public or quasi-public sectors for the publicity of our dataset usage. All of the documents and webpages are publicly available online and we follow our institutional rules to gather them. We follow our institutional rules and also consult external advisors for data collection processes.

We assume our datasets are useful for both research and development of generative language models and their applications for Japanese document question answering. We also consider our dataset with *unanswerable* questions can contribute to harnessing the hallucination problem of large language models. However, this doesn’t mean that the fintuned models with *unanswerable* questions do not perform hallucinations at all.

## 7. Bibliographical References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Ankan Bansal, Yuting Zhang, and Rama Chellappa. 2020. Visual question answering on image sets. In *ECCV 2020*.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and

Kai Yu. 2021. WebSRC: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4173–4185, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L. Boyd-Graber, Hal Daumé, and Larry S. Davis. 2016. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. 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)*.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Dongxu Li, Junnan Li, and Steven CH Hoi. 2023a. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. 2023b. Lavis: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021a. Docvqa: A dataset for vqa on document images. In *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pages 2199–2208.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. *DocVQA: A dataset for VQA on document images*, Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, pages 2199–2208. Institute of Electrical and Electronics Engineers Inc. Funding Information: We thank Amazon for supporting the annotation effort, and Dr. R. Manmatha for many useful discussions and inputs. This work is partly supported by MeitY, Government of India, the project TIN2017-89779-P, an Amazon AWS Research Award and the CERCA Programme. Publisher Copyright: © 2021 IEEE.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. 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)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Yusuke Miyao and Ai Kawazoe. 2013. University entrance examinations as a benchmark resource for NLP-based problem solving. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1357–1365, Nagoya, Japan. Asian Federation of Natural Language Processing.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Pro-*

*ceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. Technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4634–4642.

Norio Takahashi, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2019. Machine comprehension improves domain-specific Japanese predicate-argument structure analysis. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 98–104, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *AAAI*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Chao-Yue Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19254–19264.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rubèn Pérez Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document collection visual question answering. In *IEEE International Conference on Document Analysis and Recognition*.

Rubèn Pérez Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for multi-page docvqa. *ArXiv*, abs/2212.05935.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. 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, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.