JDocQA：用于生成语言模型的日语文档问答数据集

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. 介绍

透彻了解由文本和图形元素（如幻灯片、报告、网页和小册子）组成的文档对于处理多媒体文档并回答有关此类文档的一些问题的智能代理至关重要。为了实现对此类文档或图像中的文本和视觉元素的联合理解，已经研究了文档的视觉理解，包括书籍封面（**?**）、带有字符的场景图像（**?**）、网页（**?**）、表格（**?**）和幻灯片（**?**）。这些数据集受到了极大的关注，因为文档是英语领域各种工业、公共和私营部门的常见形式。同样值得注意的是，尽管文档视觉问答任务在行业中很重要，但它们仍然非常困难，因为它们严重依赖文本和视觉模式，因为文档通常包括数字、图表和插图上文本的复杂视觉对齐。特别是在文档问答中，需要模型连接多种模态来找出答案。在相当有限的数据集中，需要视觉和文本信息来回答文档上的问题。还有一个问题是，尽管这些任务很重要，但这些数据集的主要关注点仅限于英语领域，而其他语言的数据集结构仍然有限。作为文档问答任务，日语文档与英语文档相比具有几个特点。日语文档处理的主要难点之一在于日语的两种官方书写风格：一种是从左到右的水平风格，另一种是上到下的垂直风格，这需要数据集中的两种书写风格理解。

如今，生成式大型语言模型（LLM）和多模态模型取得了重大进展。GPT-4（**?**）允许在语言相关甚至多模态任务中实现零样本应用。InstructBLIP（**?**）接受文本和图像输入，并按照文本提示生成文本，例如图像标题或视觉问答。LLM 的成功也引发了几个公开的日语 LLM 的竞争性发展。LLM 的指令调校可以提高其遵守某些领域或用法的能力，使它们更适合特定的应用程序，而不是保持其对语言的一般理解和有限的专业知识的能力（**??**）。虽然已经有许多尝试来微调教学以适应高度技术和专业的适应（**??**），但仍然缺乏充分准备的高质量视觉问答数据集，这些数据集可用于基于生成语言模型的问答，特别是在英语领域之外开发。

为了满足对大规模和完全注释的日语文档问答数据集的需求，我们引入了 JDocQA 数据集，从开放获取资源（包括多种格式的文档）收集 PDF 样式的日语文档，包括幻灯片、报告、网站和小册子，并在其上手动注释问答对。JDocQA 由收集的 5,504 个文档上的 11,600 个问答对组成，作为回答问题的参考、四个不同的问题类别和 1,000 个多页问题。每个问题都旨在引用注释者的文本和视觉组件，例如表格或图形。我们还引入了无法回答的问题：在参考文档中没有答案线索的问题。在实验中，我们首先展示了使用数据集微调 LLM 的有效性。我们还建议，纳入这些无法回答的问题有助于减轻幻觉，这在 LLM 生成过程中经常观察到。

2. 相关工作

多模态问答数据集。视觉问答是在给定的视觉上下文（例如文本查询之后的图像）进行问答的任务（**??**）。早期的 VQA 研究并不局限于图像，而是涵盖了各种形式的媒体，如教科书（**?**）、食谱（**?**）、漫画书（**?**）、电影（**?**）。其中，文档 VQA 是为嵌入在真实世界图像中的文本而设计的任务，引起了人们对从视觉和文本两个方面对文档的全面理解的广泛关注。目前已经发布了一些有用的文档 VQA 数据集，如 OCR-VQA（**?**）、TextVQA（**?**）、DocVQA（**?**）、VisualMRC（**?**）、WebSRC（**?**）、InfographicVQA（**?**）。这些研究大多集中在单图像 VQA 上，其中每个问答对都有一个相关的图像，该图像始终包含足够的问答信息。与单图像 VQA 不同，理解多个页面或图表以回答问题的能力对于理解人们在日常工作中阅读的幻灯片和文档更实用。

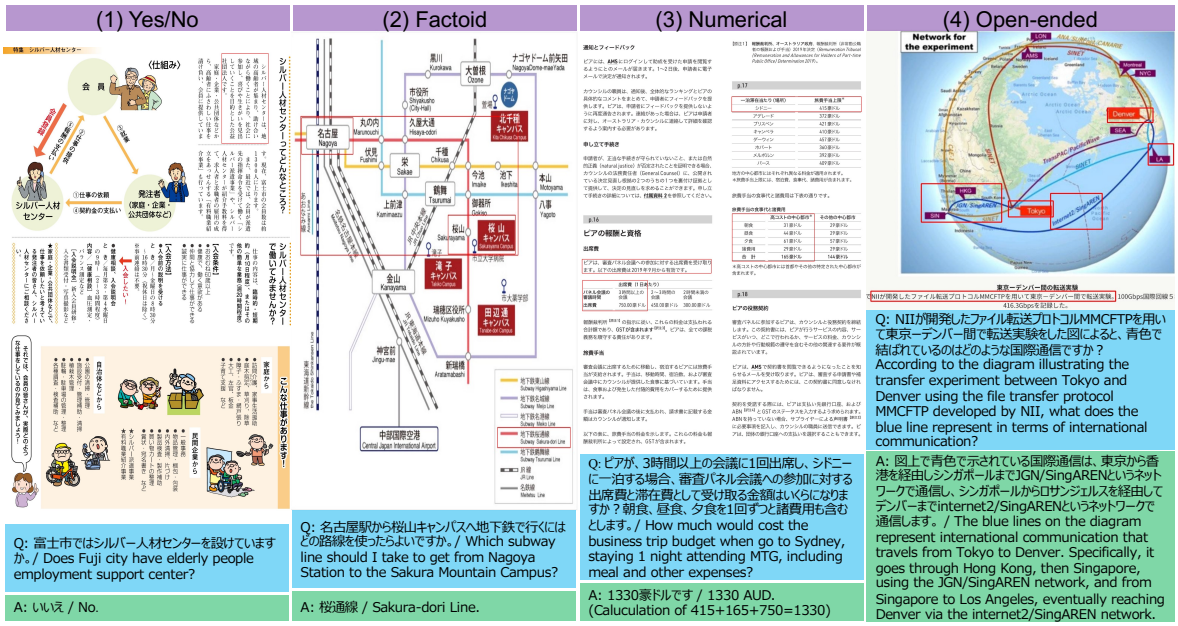

Figure 1: 四个问题类别中的 JDocQA 示例问答实例，支持事实的注释边界框为红色。

为了解决这种多图像 VQA，MultiModalQA (**?**) 、MP-DocVQA (**?**) 和 SlideVQA (**?**) 专注于多跳和数值推理，同时考虑与以前的工作类似的多模态上下文 (**???**) 。同样值得注意的是，在文档问答中，**?** 提出了通用文档处理（UDOP），通过视觉-文本布局转换器统一输入文档的视觉、文本和布局。

基于文本的日语问答。研究了日语文本问答的一些相关任务 (**????**) 。**?** 创建了一个日语图像字幕数据集，这是 MS-COCO 字幕数据集的日语版本，并证明使用两种双语数据集都优于使用单语数据集。**?** 介绍了一个日语文本蕴涵数据集，并强调许多专注于英语的现有模型没有充分考虑日语特征。**?** 提出了一个基于与驾驶相关的日语博客的 QA 数据集，目的是创建一个可以理解日语句子或文本含义的模型。值得一提的是，**?** 相关的东大机器人项目 [1] 安排了日本大学入学考试的数据集。在数学和物理科目中，他们的数据集包括 DTD 文件格式的有限多模态内容，尽管它不包括 JDocQA 的一般域。在与日语相关的数据集中，JGLUE (**?**) 在数据集的目标方面与我们的工作相似。JGLUE 是一个大规模的自然语言理解（NLU）基准测试，旨在评估 LLM。它包括各种任务，例如文本分类、句子对分类和评估日语理解的 QA。与 JGLUE 相比，我们的数据集提供了多种问题类型，可用于指令调整，并包含可用于多模态模型的日语文本和图像数据。JGLUE 的另一个关键区别是 JDocQA 包含无法回答的问题来帮助抑制幻觉。

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

3. 数据

3.1. 任务概述和制定

我们考虑生成性问答，其中模型根据文档上下文和文本问题生成文本答案。为了实际应用文档的各种用户问题，我们准备了四类问题：(1) 是/否，(2) 事实，(3) 数字，和 (4) 开放式。在是/否问题中，答案是“是”或“否”。在类事实问题中，答案是一些事实，例如命名实体，通常出现在给定文档中。在数值问题中，答案是数值，通常包括一些数字（一些单位，例如公里或日语数字，例如“UTF8ipxm 個（物体）”和“UTF8ipxm 人（人）”）。这些数值写在文档中，或根据文档中的其他数字计算得出。在开放式问题中，需要自由形式的回答。对于此类问题，我们旨在评估复杂的理解能力，例如根据提供的上下文和问题形成意见或简要解释的能力。图 **??** 给出了这四类问题的样本。所有示例都包括与收集到的一些日语文档相关的各种图像和问题类型。我们还为每个问题类别提供了无法回答的问题。

在问答的实际应用中，在引用的文档中找不到答案。因此，预计此类问题的正确回答“在文本中未提及”。无法回答的问题的预测在以前的日语问答数据集中没有得到解决，例如 **?** 。

3.2. 数据集统计

JDocQA 数据集包含 5,504 个文件和 11,600 个日语问答对。分类题类统计如下：（1）是/否题：1,855，(2) 事实题：2,052，(3) 数字题：1,866，(4) 开放式题：5,827。此外，有 1,788 个问题需要参考多页才能回答，而在 1,000 个问题中，正文中没有提到正确答案，如表 **??** 所示。某些 PDF 文档在同一文档中同时包含幻灯片和报告格式。对于

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: 按四个问题类别、多页问题和无法回答的问题分列的文档样式和问答对的数量。

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: 平均字符长度。

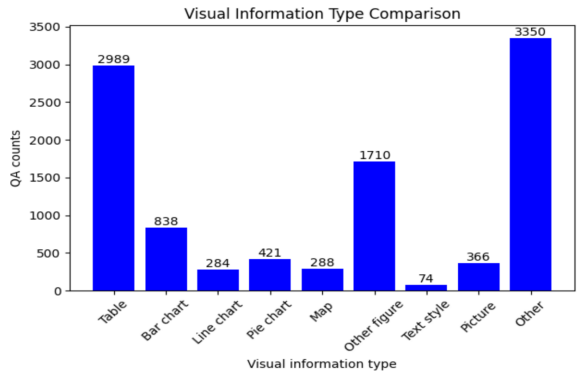

Figure 2: 视觉信息类别的数量。

此类文档，我们在计算 [2] 文档的总数时，将它们计入幻灯片和报告格式的两种类别。Table **??** 表示我们数据集中上下文、问题和答案的平均长度，Figure **??** 表示我们数据集中问题或答案所引用的视觉信息的类别。文档问答数据集的比较见表 **??** 。

3.3. 数据集创建

数据集的整体创建和标注过程如图 **??** 所示。

PDF 集合。 我们收集由日本政府机构或地方政府制作的公共文件，例如市政宣传册和网站。我们从开放获取资源中手动收集了 PDF 文档，例如日本国立国会图书馆（NDL）的数字馆藏，网络档案项目（WARP）[3] 和日本政府各部委的网站。我们通过 WARP 手动收集公共或准公共部门（如地方政府或公立大学）发布的报告、小册子或网站等文件。我们还根据政府机构的政策从他们的网站上收集日本的部门文件，例如幻灯片和报告。这些文件涵盖了广泛的主题，例如经济政策、教育政策、劳工问题、健康和卫生、农业、林业、渔业、文化艺术、历史、与政府政策或政策指导方针有关，以及地方政府的日常事务。这些文件还包括视觉元素，如数字、表格、图表、图片或曼荼罗图，这些复杂的数字与日本公共行政部门官方文件中常见的文

[2] 由于这种计数，总文档数不是表 **??** 中子类别数的总和。

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

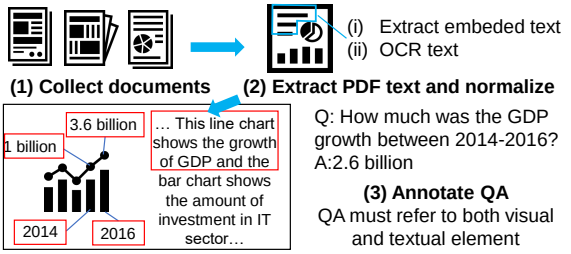

Figure 3: 注释过程。

本和对象的组合。我们将这些文件分为四类，即小册子、幻灯片、报告和网站。

文本提取 & 规范化。 我们使用 PyPDF2 [4] 从 PDF 文档中提取文本。我们还注意到，某些 PDF 文档可能是从纸质扫描创建的，我们无法从此类文档中提取嵌入文本。因此，我们通过 OCR（光学字符识别）从文档页面图像中提取文本作为替代来源。在文本提取或 OCR 之后，我们删除了错误识别的符号和表情符号，或者当同一字符连续重复出现五次以上时，从文本中重复出现的重复字符。

注释过程。 我们总共要求 43 名注释者对文档进行问答对注释。由于文档包含丰富的文本和视觉元素（例如，图形、图表、地图、插图以及垂直和水平书面文本的混合），因此我们制作了与文本和视觉信息相关的问答对。我们要求注释者在每个文档中写两到四个问答注释。我们还要求在注释过程中不要使用任何 AI 工具，例如 OpenAI ChatGPT。每个问题都附有图 **??** 和图 **??** 中红色标出的支持事实。我们将在多个页面中具有多个支持事实的问题子集归类为多页问题。与单页问题相比，多页问题要困难得多。对于无法回答的问题，我们要求注释者在文档中写下缺乏支持事实的问题，使其无法根据给定的文档进行回答。

可视输入和边界框。 我们准备了三种类型的图像作为多模态模型的视觉输入。第一种类型的图像是文档的整个页面的图像，包括带注释的问答对。第二种类型的图像是由注释者基于其答案的边界框裁剪的图像，例如页面的表格或图形。当多个边界框注释到单个问答对时，多个裁剪的图像将在此处组合成一个图像。第三种类型的图像是用于消融研究的空白（白色）图像。

[4] 我们还检查了 PyMuPDF。但是，提取文本的质量并没有太大变化。

Dataset	# Questions	# Images	# BBoxes	Language	Multihop
OCR-VQA (**?**)	1002k	207k	-	English	-
DocVQA (**?**)	50k	12k	-	English	-
InfographicVQA (**?**)	5.9k	30k	-	English	-
MP-DocVQA (**?**)	46k	48k	-	English	✓
SlideVQA (**?**)	14.5k	52k	890k	English	✓
JDocQA (Ours)	11.6k	268k	11k	Japanese	✓

Table 3: 文档问答数据集的比较。

4. 实验

4.1. 问答任务

我们的数据集旨在评估文档上下文的问答能力，包括文本和视觉信息，以及通过开放式文本生成的问题。正如 Sec. **??** 中所讨论的，我们的数据集由四种形式的问题组成：是/否 [5] 、事实、数字和开放式。所有这四个类别的问题都包括无法回答的问题，这些问题不能仅从给定的文档文件中回答。模型有望在任何这些问题类型中以开放式文本生成方式生成答案。文本模型输入，或简称提示，由文档的嵌入文本或 OCR 结果组成，如 Sec. **??** 和问题中所述。我们还在提示中加入了答案格式指南，例如“请以是/否形式回答”、“请回答文档中提到的事实”、“请根据文档中的数字信息回答”和“请以开放式格式写下答案”。对于无法回答的问题，我们为所有问题类别准备一个特殊答案：“UTF8ipxm 本文中に記載がありません （文中未提及）。

4.2. 模型

我们使用文本输入模型以及文本和视觉输入的多模态模型进行实验。对于模型训练，我们使用监督微调。使用训练集和验证集搜索最佳超参数，然后使用最佳超参数评估模型性能。

带有文本输入的模型。我们采用了多达 130 亿（13B）个模型参数量表的日语大型语言模型进行实验。我们试验了以下仅接受文本输入的代表性日语模型：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] 和 Stability AI Japanese-StableLM-Instruct-Alpha-7B [13] . 我们还包括 Llama-2-7B [14] 的多语言大语言模型。我们训练和评估了具有 1024 个令牌长度的模型，以实现公平的比较和计算效率，但 rinna japanese-gpt-4B-8k 最多使用 8192 个令牌进行训练。对于具有较长标记长度的分析，我们训练具有 2048、4096 和 8192 个标记的 rinna japanese-gpt-4B-8k 模型。

具有多模态输入的模型。JDocQA 的目的是通过文本和视觉感知来分析文档。为了评估同时使用图像和文本对 JDocQA 数据集的影响，我们应用了从图像和文本中获取输入的多模态模型。为此，我们使用了 Stability AI Japanese-StableLM-Instruct-Alpha-7B [15] ，这是 InstructBLIP（**???**）的日语版本，因为它们适用于日语文本和图像输入。我们训练和评估了这个模型，其最大容量后有 512 个令牌长度。我们开发了三种不同的模型，具有三种不同的视觉输入，如 Sec. **??** 中所述。第一个模型采用空白图像的视觉输入，该图像始终与消融研究的 800x600 像素大小的白色图像相同。第二个模型采用与注释中的问答相关的整个文档页面的图像。这些图像也被缩放到 800 像素的宽度。第三个模型根据带注释的支持事实对图像进行输入。在带注释的边界框之后，我们裁剪页面图像的引用区域，组合边界框并缩放它们以用于模型视觉输入。由于某些问题的答案具有多个带注释的支持事实，因此组合图像可能包含带注释的边界框的多个区域。所有这些多模态模型还采用类似于文本输入模型的文本提示。

OpenAI GPT 基线。我们还展示了 OpenAI GPT 性能作为基线。在这里，我们使用 gpt-3.5-turbo-16k 和 gpt-4 模型 [16] 。它们采用与文本输入模型类似的提示。然而，由于它们是我们任务的零样本模型，我们观察到它们对提示非常敏感。为了提高它们的性能，我们手动调整 OpenAI GPT 模型的提示。我们避免微调 OpenAI GPT 模型，尽管由于以下原因，微调它们可能会大大提高性能。首先，我们的目的是开发在有限的计算资源上工作的本地模型，其次，这些模型无法获得微调的细节，最后是由于 API 成本问题。

[5] 当模型始终标记为“是”时，包括无法回答的问题（包括无法回答的问题）的几率为 61.57。

[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

[16] 2023 年 10 月 9 日的最新型号。

	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: 所有微调模型和 OpenAI GPT zeroshot 的结果。平均值是分数的加权平均值。

4.3. 评估方法

对于是/否、事实和数字问题，我们在通过简单的规则修剪了标点符号的存在与否或日语后缀短语（如“UTF8ipxm です （是）”等琐碎差异后，使用了完全匹配指标。我们还研究了完全匹配的变化，例如模型预测短语是否包含在正确答案短语中的比率。但是，我们意识到该指标的性能与完全匹配评估非常相似，并且验证集中的差异通常小于 10 个问答对。对于开放式问题，答案通常很长，例如，日语答案的平均长度为 65.97 个字符，因此完全匹配不适用于评估。因此，我们使用 MeCab [17] 标记 BLEU 分数 [18] 来自动评估开放式问题。

4.4. 实验设置

我们的数据集包括所有问题类别中无法回答的问题。虽然预计对具有无法回答的实例的模型进行微调可能会利用模型来超越称为幻觉的幻觉答案，但截至 **?** 年，检测无法回答的问题也是出了名的困难。因此，我们为所有实验的基础模型准备了两种类型的模型：使用所有问答对（包括不可回答的实例）进行微调的模型，以及在没有不可回答实例的情况下微调的模型。在评估中，我们同样准备了两个独立的验证和测试集：由所有问答对组成的标准验证和测试集，以及删除无法回答的问题的较小验证和测试集。

4.5. 结果

Table **??** 显示了所有问题类型在有效拆分和测试拆分上的文本输入模型的性能。

使用所有实例训练的模型。 我们比较表 **??** 中的第一个和第三个块。它们是所有 JDocQA 实例的结果，由零样本和微调模型组成，这些模型使用所有 JDocQA 训练实例（包括“无法回答的问题”）进行训练。我们意识到，微调模型的性能优于 gpt-3.5 和 gpt-4 结果，尤其是当模型大小较大时。接下来，我们比较表 **??** 中的第二块和第四个块。它们与之前在 JDocQA 上的区块评估相同，没有无法回

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

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

	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: 多模态输入模型的结果。平均值为加权平均值。

	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: 不同标记长度的 rinna bi-4B-8k 模型的结果。平均值为加权平均值。

答的问题。我们观察到所有实例评估都有类似的趋势，这表明对不可回答的实例进行微调的模型在可回答和不可回答的问题中都表现出相似的表现。其中，StableLM 模型尽管参数大小为 7B，但表现最佳。尽管参数大小很大，但 rinna bi-4B-8k 型号也表现良好。我们将其归因于其代币长度大小 8192，稍后将对此进行讨论。我们注意到（1）是/否问题相对容易，尽管它们对平均分数（平均）没有太大影响，平均分数（平均）主要遵循最常见的问题类别（4）开放式。

模型训练没有无法回答。 如 Sec. **??** 中所述，我们还准备了在训练实例中微调的模型，没有无法回答的问题。我们在表 **??** 的第五块和第六块中提出了包括和排除无法回答的问题在内的评估结果。当我们比较表 **??** 中的第三块和第五块结果时，这是非常有趣的，因为它们共享相同的评估集，包括无法回答的问题，而模型在有和没有不可回答的实例的情况下进行了微调。比较第三块和第五块结果，我们注意到几乎所有用无法回答的问题进行微调的模型在平均分数上都比仅可回答的微调模型表现得更好。OpenCALM-7B、weblab-10B 和 Llama2-7B 模型除外，我们将在下一段中讨论。我们将其归因于幻觉的概念，其中模型生成的答案不会出现在上下文文本中。我们将在定性分析段落中介绍一个例子。在这种实验比较中，兴奋剂无法回答的情况可能有助于利用幻觉。

OpenCALM-7B 和 weblab-10B 不会预测问题无法回答。 在 Table **??** 的第三个块中，我们注意

	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-AI.-7B	26.96	20.40	24.04	4.64
Inst-AI.-7B	27.69	23.26	23.01	5.71
InstBLIP-AI (blank)	25.49	22.04	21.48	3.79
InstBLIP-AI (img)	25.00	21.63	23.78	3.75
InstBLIP-AI (bbox)	25.00	22.04	22.50	4.20

Table 7: 详细的文件类型结果。” 网站” 仅作为域外集包含在测试集中。

到有趣的现象：OpenCALM-7B 和 weblab-10B 尽管参数大小很大，但性能不佳。当我们仔细检查这些模型的输出时，我们意识到这些模型，对所有实例进行了微调，预测的 “UTF8ipxm 本文中に記載がありません （文中未提及）” 比其他模型少得多。OpenCALM-7B 和 weblab-10B 预测所有实例的 11.9 个% 和 13.7 个% 是无法回答的，而 Table 的第三个块中的其他模型 **??** 预测大约 20 个% 。由于他们使用相同的数据集进行训练，我们怀疑这是由于他们的预训练。同样值得注意的是，将问题预测为无法回答的问题对于模型来说是一项艰巨的任务，并且通常会影响整体性能。

多模态模型结果。 我们在表 **??** 中给出了具有 StableLM-InstructBLIP-Alpha 的多模态输入模型的结果。模型性能得到增强，特别是当我们使用参考表格或图形的裁剪图像时（bbox）。我们还注意到，具有黑色图像输入的模型在某种程度上表现接近视觉模型，这表明文本输入在我们的任务中的有效性。我们还仔细注意到，StableLM-InstructBLIP-Alpha 的最大 token 长度为 512，这可能会限制当前多模态模型的文本理解能力。

令牌长度依赖关系。 我们调查了令牌长度对性能的影响。为此，我们微调了三种不同标记长度的 rinna bi-4B-8k 模型，例如，分别针对所有训练实例和没有不可回答条件的 2048、4096 和 8192。我们在 Table **??** 中呈现表演。我们注意到微调令牌长度肯定会影响最终结果，尽管我们也注意到长令牌长度的微调模型比短令牌长度模型的计算成本高得多，这可能是 rinna bi-4B-8k（8192 个令牌）模型在表 **??** 中表现良好的原因。

文档类型的详细分析。 表 **??** 显示了每种文件类型之间的性能比较。文件类型分为公共关系小册子或杂志等日语小册子、演示材料等幻灯片和图表等报告文件。我们还准备了网站扫描的域外测试集，其中模型的表现仍然较差。

定性分析。 图 **??** 提供了两个模型世代的示例实例。在上面的例子中，我们分别提出了问题、注释答案和三个模型的生成，这些问题包括不可回答的

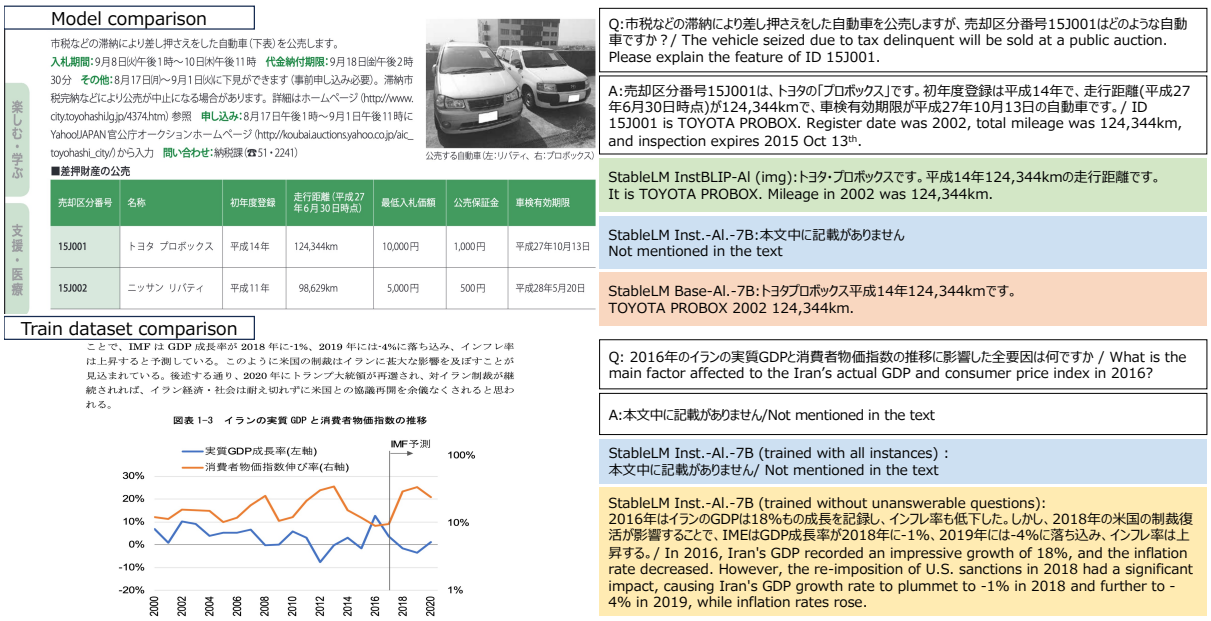

Figure 4: 开放式问答的定性分析。

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: 对采样集进行人工评估。

问题：StableLM InstructBLIP-Alpha （img）、StableLM Instruct-Alpha-7B 和 StableLM Base-Alpha-7B。StableLM InstructBLIP-Alpha （img）可以通过视觉输入查看表格的对齐情况，并为汽车生成合理的描述。StableLM Base-Alpha 也生成了类似的答案，但它无法归因于表中显示的 123,334 的运行里程。在 Figure **??** 的底部，我们展示了两个微调模型的比较，它们来自同一个预训练模型，并且训练时有和没有无法回答的问题。所有实例模型准确地预测文本中没有写有答案，而没有无法回答问题的模型会错误地生成开放式答案，从而导致称为幻觉的现象。

人工评估。 最后，我们对一些代表性模型的抽样结果进行了人工评估。为此，我们从测试集中抽取了 100 个开放式问题。我们选择了 PLaMo-13B 和 StableLM Instruct-Alpha-7B 两种文本输入模型。我们还选择具有背面和图像输入的 StableLM InstructBLIP-Alpha 的多模态输入模型。我们要求注释者为两个标准附加从 0 到 2 的分数：生成的答案是否包含带注释的答案，以及生成的答案是否不包括错误陈述作为问题的答案。结果如表 **??** 所示。我们注意到 StableLM Instruct-Alpha-7B 的性能优于 PLaMo-13B，图像模型的性能优于空白图像模型。

5. 结论

我们引入了 JDocQA 数据集，该数据集集中于日语问答中视觉和语言线索的整合。我们从给定的文件中纳入了无法回答的问题，我们证实这些问题在我们的实验中在某种程度上可以有效地利用幻觉的产生。我们的详细评估揭示了我们的数据集在从是/否到开放式的广泛问题类别中的有效性，并且对无法回答的问题的预测可以成为提高模型性能的线索，说明了 JDocQA 数据集在现实应用中的有效性，其中多类问题是可行的，并且某些程度的问题在文档中没有明确的书面答案。

6. 确认

这项工作得到了 JSPS 青年科学家补助金（# 22K17983），JSPS 促进联合国际研究（A）（# 22KK0184）和 JST PRESTO（# JPMJPR20C2）的支持。

7.

道德声明 & 限制 在数据收集方面，我们从日本国立国会图书馆（NDL）的数字馆藏、网络档案项目（WARP）和日本政府各部委的网站收集文件 PDF 文件和网页。地方政府或大学发布的行政 PDF、文件、小册子或网站是通过 WARP 收集的。我们小心翼翼地避免使用私人文件，并选择由公共或准公共部门发布的大量公共文件来宣传我们的数据集使用情况。所有文件和网页都在网上公开提供，我们遵循我们的机构规则来收集它们。我们遵守我们的机构规则，并就数据收集过程咨询外部顾问。

我们假设我们的数据集对于生成语言模型的研究和开发及其在日语文档问答中的应用都很有用。我们还认为，带有无法回答问题的数据集有助于利

用大型语言模型的幻觉问题。然而，这并不意味着具有无法回答的问题的精细模型根本不会产生幻觉。

8. 参考书目