HakushoBench: A Japanese Chart and Table VQA Benchmark from Governmental White Papers

Issa Sugiura[♡,♠] **Shuhei Kurita**[♢,♠] **Yusuke Oda**[♠] **Naoaki Okazaki**[♡,♠,♣]
♡Institute of Science Tokyo ♠NII LLMC ♢NII ♣AIST

Abstract

Understanding chart and table images is essential for applying vision-language models (VLMs) to real-world document understanding. While English benchmarks have advanced rapidly, non-English counterparts remain scarce, leaving it unclear whether this progress generalizes across languages. A key obstacle is the difficulty of collecting realistic and diverse non-English chart and table images at scale. To address this, we leverage governmental white papers as a source for benchmark construction, as they contain naturally occurring charts and tables across diverse formats and domains and are freely accessible in many countries. As a first instantiation, we introduce HakushoBench, a Japanese chart and table VQA benchmark built from 33 governmental white papers. HakushoBench contains 2,053 images spanning over 10 image types, with manually annotated and independently verified QA pairs designed to assess holistic understanding of charts and tables rather than local visual cues alone. Experiments across a broad range of VLMs show that HakushoBench is substantially harder than the existing Japanese benchmark and remains challenging for open-weight models: sub-10B open-weight models reach at most 58.6% accuracy, and even the flagship open-weight model Qwen3.5-397B-A17B trails Gemini 3 Pro by 8.1 points (85.8% vs. 93.9%), highlighting substantial room for improvement in complex chart and table understanding. We release our dataset and code.

🤗 **Dataset:** llm-jp/HakushoBench

Code: github.com/llm-jp/simple-evals-mm

1 Introduction

Vision-language models (VLMs) (Liu et al., 2023; OpenAI, 2024; Qwen Team, 2025; Kimi Team, 2025) have rapidly advanced as general-purpose models capable of solving a wide range of vision-language tasks (Antol et al., 2015; Masry et al., 2022; Lu et al., 2024; Xie et al., 2024). Among these tasks, chart and table visual question answering (VQA) represents a fundamental capability for VLMs, as charts and tables are ubiquitous across a wide range of real-world documents, including public reports, financial documents, and scientific articles, often conveying information that cannot be expressed in text alone (Masry et al., 2022; Kantharaj et al., 2022).

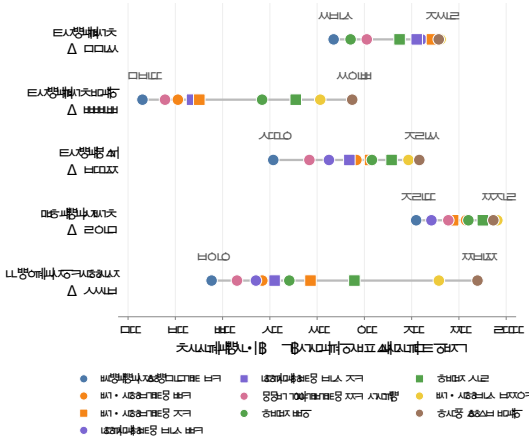


Figure 1: Score distribution of the evaluated models on each benchmark (best of Direct/CoT; Δ = max − min). Every model scores lower on HakushoBench than on the existing Japanese benchmark JGraphQA, where scores cluster above 80%.

To evaluate chart and table understanding in VLMs, numerous benchmarks have been developed. Early datasets such as ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) mainly focused on relatively simple chart types and questions requiring straightforward extraction of numerical values or labels. As VLM performance rapidly improved, these benchmarks became increasingly saturated and less effective at distinguishing model capabilities (Ho et al., 2025). More recent benchmarks, such as ChartQAPro (Masry et al., 2025) and CharXiv (Wang et al., 2024), address this limitation by introducing more di-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2273–2290
November 6-10, 2026 ©2026 Association for Computational Linguistics

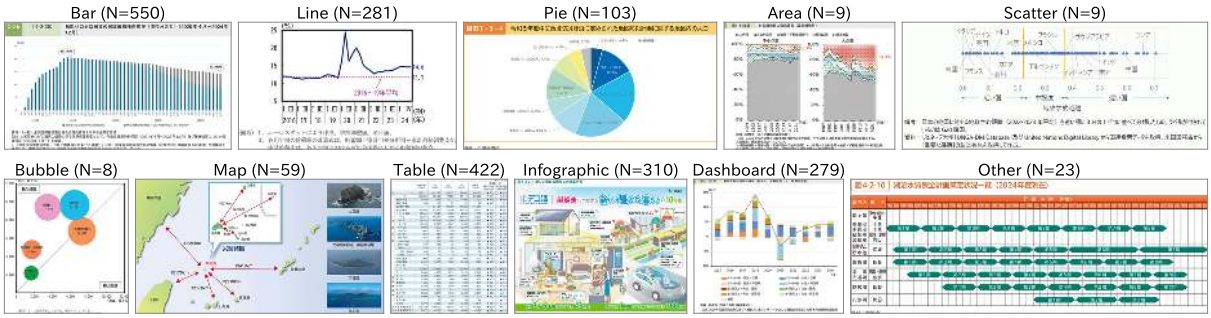


Figure 2: Diversity of image types in HakushoBench. One randomly sampled example is shown for each image type included in HakushoBench. HakushoBench contains diverse, well-designed chart and table images that present rich information in an accessible and visually understandable manner for general readers of government white papers.

verse real-world chart images and more challenging question-answer pairs.

However, existing chart and table benchmarks are heavily biased toward English-centric visual conventions and document styles (Masry et al., 2022; Wang et al., 2024; Masry et al., 2025). Real-world chart understanding varies across languages and cultures, affecting visual composition, textual structure, and reasoning requirements (Tang et al., 2025; Xu et al., 2026). For example, chart and table images may differ in geographic conventions used in map-based figures, language-specific terminology in tables, writing direction (e.g., mixed vertical and horizontal text), and overall information density and layout structure (Wei et al., 2025; Sasagawa et al., 2025; Onami et al., 2024). As a result, strong performance on existing English benchmarks does not imply robust multilingual chart understanding (Pouget et al., 2024).

In Japanese, JGraphQA (Kinoshita, 2025) is currently the primary benchmark for chart and table understanding. However, it consists of only around 200 instances with limited visual diversity and relatively simple questions, leading to performance saturation where even 3B-scale VLMs already achieve over 80% accuracy (Figure 1).

To address these limitations, we leverage governmental white papers (*Hakusho*, 白書, in Japanese) as an underexplored yet valuable source for chart and table benchmark construction. These reports contain large amounts of real-world charts and tables spanning diverse domains and visual formats, and are publicly available across many countries and languages (Japanese Digital Agency, 2026; U.S. Government Publishing Office, 2026).

As a first instantiation of this approach, we introduce **HakushoBench**, a challenging benchmark for Japanese chart and table understanding constructed from 33 white papers published by Japanese governmental agencies. The benchmark contains 2,053 unique images spanning more than 10 image types, with manually annotated QA pairs designed to be challenging, including questions that require integrating information across the entire image and multi-hop reasoning beyond simple data extraction.

Benchmark	Real	Image Types	Lang.
PlotQA (Methani et al., 2020)	✗	3	EN
ChartQA (Masry et al., 2022)	✓	3	EN
ChartQAPro (Masry et al., 2025)	✓	>10	EN
CharXiv (Wang et al., 2024)	✓	>10	EN
JGraphQA (Kinoshita, 2025)	✓	5	JA
HakushoBench (Ours)	✓	>10	JA

Table 1: Comparison of chart and table VQA benchmarks. "Real" indicates that images are collected from real-world sources rather than programmatically generated. HakushoBench provides realistic and visually diverse Japanese chart and table VQA.

We evaluate a broad suite of open-weight and proprietary VLMs on HakushoBench and compare against existing Japanese and English chart and table benchmarks. HakushoBench is substantially harder than JGraphQA: sub-10B open-weight models reach at most 58.6% accuracy (Qwen3-VL 8B (Qwen Team, 2025)), and even the flagship open-weight model Qwen3.5-397B-A17B (Qwen Team, 2026), which matches or exceeds Gemini 3 Pro on JGraphQA and ChartQA, trails it by 8.1 points on HakushoBench (85.8% vs. 93.9%). The resulting 56-point spread across models distinguishes them far more clearly than JGraphQA, where scores cluster between 81% and 98%.

2 Related Work

Chart and table VQA benchmarks. As vision-language models have become increasingly capable

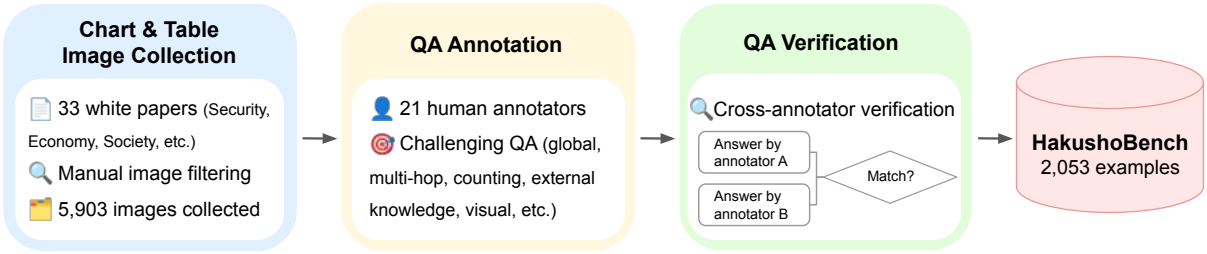


Figure 3: Construction pipeline of HakushoBench. Chart and table images are collected from 33 Japanese white papers and filtered to 5,903 candidates. Annotators then create one high-difficulty QA pair per image, followed by independent verification, yielding 2,053 VQA pairs.

and general-purpose (Liu et al., 2023; Qwen Team, 2025), chart and table VQA benchmarks have also evolved toward greater diversity and more challenging reasoning tasks.

Early datasets such as FigureQA (Kahou et al., 2018) and DVQA (Kafle et al., 2018) focused on synthetic bar charts, while later work such as PlotQA (Methani et al., 2020) expanded chart types and task complexity. To bridge the gap with real-world documents, ChartQA (Masry et al., 2022) introduced human-annotated QA pairs on real chart images collected from online platforms such as Statista (Statista, 2026) and Our World in Data (OWID, 2026), which present diverse topics across a broad range of domains.

More recent benchmarks have further expanded the scope of image sources: ChartQAPro (Masry et al., 2025) collected charts from a larger set of online platforms, covering more varied types such as dashboards and infographics, while CharXiv (Wang et al., 2024) extracted chart images from papers across eight arXiv categories (including Computer Science, Economics, and Physics) to construct a natural and challenging benchmark. ChartMuseum (Tang et al., 2026) further targets questions that require visual rather than textual reasoning, and MTabVQA (Singh et al., 2025) evaluates multi-hop reasoning over multiple rendered tables. Despite these advances, chart and table benchmark development has primarily focused on English.

Japanese chart and table VQA benchmarks. In Japanese, several benchmarks have been proposed for document understanding. JDocQA (Onami et al., 2024) built a document VQA benchmark from diverse PDFs published by Japanese public institutions such as municipal offices, and BusinessSlideVQA (Stockmark, 2025) constructed a slide understanding benchmark from business slides released by Japanese companies.

The most closely related prior work, JGraphQA (Kinoshita, 2025), focuses on Japanese chart and table question answering and is constructed from Investor Relations (IR) presentation slides of Japanese companies. However, this source domain is relatively narrow, and the benchmark contains only around 200 examples. Furthermore, following the QA design of ChartQA, questions are comparatively simple and focused primarily on data extraction and basic arithmetic rather than complex reasoning. As a result, even 3B-scale VLMs already achieve over 80% accuracy, suggesting that the benchmark is becoming saturated.

Table 1 summarizes existing chart and table VQA benchmarks alongside HakushoBench. In contrast to prior work, HakushoBench leverages governmental white papers as an image source, enabling a realistic, visually diverse, and challenging Japanese chart and table VQA benchmark.

3 Construction of HakushoBench

HakushoBench is constructed through a three-stage pipeline, illustrated in Figure 3.

3.1 Chart and Table Image Collection from Japanese White Papers

White papers as a data source. White papers published by Japanese governmental agencies summarize official statistics and policy analyses on national topics such as defense, energy, welfare, and education for a general readership. As a result, the images they contain are professionally designed, information-dense, and visually diverse across a wide range of domains, making them a valuable source for evaluating the generalizability of VLMs across heterogeneous chart and table types. Furthermore, since many governments around the world

similarly publish official white papers (U.S. Government Publishing Office, 2026), the same source-level methodology could be applied to other languages. We demonstrate it for Japanese in this paper and leave multilingual instantiation to future work.

Scope and edition selection. We collect chart and table images from such white papers, which are publicly available through the Japanese government's e-Gov portal (Japanese Digital Agency, 2026). Each Japanese governmental agency publishes its own annual white paper spanning a broad range of policy domains. While these are primarily distributed as PDFs, accurately extracting figures and tables from PDFs remains challenging even for current OCR models (Wei et al., 2025; Cui et al., 2025). We therefore restrict our dataset to white papers that also provide HTML editions, from which chart and table images can be collected directly via their URLs. Since different yearly editions of the same white paper often contain highly similar charts and tables (e.g., annual updates of population statistics), we use only the most recent edition from each white paper series. This choice also helps reduce potential contamination risk (Oren et al., 2024). As a result, our benchmark is constructed from 33 distinct governmental white paper series. The full list is provided in Appendix C.

Image filtering and candidate selection. From these white papers, we initially collect 18,539 images. However, many of them are unsuitable for chart and table QA annotation, including photographs such as group pictures, low-resolution images whose contents are difficult to read, and near-duplicate images appearing across multiple pages. We therefore manually remove such inappropriate images and retain 5,903 candidate chart and table images for annotation.

3.2 QA Annotation

Building on the collected chart and table images, we annotate them with question-answer pairs as follows. QA annotation is carried out by 21 native Japanese-speaking annotators hired through a professional annotation agency. Each annotator is given a set of images and writes one QA pair per image, following the requirements below.

QA requirements. Questions are required to (a) be unanswerable from the question text alone, requiring the image to resolve; (b) be natural and well-posed; and (c) admit a single, unambiguous answer expressible in a word, phrase, or short sentence. These constraints follow prior work showing that ambiguous questions make evaluation difficult and hinder fair assessment of model capabilities (Chen et al., 2024; Joshi et al., 2026).

We adopt a short answer format (Wang et al., 2024), avoiding multiple-choice and yes/no questions, which can be solved without the image or by random guessing (Chen et al., 2024). Answers are scored using an LLM-based judge that classifies each response as correct or incorrect, enabling robust evaluation that accommodates surface-level variations (e.g., "7" vs. "seven").

Question-type flags. To encourage diverse and reasoning-intensive QA pairs, each accepted question must satisfy at least one of the following question types:

- **Global**: requiring integration of information spread across multiple regions of the image;
- **Multi-hop**: requiring multiple reasoning steps, such as extracting multiple values and then computing their ratio;
- **Counting**: requiring counting or indexing of objects;
- **External knowledge**: requiring knowledge beyond the image to answer correctly, such as knowledge of Japanese geography;
- **Visual**: requiring fine-grained visual perception, such as object color and shape;
- **Other**: covering difficult questions that did not fit the above categories.

Annotators assign one or more flags per QA pair. The flags are non-exclusive by design: a single question frequently carries several (e.g., extracting values from distant regions of a table and then computing their ratio is both Global and Multi-hop).

Annotators are allowed to skip images for which no sufficiently challenging and well-defined question could be written, preventing the inclusion of artificially simple QA pairs.

3.3 QA Verification

QA annotation is prone to producing ambiguous questions and incorrectly annotated answers, and verification by independent annotators is widely used to improve benchmark quality (Chen et al., 2024; Masry et al., 2025). Following this practice, we introduce a separate verification stage for all annotated QA pairs.

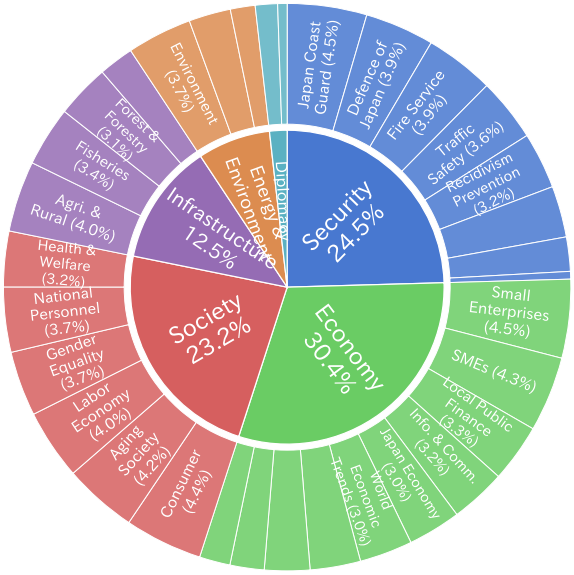


Figure 4: Distribution of QA pairs in HakushoBench. The inner ring represents the topic-wise distribution of examples, while the outer ring represents the white-paper-wise distribution.

For each QA pair, a second annotator (different from the original author) is given only the image and question and asked to independently answer it without access to the original answer. If the two answers match up to minor surface variations, the example is accepted; otherwise, the QA pair is revised or discarded. After completing the full pipeline, we obtained 2,053 QA pairs.

4 Exploring HakushoBench

Statistics of HakushoBench. HakushoBench contains 2,053 examples collected from 33 white papers, grouped into six topics (Security, Economy, Society, Infrastructure, Energy & Environment, and Diplomacy). Figure 4 shows the distribution of QA pairs per topic and per white paper, and Appendix C lists the per-white-paper breakdown. The largest topic is Economy (30.4%), followed by Society (23.3%), while the number of QA pairs per white paper is relatively balanced. Table 3 reports the distribution of question-type flags. Each type has at least 100 examples, and Global is the most frequent, as multiple flags are allowed per question and Global tends to overlap with other types: 1,430 questions (69.7%) carry two or more flags, most often Global together with Visual (801) or Multi-hop (604). Appendix G gives the full co-occurrence matrix and one example per type. Questions are short (mean 39.9 characters, median 37, range 10–109), and answers are shorter still (median 5 characters).

Visual diversity. To analyze the visual diversity of HakushoBench, we categorize images from HakushoBench, JGraphQA, ChartQA (test), ChartQAPro, and CharXiv (val) into 11 visual-format categories: Bar, Line, Pie, Area, Scatter, Bubble, Map, Table, Infographic, Dashboard, and Other. Our taxonomy follows that of ChartQAPro (Masry et al., 2025), with the addition of Map and Table. We use Gemini 3 Pro (Google DeepMind, 2025) to classify image types by providing each image together with a classification prompt, which is described in Appendix F. To validate the classifier, an author blindly re-annotated 100 randomly sampled images without seeing the model's labels; the two agree on 83, with near-perfect agreement on single-chart types and disagreements concentrated on composite images for which a single label is inherently ambiguous (Appendix D). Table 2 shows the image types distribution for each benchmark. Compared to JGraphQA, HakushoBench contains more than 10× as many images and covers more than 10 image types, including types absent from JGraphQA such as Map and Infographic, enabling evaluation of more diverse visual understanding capabilities.

We further measure visual diversity using image embeddings extracted by the SigLIP2 vision encoder (`siglip2-so400m-patch16-512`) (Tschannen et al., 2025). Computing the mean pairwise cosine distance between embeddings shows that HakushoBench achieves higher diversity than JGraphQA (0.365 vs. 0.275).

Dataset showcase. Figure 2 presents a randomly selected image from each image type in HakushoBench, illustrating its diversity. As shown, HakushoBench contains a wide variety of chart and table image types spanning diverse domains, reflecting the breadth of governmental white papers as a benchmark resource. Figure 5 shows representative VQA pairs, demonstrating that HakushoBench contains challenging questions over information-dense chart and table images. For example, some questions require extracting and computing multiple numerical values from a table, others demand understanding spatial relationships between map locations and legends, and still others involve reading and comparing textual descriptions embedded in figures. These examples illustrate that HakushoBench often requires multi-hop reasoning and holistic image understanding beyond simple value extraction.

	Bar	Line	Pie	Area	Scatter	Bubble	Map	Table	Infographic	Dashboard	Other	Total
HakushoBench	550	281	103	9	9	8	59	422	310	279	23	**2,053**
JGraphQA	50	30	44	0	0	0	0	50	0	22	0	196
ChartQA	1,220	211	78	0	0	0	0	0	0	0	0	1,509
ChartQAPro	380	317	22	39	18	12	9	1	33	416	5	1,252
CharXiv	108	443	5	18	126	6	9	5	1	105	153	979

Table 2: Distribution of image types across benchmarks. Compared with JGraphQA, HakushoBench contains over ten times more examples and covers a broader range of image types, while remaining competitive with benchmarks such as ChartQAPro and CharXiv in both scale and diversity.

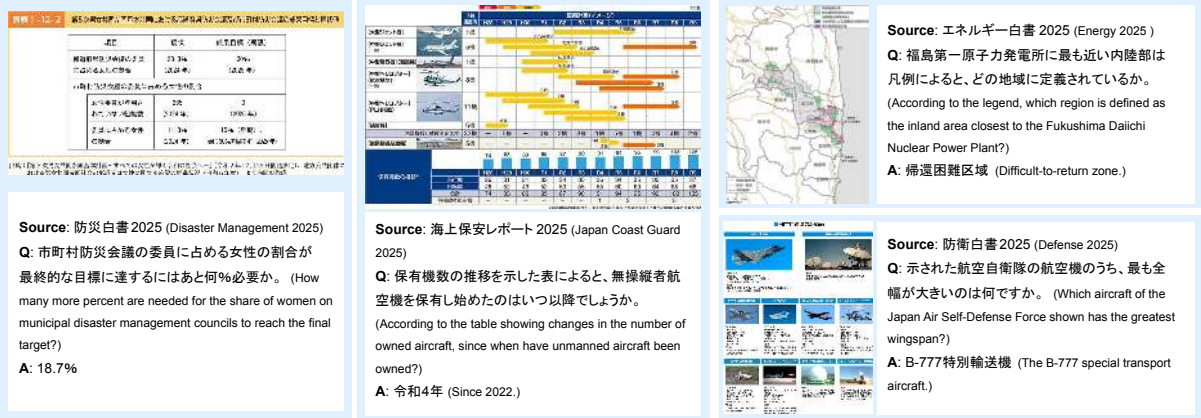


Figure 5: Representative VQA pairs in HakushoBench, requiring multi-hop reasoning and global image understanding rather than local visual cues alone.

Question Type	Examples
Global	1,743
Multi-hop	634
Counting	293
External knowledge	148
Visual	897
Other	132

Table 3: Distribution of question-type flags over the 2,053 verified QA pairs in HakushoBench. Each QA pair may have multiple flags, so counts are not mutually exclusive and do not sum to 2,053.

5 Experiments

We evaluate a broad set of open-weight and proprietary models on HakushoBench, and compare its characteristics against existing benchmarks.

5.1 Experimental Settings

Models. We evaluate a diverse set of open-weight and proprietary VLMs. On the open-weight side, we include Qwen3-VL-4B, 8B (Qwen Team, 2025) and InternVL3.5-4B, 8B (InternVL Team, 2025) as general-purpose multilingual models, as well as Sarashina2.2-Vision-3B (SB Intuitions, 2025) and LLM-jp-4-VL 9B beta (Sugiura et al., 2026b) as Japanese-centric models. To gauge the frontier of open-weight models, we additionally evaluate Qwen3.5-397B-A17B (Qwen Team, 2026), a 397B-parameter mixture-of-experts VLM with 17B active parameters. On the proprietary side, we evaluate GPT-4o (`gpt-4o-2024-11-20`) (OpenAI, 2024), GPT-5.1 (`gpt-5.1-2025-11-13`) (OpenAI, 2025), and Gemini 3 Pro (`gemini-3-pro-preview`) (Google DeepMind, 2025).

Prompt settings. To examine the effect of reasoning-oriented prompting, we evaluate two settings per model. The **Direct** setting asks models to answer the question accurately and concisely, while the **CoT** setting additionally instructs models to think step by step before producing the final answer (Wei et al., 2022; Kojima et al., 2022). Full prompts are provided in Appendix F.

Inference settings. We set the temperature to 0 for all models except GPT-5.1, which does not support temperature control, and use a maximum generation length of 8,192 tokens. Open-weight models are evaluated on NVIDIA A100 GPUs, except Qwen3.5-397B-A17B, which is served with SGLang on 8 NVIDIA H200 GPUs with its native thinking mode enabled and the reasoning trace

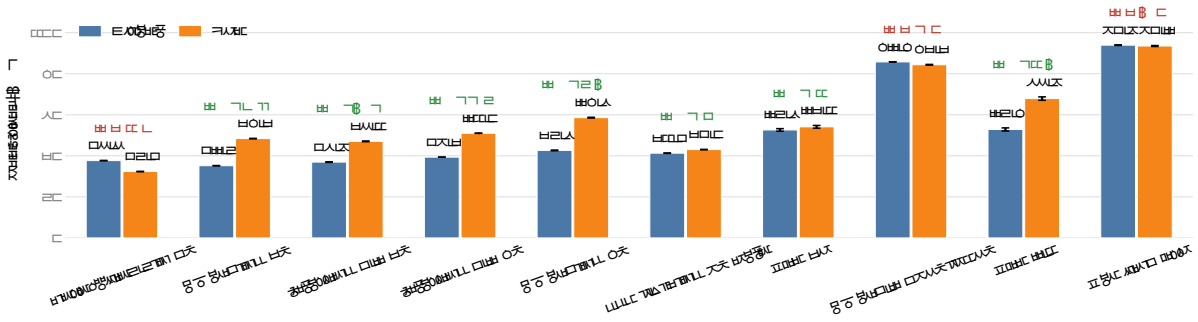


Figure 6: Performance of each model on HakushoBench under the Direct and CoT settings. Gemini 3 Pro and Qwen3.5-397B-A17B reason internally in both settings, so their Direct bars already include native thinking. All sub-10B open-weight models remain below 60% accuracy, highlighting the challenging nature of HakushoBench.

stripped before grading. Proprietary models are accessed through their official APIs. For reasoning settings, we use the `medium` reasoning mode for Gemini 3 Pro. For GPT-5.1, we use `none` for the Direct setting and `medium` for the CoT setting.

Evaluation metric. We report accuracy, scoring each model output as correct or incorrect with GPT-5.1 (`gpt-5.1-2025-11-13`) as an LLM judge, which tolerates minor surface-level variations in phrasing. To account for stochastic variation in both model outputs and LLM judgment, each evaluation is repeated three times, and we report the mean and standard deviation. The judge prompt is given in Appendix F, and Appendix E validates the judge against human judgment (95% agreement, Cohen's $\kappa = 0.90$).

Compared benchmarks. To better understand the characteristics of HakushoBench, we additionally evaluate the same set of models on four existing chart and table benchmarks: JGraphQA (Kinoshita, 2025), ChartQA (test) (Masry et al., 2022), ChartQAPro (Masry et al., 2025), and CharXiv (val) (Wang et al., 2024). For JGraphQA, we use JGraphQA-Verified (Sugiura et al., 2026a), a cleaned and corrected version of the original benchmark. Per-model results on these benchmarks are visualized in Appendix I, and representative images from each benchmark are shown in Appendix J.

6 Results

6.1 Main Results

Figure 6 shows model performance on HakushoBench.

Gemini 3 Pro leads; the flagship open-weight model narrows but does not close the gap. Gemini 3 Pro achieves the highest score of 93.9%. Among open-weight models, Qwen3.5-397B-A17B reaches 85.8%, clearly surpassing GPT-5.1 (67.9%) but still 8.1 points below Gemini 3 Pro. All sub-10B open-weight models remain below 60%, with Qwen3-VL 8B the best at 58.6%. GPT-5.1 scores only 67.9%, revealing a substantial gap even among proprietary models in Japanese chart and table understanding. These results suggest considerable room for improvement in open-weight models, and in particular in the small models that are most widely deployed.

CoT prompting effectiveness varies by model. CoT prompting improves accuracy for most models, with GPT 5.1, Qwen3-VL, and InternVL3.5 all gaining more than 10 points. In contrast, GPT-4o, LLM-jp-4-VL 9B beta, and Sarashina2.2-Vision 3B show limited gains or even degradation. Manual analysis suggests these models often fail to engage in reasoning or produce repetitive and incoherent chains, indicating limited reasoning ability. The two models with native thinking, Gemini 3 Pro and Qwen3.5-397B-A17B, already reason internally in the Direct setting and gain nothing from the additional CoT prompt (-0.4 and -1.4 points).

6.2 Comparison with Existing Benchmarks

Table 4 reports the performance of each model on all benchmarks, and Figure 1 summarizes the score range across models.

HakushoBench is more challenging than JGraphQA. Compared to JGraphQA, HakushoBench is much harder for every model: Qwen3-VL 8B reaches 58.6% on HakushoBench versus 88.8% on JGraphQA, Sarashina2.2-Vision 3B 37.7% versus 81.0%, and even Qwen3.5-397B-A17B 85.8% versus

Model	HakushoBench		JGraphQA		ChartQA		ChartQAPro		CharXiv	
	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT	Direct	CoT
Open-weight Models										
Sarashina2.2-V 3B	37.7 ± 0.1	32.3 ± 0.1	80.4 ± 0.3	81.0 ± 0.8	47.8 ± 0.2	63.5 ± 0.1	23.0 ± 0.0	23.0 ± 0.1	48.6 ± 0.0	50.7 ± 0.0
Qwen3-VL 4B	35.2 ± 0.1	48.4 ± 0.1	84.2 ± 0.5	91.5 ± 0.3	67.0 ± 0.1	83.7 ± 0.0	30.5 ± 0.0	30.5 ± 0.1	65.3 ± 0.0	68.3 ± 0.1
InternVL 3.5 4B	36.9 ± 0.1	47.1 ± 0.2	84.2 ± 0.0	84.2 ± 0.0	73.7 ± 0.1	82.1 ± 0.1	34.7 ± 0.1	34.8 ± 0.1	62.5 ± 0.0	61.9 ± 0.0
InternVL 3.5 8B	39.4 ± 0.0	51.0 ± 0.0	85.9 ± 0.3	88.4 ± 0.3	75.0 ± 0.2	81.1 ± 0.3	33.5 ± 0.0	33.5 ± 0.1	66.8 ± 0.0	65.7 ± 0.1
Qwen3-VL 8B	42.6 ± 0.2	58.6 ± 0.2	86.6 ± 0.3	88.8 ± 0.5	70.3 ± 0.0	84.3 ± 0.3	35.0 ± 0.1	35.1 ± 0.1	70.1 ± 0.1	71.2 ± 0.1
LLM-jp-4-VL 9B beta	41.3 ± 0.1	43.0 ± 0.1	87.8 ± 0.5	85.2 ± 0.0	70.5 ± 0.1	68.0 ± 0.1	27.8 ± 0.2	27.8 ± 0.1	58.4 ± 0.1	56.6 ± 0.1
Qwen3.5 397B-A17B	85.8 ± 0.1	84.4 ± 0.2	98.0 ± 0.5	**98.1 ± 0.3**	**86.1 ± 0.3**	79.6 ± 0.1	60.5 ± 0.2	60.7 ± 0.3	79.3 ± 0.1	79.0 ± 0.0
Proprietary Models										
GPT 4o	52.6 ± 0.6	54.1 ± 0.7	92.0 ± 0.6	84.7 ± 1.0	62.1 ± 0.3	67.1 ± 0.3	44.1 ± 0.2	48.4 ± 0.1	71.6 ± 0.1	71.4 ± 0.2
GPT 5.1	52.8 ± 0.8	67.9 ± 0.8	95.1 ± 1.1	94.9 ± 1.0	64.9 ± 0.7	77.4 ± 0.3	42.0 ± 0.4	55.5 ± 0.7	68.9 ± 0.2	75.8 ± 0.1
Gemini 3 Pro	**93.9 ± 0.1**	93.5 ± 0.2	97.3 ± 0.3	96.9 ± 0.5	80.5 ± 0.4	85.7 ± 0.4	**67.4 ± 0.2**	65.8 ± 0.2	81.4 ± 0.2	**81.6 ± 0.1**

Table 4: Performance of each model on chart and table benchmarks under the Direct and CoT settings, reported as accuracy (%) averaged over three runs (mean ± standard deviation). For each benchmark, the best score across all (model, setting) pairs is shown in bold. Gemini 3 Pro and Qwen3.5-397B-A17B reason internally in both settings, so their Direct columns already include native thinking.

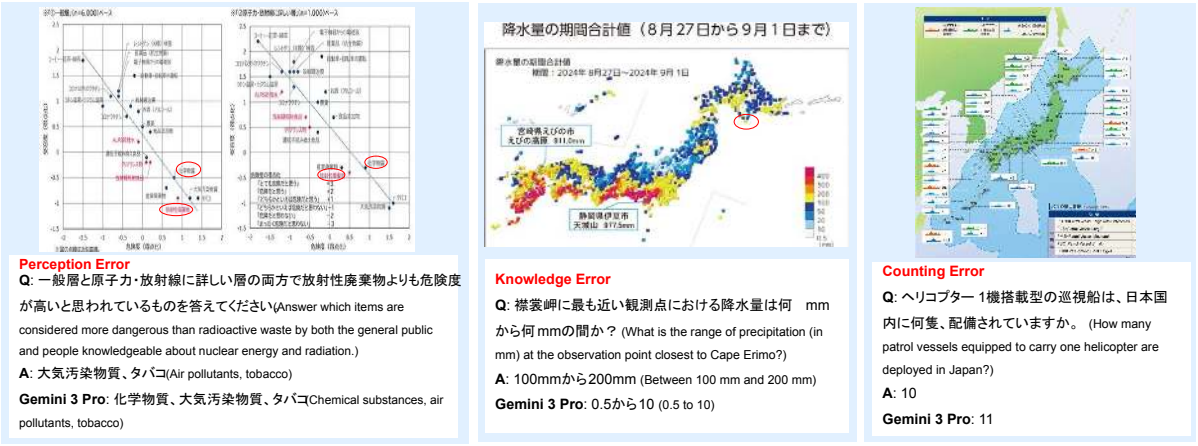


Figure 7: Representative failure cases of Gemini 3 Pro on HakushoBench. Left: a perception error in reading scatter plot conditions. Middle: a knowledge error in identifying the map location closest to Cape Erimo (襟裳岬). Right: a counting error resulting in an off-by-one prediction.

98.1%. Appendix H shows the accuracy spread across image types and across question types. The spread is consistent across both dimensions, with infographics being a notable exception, suggesting that HakushoBench evaluates a broad range of model capabilities.

HakushoBench exposes performance gaps that saturated benchmarks hide. On JGraphQA and ChartQA, Qwen3.5-397B-A17B matches or exceeds Gemini 3 Pro (98.1 vs. 97.3 and 86.1 vs. 85.7), so these benchmarks no longer separate the two, whereas on HakushoBench an 8.1-point gap remains. The same holds across the model range: the spread between the weakest and strongest model is 56.3 points on HakushoBench (37.7–93.9), compared with 17.2 on JGraphQA (81.0–98.1) and 22.6 on ChartQA (63.5–86.1) (Figure 1). In this respect HakushoBench behaves like the recent English benchmarks designed to remain challenging, CharXiv (30.9) and ChartQAPro (44.4), which likewise keep models clearly distinguishable.

6.3 Does HakushoBench Require Visual Understanding?

To verify that HakushoBench requires understanding the visual content of the image, not only the text it contains, we evaluate Qwen3-VL 8B in the Direct setting under three conditions that keep the question fixed and vary only the visual input (Table 5): the original image, the image replaced by its OCR text, and no visual input at all. OCR text is produced by PaddleOCR-VL-1.6 (Zhang et al., 2026), a state-of-the-art document parsing model.

Replacing the image with its OCR text drops

Visual input	Accuracy (%)
Original image (default)	42.6 ± 0.2
OCR text of the image	28.2 ± 0.1
None (question only)	6.1 ± 0.0

Table 5: Input ablation of Qwen3-VL 8B on HakushoBench (Direct setting). The question is always given; only the visual input varies. OCR text is produced by PaddleOCR-VL-1.6.

accuracy from 42.6% to 28.2%, so text extraction alone is not sufficient and HakushoBench requires visual and layout understanding beyond the textual content of the image. Removing the visual input entirely drops accuracy to 6.1%, so HakushoBench cannot be solved from textual priors or memorized statistics.

6.4 Error Analysis of Gemini 3 Pro

To better understand the remaining challenges in HakushoBench, we manually analyze 50 randomly sampled questions incorrectly answered by Gemini 3 Pro (CoT setting), the best-performing model on the benchmark. Of the 50 cases, 22 are perception errors, 3 are reasoning errors, and one each are OCR, external knowledge, and counting errors; the remaining 22 are attributable to gold-answer ambiguity (13) or judge false negatives (9), so the true error rate of Gemini 3 Pro is lower than the reported 6–7%. Figure 7 shows representative examples of three error types. In the *Perception* example, the model fails to correctly interpret spatial relationships in a scatter plot. The *Knowledge* example requires external geographic knowledge to identify Cape Erimo on a thematic precipitation map. The *Counting* example illustrates an off-by-one error in fine-grained counting. Even the best-performing model thus still makes diverse errors, dominated by fine-grained perception.

7 Conclusion

We presented HakushoBench, a Japanese chart and table VQA benchmark constructed from 33 governmental white papers, containing 2,053 VQA pairs over 10 distinct image types. We demonstrated that governmental white papers serve as a valuable source for benchmark construction, offering broad domain coverage and visual diversity. HakushoBench is substantially harder than JGraphQA and remains challenging for open-weight models: sub-10B models reach at most 58.6% accuracy, and even the flagship Qwen3.5-397B-A17B trails Gemini 3 Pro by 8.1 points despite matching it on JGraphQA and ChartQA.

Limitations

Language and domain coverage. HakushoBench addresses the lack of challenging and visually diverse Japanese chart and table QA benchmarks. However, the dataset is currently limited to Japanese and does not directly address the scarcity of benchmarks for other low-resource languages. In addition, our dataset is constructed exclusively from governmental white papers, which may not fully cover the diversity of visual styles and domains found in other real-world documents. Because many countries publish analogous governmental reports and white papers (U.S. Government Publishing Office, 2026), the same source-level methodology could be applied to other languages, but we have demonstrated it only for Japanese and leave multilingual instantiation to future work.

Potential data contamination. We use the most recent edition of each white paper to mitigate contamination risk (Oren et al., 2024). However, we cannot completely rule out the possibility that some images or related information appeared in derivative web content included in model pretraining corpora. That said, all QA pairs in our benchmark were newly constructed through manual annotation, making it highly unlikely that they were included in any model's pretraining data.

Saturation at the frontier. While HakushoBench proves challenging for open-weight models, Gemini 3 Pro achieves 93.9%, leaving limited headroom to discriminate among frontier models. A natural remedy is to construct a harder subset by filtering out questions solved by all evaluated models, following Phan et al. (2026), or to collect more demanding QA pairs targeting capabilities beyond current frontier models. However, increasing difficulty introduces a risk of producing unnatural questions that deviate from realistic use cases. Balancing difficulty and practical relevance is therefore a non-trivial design challenge that we leave for future work.

Ethical Considerations

Public data sources and safety. All images are collected from white papers publicly released by

Japanese governmental agencies. Given this source, risks related to privacy, personally identifiable information, and NSFW (Not Safe For Work) content are negligible. This was further confirmed during manual filtering, where no problematic content was observed.

Acknowledgments

In this research work, we used the "mdx: a platform for building data-empowered society". We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use". This research was supported by the Science Tokyo Support Program for Doctoral Students, funded by the Universities for International Research Excellence.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *NeurIPS*.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Google DeepMind. 2025. Gemini 3 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Accessed: 2026-05-19.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. 2025. A rosetta stone for AI benchmarks. *Preprint*, arXiv:2512.00193.

InternVL Team. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Japanese Digital Agency. 2026. e-Gov. https://www.e-gov.go.jp/about-government/white-papers.html. Accessed: 2026-05-19.

Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Darren Teh, David Schwab, Fan Pan, Haakon Mongstad, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, and 12 others. 2026. DatBench: Discriminative, faithful, and efficient VLM evaluations. *Preprint*, arXiv:2601.02316.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *CVPR*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-Text: A large-scale benchmark for chart summarization. In *ACL*.

Kimi Team. 2025. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Akira Kinoshita. 2025. JGraphQA. https://huggingface.co/datasets/r-g2-2024/JGraphQA. Accessed: 2026-05-19.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *ACL (Findings)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL (Findings)*.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. In *WACV*.

Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. JDocQA: Japanese document question answering dataset for generative language models. In *LREC-COLING*.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. GPT-5.1: A smarter, more conversational ChatGPT. https://openai.com/index/gpt-5-1. Accessed: 2026-05-19.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *ICLR*.

OWID. 2026. Our world in data. https://ourworldindata.org/. Accessed: 2026-05-19.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 281 others. 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No Filter: Cultural and socioeconomic diversity in contrastive vision-language models. In *NeurIPS*.

Qwen Team. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. Accessed: 2026-09-16.

Keito Sasagawa, Shuhei Kurita, and Daisuke Kawahara. 2025. Evaluating multimodal large language models on vertically written japanese text. *Preprint*, arXiv:2511.15059.

SB Intuitions. 2025. Sarashina2.2-Vision-3B. https://huggingface.co/sbintuitions/sarashina2.2-vision-3b. Accessed: 2026-05-19.

Anshul Singh, Chris Biemann, and Jan Strich. 2025. MTabVQA: Evaluating multi-tabular reasoning of language models in visual space. In *EMNLP (Findings)*.

Statista. 2026. Statista. https://www.statista.com/. Accessed: 2026-05-19.

Stockmark. 2025. BusinessSlideVQA. https://github.com/stockmarkteam/business-slide-questions. Accessed: 2026-05-19.

Issa Sugiura, Koki Maeda, Shuhei Kurita, Yusuke Oda, Daisuke Kawahara, and Naoaki Okazaki. 2026a. JAMMEval: A refined collection of Japanese benchmarks for reliable VLM evaluation. *Preprint*, arXiv:2604.00909.

Issa Sugiura, Keito Sasagawa, Keisuke Nakao, Koki Maeda, Ziqi Yin, Zhishen Yang, Shuhei Kurita, Yusuke Oda, Ryoko Tokuhisa, Daisuke Kawahara, and Naoaki Okazaki. 2026b. Jagle: Building a large-scale Japanese multimodal post-training dataset for vision-language models. *Preprint*, arXiv:2604.02048.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. MTVQA: Benchmarking multilingual text-centric visual question answering. In *ACL (Findings)*.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Rea Sprague, Ramya Namuduri, Bodun Hu, Juan Diego Rodriguez, Puyuan Peng, and Greg Durrett. 2026. ChartMuseum: Testing visual reasoning capabilities of large vision-language models. In *NeurIPS (Datasets and Benchmarks Track)*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

U.S. Government Publishing Office. 2026. Economic report of the president. https://www.govinfo.gov/app/collection/erp. Accessed: 2026-05-19.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *NeurIPS (Datasets and Benchmarks Track)*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. DeepSeek-OCR: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS (Datasets and Benchmarks Track)*.

Yichen Xu, Liangyu Chen, Liang Zhang, Jianzhe Ma, Wenxuan Wang, and Qin Jin. 2026. POLYCHARTQA: Benchmarking large vision-language models with multilingual chart question answering. *Preprint*, arXiv:2507.11939.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. 2026. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *Preprint*, arXiv:2606.03264.

A Licenses for Our Resources

HakushoBench and its evaluation code are released under the Apache 2.0 License. Note that we distribute only image URLs rather than the raw image data.

B Use of AI Assistants

We used AI assistants to correct typographical errors, improve the clarity and naturalness of expressions, and generate scripts for plotting figures.

C White Paper Sources

Table 6 lists the 33 Japanese governmental white papers used in HakushoBench.

D Validation of the Image-Type Classifier

To validate the Gemini 3 Pro image-type labels used in Section 4, an author re-annotated 100 randomly sampled HakushoBench images without seeing the model's labels. Table 7 shows the confusion matrix. The two agree on 83 of 100 images. Agreement on single-chart types is near-perfect (Table 19/19, Infographic 9/9, Map 3/3, Pie 2/2, Bar 29/30, Line 15/16), and 15 of the 17 disagreements involve the Dashboard class: composite images with several panels or mixed chart types, where the human tends to label the dominant panel while the model assigns Dashboard, and no single ground-truth label exists. The labels are therefore not a strict taxonomy for composite figures, but they are sufficiently accurate for their purpose in this paper, namely comparing the image-type diversity of benchmarks.

E Validation of the LLM Judge

We validate the GPT-5.1 judge against human judgment on 100 model answers sampled uniformly at random from the outputs of all evaluated models and settings. An author graded each answer given the image, question, and gold answer. The judge and the human agree on 95 of 100 answers (Cohen's $\kappa = 0.90$). All five disagreements are in the same direction: the human judged the answer correct while the judge marked it incorrect, typically because of surface-form variation or because the answer is valid given the image content, which the text-only judge cannot see. There is no case in which the judge accepted an answer the human rejected, so the reported accuracies are, if anything, slightly conservative. Providing the image to the judge is a possible refinement that we leave for future work.

F Prompt

We show the prompts used in the experiments below.

Image type classification prompt

You are an expert at classifying figures in documents. Look at the image and choose exactly ONE label from the following 11 categories that best describes the primary visualization type:
- Bar : Bar chart (vertical, horizontal, grouped, or stacked bars).
- Line : Line chart (including multi-line time series).
- Pie : Pie chart or donut chart.
- Area : Area chart or stacked-area chart.
- Scatter : Scatter plot.
- Bubble : Bubble chart.
- Map : Geographic / thematic / choropleth map.
- Table : A pure table with no charts.
- Infographic : Text-heavy figure with icons, illustrations, or flow elements; not a chart or table.
- Dashboard : A composite figure that combines two or more of the above in one image (e.g., a chart together with an inline table, or multiple charts of different types laid out together).
- Other : Does not fit any of the above.

Rules: - Choose EXACTLY ONE label.
- If the image contains multiple chart components (e.g., a bar chart and an inline table, or two different chart types side by side), use "Dashboard".
- Use "Other" only as a last resort.

Respond with ONLY the label (e.g., "Bar"). No explanation, no quotes.

Direct prompt

上記の質問に対して、正確かつ簡潔に答えてください。
(For the above question, answer accurately and concisely.)

CoT prompt

上記の質問に対して、ステップバイステップで考えてから答えてください。
最後の行は 'Answer: $ANSWER' の形式で、正確かつ簡潔に回答してください。
(For the above question, think step by step before answering.

Japanese Title	English Title	Topic Group	Examples
海上保安レポート2025	Japan Coast Guard 2025	Security	93
防衛白書2025	Defence of Japan 2025	Security	81
消防白書2024	Fire Service 2024	Security	81
交通安全白書2025	Traffic Safety 2025	Security	74
再犯防止推進白書2024	Recidivism Prevention 2024	Security	66
防災白書2025	Disaster Management 2025	Security	60
犯罪被害者白書2024	Crime Victims 2024	Security	40
警察白書2025	Police 2025	Security	9
小規模企業白書2025	Small Enterprises 2025	Economy	92
中小企業白書2025	Small and Medium Enterprises 2025	Economy	88
地方財政白書2025	Local Public Finance 2025	Economy	68
情報通信白書2025	Information & Communications 2025	Economy	65
日本経済レポート2024	Japan Economy 2024	Economy	62
世界経済の潮流2025-01	World Economic Trends 2025	Economy	62
通商白書2025	International Economy and Trade 2025	Economy	59
年次経済財政報告2025	Japanese Economy and Public Finance 2025	Economy	53
地域課題分析レポート2025-08	Regional Issues 2025	Economy	40
科学技術白書2025	Science & Technology 2025	Economy	36
消費者白書2023	Consumer 2023	Society	91
高齢社会白書2025	Aging Society 2025	Society	86
労働経済白書2024	Labor Economy 2024	Society	82
男女共同参画白書2025	Gender Equality 2025	Society	76
人事院白書2024	National Personnel 2024	Society	76
厚生労働白書2025	Health, Labour & Welfare 2025	Society	65
食料農業農村白書2024	Food, Agriculture and Rural Areas 2024	Infrastructure	83
水産白書2024	Developments in Japan's Fisheries 2024	Infrastructure	70
森林林業白書2024	Forest & Forestry 2024	Infrastructure	63
国土交通白書2024	Land, Infrastructure, Transport and Tourism 2024	Infrastructure	41
環境白書2025	Environment 2025	Energy & Environment	76
原子力白書2023	Nuclear Energy 2023	Energy & Environment	49
エネルギー白書2025	Energy 2025	Energy & Environment	29
開発協力白書2023	Development Cooperation 2023	Diplomacy	26
外交青書2025	Diplomatic Bluebook 2025	Diplomacy	11
Total			2,053

Table 6: The 33 Japanese white papers used in HakushoBench.

The final line should be in the format 'Answer: $ANSWER', and provide the answer accurately and concisely.)

Judge prompt

Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below.
When judging equivalence, allow variations in script or notation that convey the same meaning (e.g., '2羽' and '二羽' should be considered equivalent).
Treat the following cases as correct: - The extracted answer includes additional context (e.g., series name, author name, location, broader category) while still containing the correct_answer exactly or as its unambiguous, specific instance. (For example, "富嶽三十六景江戸日本橋" is correct if the correct_answer is "江戸日本橋".) - The extracted answer is more specific than the [correct_answer] while remaining consistent with it. - The extracted answer is an alternate name, synonymous phrasing, or another commonly accepted way to refer to the same concept, object, place, or artwork. - The extracted answer omits information that is not essential to the correctness of the question. - Allow minor variations in spacing, capitalization, or script, as long as the core correct_answer is unambiguously present.
[question]: question
[response]: response
Your judgement must be in the format and criteria specified below:
extracted_final_answer: The final exact answer extracted from the [response]. Put the extracted answer as 'None' if there is no exact, final answer to extract from the response.
[correct_answer]: correct_answer
reasoning: Explain why the extracted_final_answer is correct or incorrect based on [correct_answer], focusing only on if there are meaningful differences between [correct_answer] and the ex-

Model \ Human	Bar	Dash.	Info.	Line	Map	Other	Pie	Table
Bar	**29**	1	0	0	0	0	0	0
Dashboard	9	**6**	1	1	2	1	0	0
Infographic	0	0	**9**	0	0	0	0	0
Line	0	1	0	**15**	0	0	0	0
Map	0	0	0	0	**3**	0	0	0
Other	0	0	1	0	0	**0**	0	0
Pie	0	0	0	0	0	0	**2**	0
Table	0	0	0	0	0	0	0	**19**

Table 7: Confusion matrix of image-type labels on 100 randomly sampled HakushoBench images (rows: Gemini 3 Pro; columns: human annotator). Overall agreement is 83/100.

	Glob.	Multi.	Count.	Ext.	Vis.	Other
Global (Glob.)	1,743	604	231	110	801	0
Multi-hop (Multi.)		634	40	39	206	0
Counting (Count.)			293	13	64	0
Ext. knowl. (Ext.)				148	34	0
Visual (Vis.)					897	0
Other						132

Table 8: Co-occurrence of question-type flags in HakushoBench. The diagonal gives the number of questions carrying each flag; off-diagonal cells give the number of questions carrying both flags. Other is assigned only when no other flag applies.

tracted_final_answer. Do not comment on any background to the problem, do not attempt to solve the problem, do not argue for any answer different than [correct_answer], focus only on whether the answers match.
correct: Answer 'yes' if extracted_final_answer matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer 'no' otherwise, i.e. if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect.
confidence: The extracted confidence score between 0% and 100% from [response]. Put 100 if there is no confidence score available.

G Question Types: Co-occurrence and Examples

Question-type flags are non-exclusive. Table 8 reports how often each pair of flags is assigned to the same question, and Figure 8 shows one representative question per type.

H Accuracy by Image Type and Question Type on HakushoBench

Figures 9 and 10 show the accuracy spread across models on HakushoBench, grouped by image type and by question type, respectively. Tables 9 and 10 report the underlying per-image-type and per-question-type accuracies for a representative subset of models.

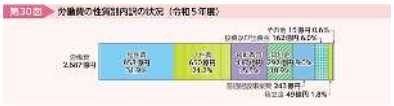
I Evaluation Results on Existing Chart and Table Benchmarks

Figures 11, 12, 13, and 14 visualize the per-model Direct and CoT accuracy on ChartQA, ChartQAPro, CharXiv, and JGraphQA as bar charts.

J Image Showcases for Comparison Benchmarks

Figures 15, 16, 17, and 18 show one randomly sampled image per image-type category for ChartQA, JGraphQA, ChartQAPro, and CharXiv, respectively.

Global



Q: 令和5年度の労働費のうち、2番目に大きな金額の費目は何ですか。 *(Among the FY2023 labor expenditure items, which item has the second-largest amount?)*
A: 人件費 *(Personnel expenses)*

Multi-hop



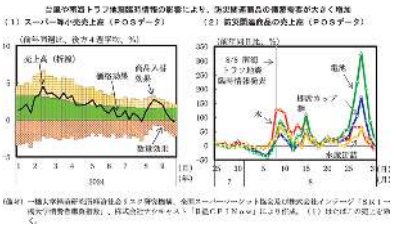
Q: 女性研究者の割合が日本の2倍以上の国は？ *(Which country has a share of female researchers more than twice that of Japan?)*
A: 英国 *(United Kingdom)*

Counting



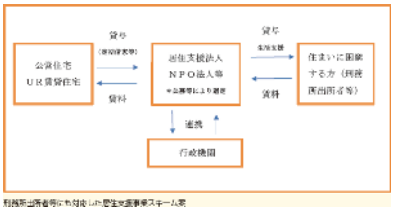
Q: この表で「AI」は何回記載されていますか？ *(How many times does "AI" appear in this table?)*
A: 2回 *(Twice)*

External knowledge



Q: (2)防災関連商品の売上高を示すグラフにおいて、水及び食料品ではないものは何ですか？ *(In graph (2), which shows sales of disaster-preparedness products, which item is neither water nor a food product?)*
A: 電池 *(Batteries)*

Visual



Q: 右から左に向かう矢印が表しているものは何ですか。 *(What does the arrow pointing from right to left represent?)*
A: 賃料 *(Rent)*

Other



地域名	都道府県名
北海道	北海道
東北	青森県、岩手県、宮城県、秋田県、山形県、福島県
関東 北関東	茨城県、栃木県、群馬県
関東 南関東	埼玉県、千葉県、東京都、神奈川県
甲信越	新潟県、山梨県、長野県
東海	静岡県、岐阜県、愛知県、三重県
北陸	富山県、石川県、福井県
近畿	滋賀県、京都府、大阪府、兵庫県、奈良県、和歌山県
中国	鳥取県、島根県、岡山県、広島県、山口県
四国	徳島県、香川県、愛媛県、高知県
九州	福岡県、佐賀県、長崎県、熊本県、大分県、宮崎県、鹿児島県
沖縄	沖縄県

Q: 示された表において、東海地域の最後に記載されている県はどこですか。 *(In the table shown, which prefecture is listed last within the Tokai region?)*
A: 三重県 *(Mie Prefecture)*

Figure 8: One representative question per question-type flag in HakushoBench, together with its source image. English glosses are ours; models are evaluated on the original Japanese questions.

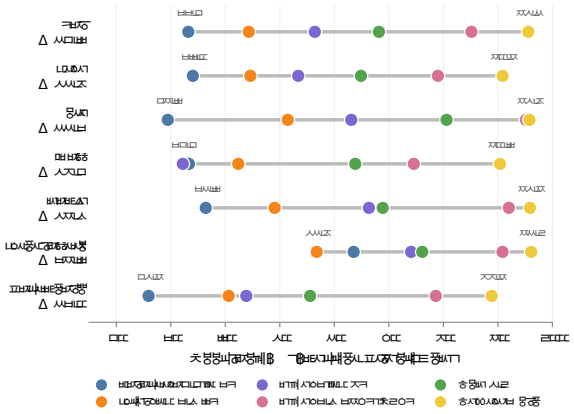


Figure 9: Accuracy spread across models on HakushoBench, grouped by image type. Categories with fewer than 50 examples (Area, Scatter, Bubble, and Other) are omitted.

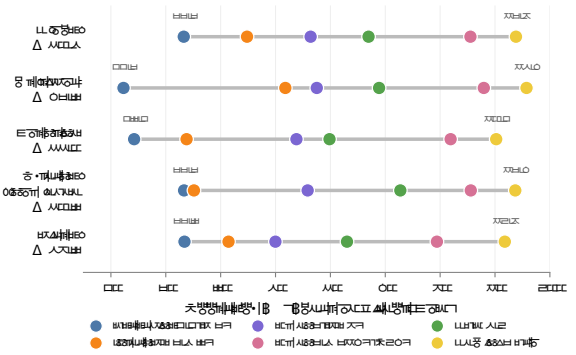


Figure 10: Accuracy spread across models on HakushoBench, grouped by question type.

Image type	N	Sarashina2.2-V 3B	InternVL 3.5 4B	Qwen3-VL 8B	Qwen3.5 397B-A17B	GPT 5.1	Gemini 3 Pro
Bar	550	33.2	44.3	56.4	85.2	68.2	**95.6**
Line	281	34.0	44.6	53.4	79.0	64.9	**90.9**
Pie	103	29.4	51.5	63.1	95.1	80.6	**95.8**
Map	59	33.3	42.4	32.2	74.6	63.8	**90.4**
Table	422	36.4	49.1	66.4	92.0	68.9	**95.9**
Infographic	310	63.5	56.8	74.1	90.9	76.1	**96.1**
Dashboard	279	25.9	40.6	43.8	78.6	55.6	**88.9**

Table 9: Per-image-type accuracy on HakushoBench (best of Direct/CoT per model). Rows are the image-type categories with at least 50 examples (Section 4); N is the number of HakushoBench examples of that type. Each cell is the mean accuracy (%) over three independent runs. The highest entry in each row is bolded.

Question type	N	Sarashina2.2-V 3B	InternVL 3.5 4B	Qwen3-VL 8B	Qwen3.5 397B-A17B	GPT 5.1	Gemini 3 Pro
Global	1743	33.3	44.8	56.4	85.5	67.0	**93.8**
Multi-hop	634	22.3	51.8	57.5	88.0	68.9	**95.7**
Counting	293	24.2	33.8	53.8	81.9	59.8	**90.2**
External knowledge	148	33.3	35.1	55.9	85.6	72.7	**93.7**
Visual	897	33.4	41.4	50.0	79.4	63.0	**91.8**

Table 10: Per-question-type accuracy on HakushoBench (best of Direct/CoT per model). Rows are the question-type flags defined in Section 3.2 (Other omitted); N is the number of HakushoBench examples whose annotator marked that flag (a single example may contribute to multiple rows). Each cell is the mean accuracy (%) over three independent runs; the highest entry per row is bolded.

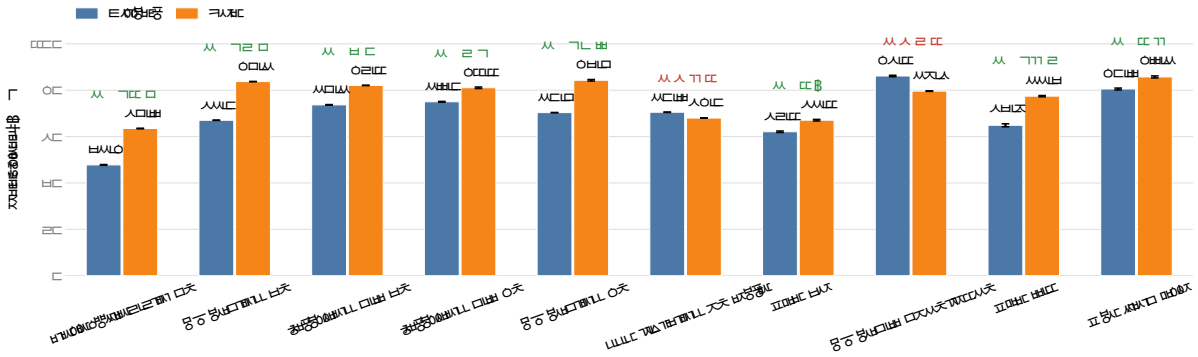


Figure 11: Performance of each model on ChartQA.

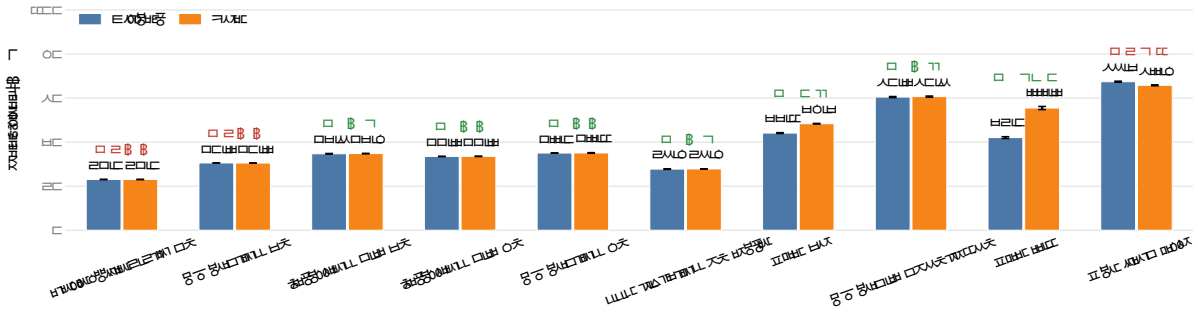


Figure 12: Performance of each model on ChartQAPro.

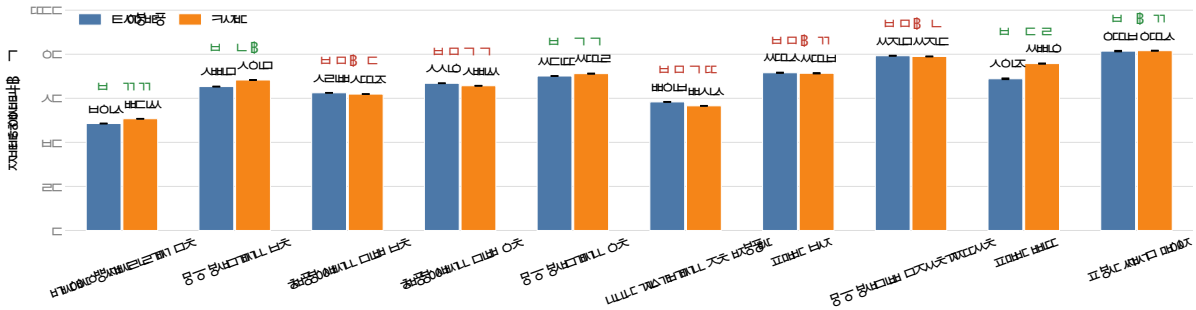


Figure 13: Performance of each model on CharXiv.

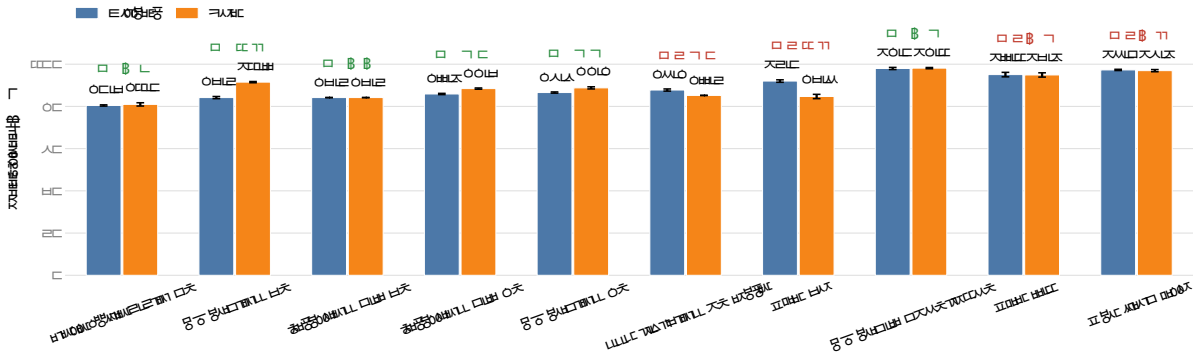


Figure 14: Performance of each model on JGraphQA.

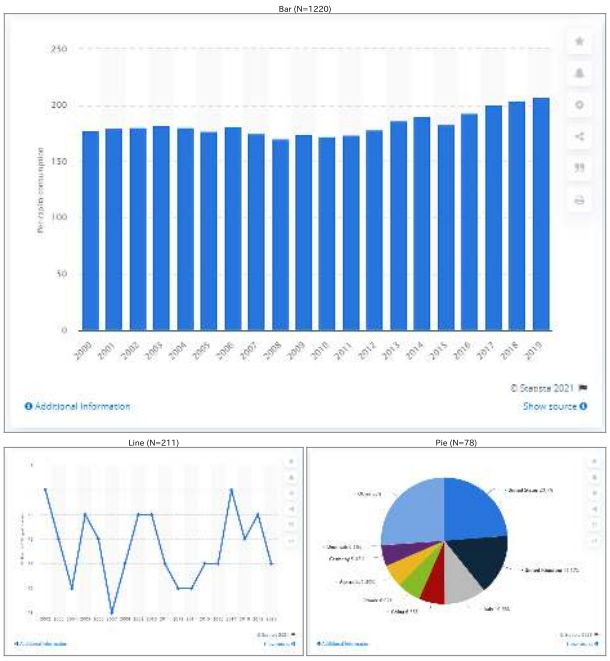


Figure 15: ChartQA: one randomly sampled image per image type.



Figure 16: JGraphQA: one randomly sampled image per image type.

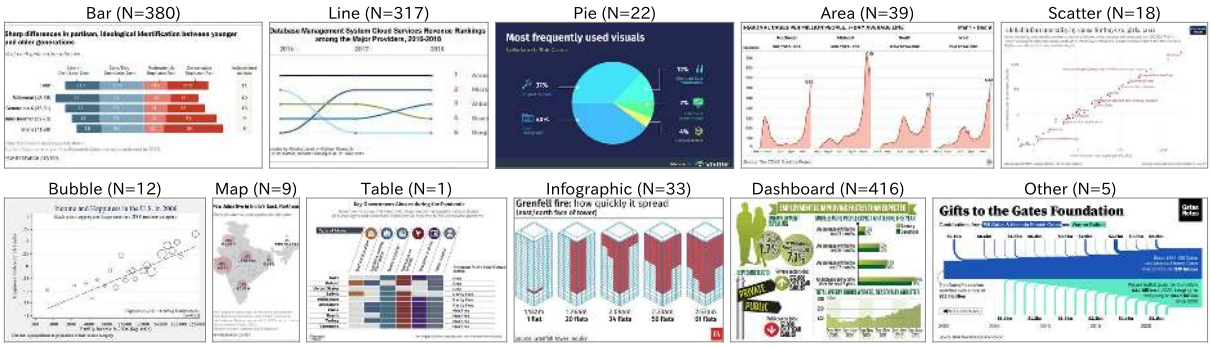


Figure 17: ChartQAPro: one randomly sampled image per image type.

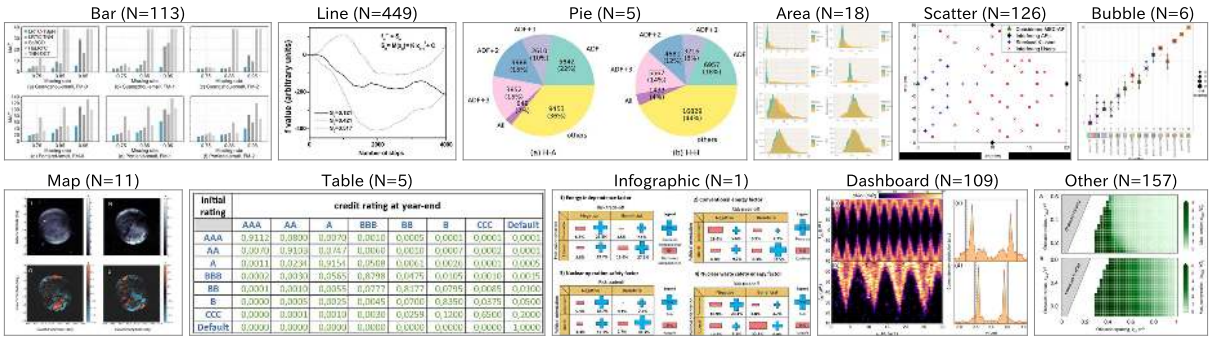


Figure 18: CharXiv: one randomly sampled image per image type.