# JaWildText: A Benchmark for Vision-Language Models on Japanese Scene Text Understanding

Koki Maeda[1,2][0009−0008−0529−3152] and Naoaki Okazaki[1,2][0000−0001−7635−6175]

[1] Institute of Science Tokyo, Tokyo, Japan
[2] Research and Development Center for Large Language Models, National Institute of Informatics, Tokyo, Japan
{koki.maeda@nlp.,okazaki@}comp.isct.ac.jp

**Abstract.** Japanese scene text poses challenges that multilingual benchmarks often fail to capture, including mixed scripts, frequent vertical writing, and a character inventory far larger than the Latin alphabet. Although Japanese is included in several multilingual benchmarks, these resources do not adequately capture the language-specific complexities. Meanwhile, existing Japanese visual text datasets have primarily focused on scanned documents, leaving in-the-wild scene text underexplored. To fill this gap, we introduce **JaWildText**, a diagnostic benchmark for evaluating vision-language models (VLMs) on Japanese scene text understanding. JaWildText contains 3,241 instances from 2,961 images newly captured in Japan, with 1.12 million annotated characters spanning 3,643 unique character types. It comprises three complementary tasks that vary in visual organization, output format, and writing style: (i) Dense Scene Text Visual Question Answering (STVQA), which requires reasoning over multiple pieces of visual text evidence; (ii) Receipt Key Information Extraction (KIE), which tests layout-aware structured extraction from mobile-captured receipts; and (iii) Handwriting OCR, which evaluates page-level transcription across various media and writing directions. We evaluate 14 open-weight VLMs and find that the best model achieves an average score of 0.64 across the three tasks. Error analyses show recognition remains the dominant bottleneck, especially for kanji. JaWildText enables fine-grained, script-aware diagnosis of Japanese scene text capabilities, and will be released with evaluation code.

**Keywords:** Japanese scene text · vision-language models · benchmark · handwriting recognition · key information extraction

## 1 Introduction

Text is ubiquitous in everyday environments: on street posters, handwritten notes, receipts, and storefronts. For decades, text-centric vision systems relied on a modular workflow that first applied OCR to convert pixels into characters and then fed the text to separate modules for downstream tasks [20]. With the rise of

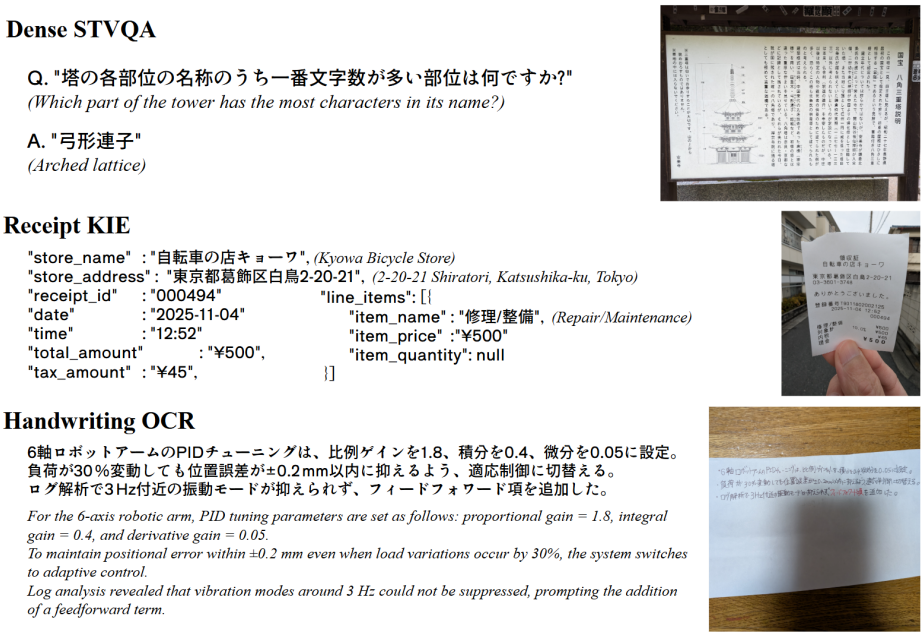



Fig. 1: Overview of the JaWildText benchmark: (i) Dense STVQA, (ii) Receipt KIE, and (iii) Handwriting OCR. We added English translations for readability.

vision-language models (VLMs) such as GPT-4V [33], this workflow is shifting toward an end-to-end approach: VLMs generate outputs of a downstream task directly from natural images. This new workflow is increasingly adopted as a practical alternative to dedicated OCR pipelines.

This shift, however, complicates evaluation. When a VLM is evaluated only by downstream task accuracy, it is often unclear whether an error stems from a failure of character recognition or from incorrect reasoning over correctly recognized text. Disentangling these two failure modes is critical because precise reading is a prerequisite for any higher-level text understanding. Benchmarks that assess reading in natural images and separate recognition failures from reasoning failures are therefore essential, yet few existing resources offer this diagnostic capability, particularly for non-English scripts.

Although multilingual benchmarks [29,30,39] have extended scene text evaluation beyond English [4,22,38], they prioritize language breadth over language-specific diagnostics. This matters particularly for Japanese, whose scene text frequently mixes kanji, hiragana, katakana, and Latin alphanumerics. Because the resulting failure patterns differ from those of other scripts, it is difficult to diagnose failures without targeted evaluation. Existing Japanese scene text resources provide recognition data at the character or word level [12, 15, 28]. In contrast, Japanese VLM benchmarks target scanned documents or knowledge-centric multimodal tasks [31, 32] without explicitly measuring the underlying

recognition ability. As a result, no existing benchmark can tell whether VLM errors on Japanese scene text arise from recognition or from reasoning.

To fill this gap, we introduce JaWildText, a fine-grained benchmark for evaluating VLMs on Japanese scene text understanding. The benchmark is designed to disentangle reading from reasoning. As shown in Figure 1, JaWildText consists of three tasks that form a compact yet comprehensive configuration. The tasks are chosen to vary three factors that commonly confound end-to-end evaluation: visual organization (from cluttered to structured), output format (from free-form to verbatim), and writing style (printed or handwritten). This design exposes failure modes that remain conflated when models are scored only by downstream task accuracy. Dense Scene Text Visual Question Answering (Dense STVQA) tests multi-region reading and cross-reference reasoning in cluttered signboards and posters. Receipt Key Information Extraction (Receipt KIE) evaluates layout-aware structured extraction from in-the-wild imagery. Handwriting OCR (page-level transcription) assesses long-context transcription of handwritten text, providing a recognition-dominant setting that complements the reasoning-heavy tasks above. JaWildText contains 3,241 evaluation instances from 2,961 newly collected images in Japan, with 1.12 million annotated characters spanning 3,643 unique characters.

We benchmark 14 open-weight VLMs on JaWildText. The experiments show that the best model achieves an average score of 0.64 across the three tasks. Our error analysis identified distinct bottlenecks: models that read text accurately may still fail at reasoning, and recognition difficulty varies drastically by script type. In summary, these results demonstrate that JaWildText provides fine-grained diagnostic evidence that is invisible in aggregated accuracy alone.

The contributions of this paper are as follows:

1. We introduce JaWildText, to our knowledge, the first benchmark dedicated to evaluating VLMs on Japanese scene text understanding across three complementary tasks grounded in real-world images.
2. We benchmark 14 open-weight VLMs, establish reproducible baselines, and quantify substantial performance gaps across architectures.
3. We provide an error analysis that disentangles recognition from reasoning failures, showing that their relative severity varies markedly across model families.

## 2 Related Work

### 2.1 Benchmarking Text Understanding in Natural Images

For English, evaluation resources have matured along three complementary tracks. *Scene text* benchmarks first targeted detection and recognition in natural images [40, 45], and then advanced to text-centric VQA [4, 18, 22, 23, 38], which requires models to read and reason over recognized text. *Receipt and document understanding* benchmarks such as SROIE [13], CORD [34], and FUNSD [17] evaluate structured key information extraction (KIE), testing whether models

can map visually organized fields to predefined categories. *Handwriting recognition* benchmarks, anchored by the IAM Handwriting Database [21], assess verbatim transcription of diverse writing styles; recent work shows that VLM performance degrades substantially on non-English handwriting [7]. Together, these tracks span a range of visual organization, output format, and writing style, forming a comprehensive evaluation ecosystem for English.

Several benchmarks extend this ecosystem to other languages, progressively broadening language coverage and task complexity. The ICDAR MLT challenges [29,30] introduce multilingual scene text detection and recognition across up to ten languages. XFUND [43] extends form understanding to seven languages. Targeting VLMs directly, MTVQA [39] shifted the focus to multilingual text-centric VQA with native annotations. OCRBench [9,19] and CC-OCR [44] broaden the scope to multilingual OCR for VLMs. While these efforts increase language coverage, they treat each language as one among many and provide limited diagnostic depth for language-specific challenges.

Among CJK languages, dedicated benchmarks have emerged for Chinese scene text recognition [5] and Korean text-centric VQA [14]. However, Japanese remains without a comprehensive evaluation despite its unique challenges, notably concurrent use of multiple scripts within a single text, complex layouts, and thousands of distinct characters.

### 2.2 Japanese Text Understanding

Existing Japanese-specific resources address isolated facets of text understanding. For scene text recognition, existing resources target scene text spotting in omnidirectional video [16], isolated character classification [12], vertical text recognition [36], and comics [2], restricted to a specific visual setting or textual granularity. For receipt understanding, existing resources support training or fine-tuning on mobile-captured receipts and post-OCR correction [10, 27], but none serve as a benchmark for assessing general-purpose VLM capabilities. For handwriting, existing datasets provide isolated characters or online stroke data [24–26, 28], or target classical cursive [6]; none covers page-level offline recognition of modern handwriting. On the reasoning side, JDocQA [31] addresses question answering over scanned documents, and JMMMU [32] benchmarks multimodal understanding centered on cultural and academic knowledge rather than text recognition ability.

Mapping these resources onto the three evaluation dimensions of visual organization, output format, and writing style reveals that none connect recognition with reasoning for Japanese scene text. JaWildText fills this gap with three complementary tasks that systematically vary these dimensions, enabling fine-grained diagnosis of where and why current VLMs fail on Japanese text in real-world images.

## 3 Dataset: JaWildText

JaWildText is designed to expose where a model fails, whether in character recognition, layout understanding, or reasoning, rather than reporting only aggregated task accuracy. To this end, it comprises three complementary tasks, each annotated to disentangle recognition errors from reasoning and formatting errors. Because such fine-grained diagnosis requires image diversity and annotation quality that are unavailable in web-scraped corpora, we collected original images and annotations tailored for this work.

### 3.1 Dense STVQA

Dense STVQA evaluates whether a model can read and reason over dense Japanese scene text, using visually complex real-world images such as signboards, bulletin boards, posters, and product packages.

**Image Collection.** To test recognition under realistic conditions, we asked workers from a data collection agency in Japan to photograph text-rich scenes with cameras and smartphones, resulting in 745 images. We instructed workers to cover indoor and outdoor locations under both daytime and nighttime lighting and avoid multiple shots of the same subject, ensuring diversity in layout, font style, and visual context. We retained natural artifacts, such as background clutter, partial occlusion, and reflections, to test recognition robustness.

**Annotation.** Annotations are structured into two layers to separate recognition from reasoning. In the first layer, annotators marked text regions with quadrilateral bounding boxes. They transcribed each region, which is defined as a line-level or column-level sequence of visually recognizable characters. Each image contains 45.1 annotated text regions on average. In the second layer, native Japanese speakers authored open-ended question-answer pairs over these transcribed regions. Annotators were encouraged to write questions that require reasoning across multiple text regions rather than extracting a single string from a single area. We exclude yes/no and multiple-choice formats to minimize chance-level correctness. Crucially, each question is linked to *evidence regions*: the minimal set of text regions necessary and sufficient to derive the answer. This linkage enables automatic diagnosis; if a model fails a question but correctly recognizes the evidence regions, the error is attributable to reasoning rather than recognition. In total, we created 1,025 question-answer pairs.

### 3.2 Receipt KIE

Receipt KIE evaluates structured field extraction from real-world photographs of Japanese receipts. This setting introduces challenges largely absent from scene text: rigid columnar layouts, mixed use of full-width and half-width characters, and domain-specific abbreviations produced by thermal printers.

**Image Collection.** Diagnostic value depends on testing under realistic capture conditions; hence, we collected photographs of consumer receipts from everyday transactions rather than flatbed scans. We retained natural artifacts such

as creases, folds, and hand-held tilt to ensure that models are evaluated against the geometric and photometric distortions encountered in practical use. To maximize visual diversity, we disallowed multiple receipts from the same store while permitting receipts from different branches of the same chain, yielding 1,151 unique receipt images.

**Annotation.** We annotate receipts at the individual field level so that evaluation can pinpoint which field types a model struggles with, rather than producing only a single per-receipt score. Our key schema builds on the four header fields of SROIE [13] (`store_name`, `date`, `store_address`, `total_amount`) and extends it in two directions tailored to Japanese receipts: three additional header fields (`receipt_id`, `time`, `tax_amount`) that are usually printed but absent from SROIE, and line-item-level tuples of `item_name`, `item_price`, and `item_quantity`, which test a model's ability to maintain structured alignment across repeated rows. For each field, annotators recorded the text string and a quadrilateral bounding box; fields absent from a receipt are explicitly marked as null, allowing evaluation to distinguish extraction errors from correct recognition of absence. In total, we annotated 56,095 text regions across 1,151 receipts.

### 3.3 Handwriting OCR

Handwriting OCR evaluates page-level recognition of multi-sentence handwritten Japanese text. Unlike isolated-word or single-line recognition, page-level evaluation requires models to handle layout interpretation, line segmentation, and script mixing simultaneously, reflecting realistic reading scenarios.

**Image Collection.** To obtain diverse yet controlled handwriting samples, we designed a collection pipeline that separates content generation from handwriting production. First, we defined over 100 genre-keyword pairs spanning everyday topics such as work planning, travel, and cooking. We generated up to 20 prompt texts per pair using a large language model.[3] Each prompt was constrained to approximately 100 characters, long enough to span multiple lines across mixed scripts yet short enough to fit naturally on a single page.

Then, we distributed the prompts to 51 native Japanese writers, who transcribed them onto designated media and photographed the results using their own devices. To systematically introduce visual variation, we specified the writing medium and writing direction (horizontal or vertical) for each instance. The media include lined paper, unlined plain paper, whiteboards, and tablets, while writers choose line-break positions freely. Each instance is accompanied by metadata recording the writer ID, writing medium, writing instrument, ink color, and writing direction, enabling fine-grained analysis of how these factors affect recognition performance.

**Annotation.** For each image, the target output is a transcription of all visible handwritten text. Annotators transcribed the text line by line and drew a quadrilateral bounding box around each region. A key design principle is that

[3] Specifically, we used `openai/gpt-oss-120b` to generate the prompt texts. The generated texts serve only as writing prompts.

Table 1: Summary statistics of JaWildText. #**Instances** denotes the evaluation unit: an instance is a question–answer pair in Dense STVQA, and a single image in Receipt KIE and Handwriting OCR. #**Regions** denotes the number of annotated quadrilateral text regions.

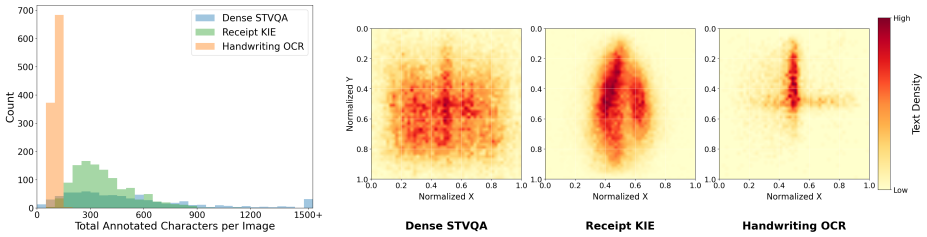
| Subset | #Instances | #Images | #Regions | #Chars | #Unique Chars |
|---|---|---|---|---|---|
| Dense STVQA | 1,025 | 745 | 33,608 | 571,979 | 3,313 |
| Receipt KIE | 1,151 | 1,151 | 56,095 | 433,558 | 2,001 |
| Handwriting OCR | 1,065 | 1,065 | 6,002 | 111,977 | 1,830 |
| **Total** | **3,241** | **2,961** | **95,705** | **1,117,514** | **3,643** |


Fig. 2: Representative images from each task, illustrating the diversity of JaWildText. Dense STVQA covers signboards, posters, and product packages under varying conditions. Receipt KIE includes receipts with creases, folds, and diverse perspectives. Handwriting OCR spans multiple writing media and directions.

the ground truth should reflect what is visually present in the image rather than what the writer intended. Accordingly, we preserved writer-introduced errors such as misspellings or omitted characters. For characters that a writer started but left incomplete due to writing errors, we assigned a dedicated symbol (□) to mark them explicitly. This annotation policy ensures that model evaluation measures visual recognition fidelity rather than error correction ability. We annotated 1,065 handwriting instances with 6,002 lines, totaling 111,977 characters.

### 3.4 Quality Control

Image collection and annotation were conducted by a professional data curation agency with compensated annotators. To calibrate annotation guidelines before full-scale production, the authors and the agency jointly reviewed the first 10%




(a) Distribution of total character length per image.

(b) Spatial distribution of text-region centers.

Fig. 3: Image-level text properties in JaWildText.

of deliverables and refined the guidelines based on observed inconsistencies. In the main phase, each instance was labeled by one annotator and independently verified by a second; disagreements were resolved through discussion. We excluded images containing non-public personally identifiable information, such as faces, vehicle license plates, or credit card numbers, while retaining publicly displayed information (e.g., store phone numbers on receipts) needed for the benchmark tasks. The dataset, including all images and annotations, will be publicly released under the Apache License 2.0.[4]

### 3.5 Dataset Statistics

Table 1 summarizes key statistics of JaWildText. The dataset comprises 3,241 instances drawn from 2,961 unique images, with 95,705 annotated text regions totaling 1,117,514 characters across 3,643 unique characters. By design, each task comprises approximately 1,000 instances, allowing for broadly comparable score precision across tasks. Figure 2 shows representative images from each task, and Figure 3 visualizes text density and layout properties discussed below.

**Text Density and Spatial Layout.** Dense STVQA and Receipt KIE are text-dense, often containing several hundred characters per image, whereas Handwriting OCR is intentionally controlled to approximately 100 characters per image (Figure 3a). Dense STVQA exhibits a long tail beyond 2,000 characters per image, reflecting the high information density of signboards and bulletin boards; this density forces models to locate and integrate information across many text regions, making the task sensitive to both recognition errors and cross-region reasoning failures. In contrast, the controlled length of Handwriting OCR isolates recognition ability from reasoning, providing a setting in which errors can be attributed almost entirely to character-level reading. The spatial distribution of text-region centers (Figure 3b) mirrors these design choices: Dense STVQA regions spread broadly across the frame, Receipt KIE regions form a narrow vertical band consistent with elongated receipt layouts, and Handwriting OCR clusters near the page center along both axes.

[4] https://huggingface.co/datasets/llm-jp/jawildtext

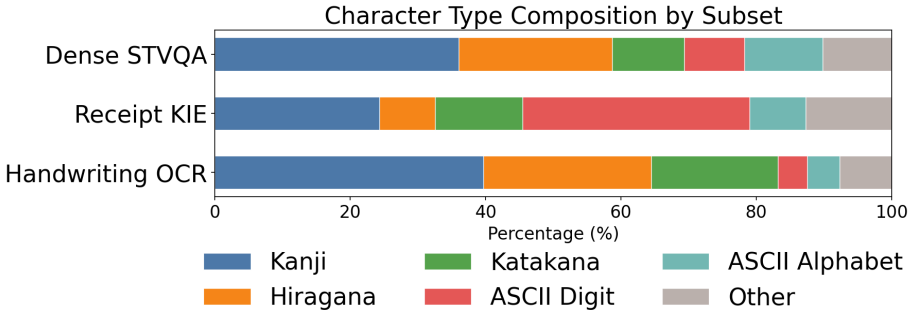



Fig. 4: Stacked character-type composition by task. Japanese scripts dominate Dense STVQA and Handwriting OCR, while Receipt KIE allocates a large fraction to ASCII digits.

Table 2: Question-type distribution in Dense STVQA.

| **Question type** | **# (%)** |
|---|---|
| Compositional retrieval | 406 (39.6%) |
| Counting | 374 (36.5%) |
| Calculation | 125 (12.2%) |
| Spatial | 84 (8.2%) |
| Other | 36 (3.5%) |

Table 3: Receipt KIE field fill rates.

| **Field** | **Fill Rate** |
|---|---|
| `store_name` | 100.0% |
| `date` | 100.0% |
| `total_amount` | 99.8% |
| `time` | 98.3% |
| `tax_amount` | 95.2% |
| `receipt_id` | 92.0% |
| `store_address` | 48.9% |

Table 4: Writing surface and direction in Handwriting OCR.

| *Medium* | *#Instances (%)* |
|---|---|
| Lined paper | 470 (43.9%) |
| Unlined paper | 344 (32.3%) |
| Tablet | 130 (12.2%) |
| Whiteboard | 121 (11.4%) |
| *Direction* | *#Regions (%)* |
| Horizontal | 4,222 (70.3%) |
| Vertical | 1,780 (29.7%) |

**Character-type Composition.** Character-type distributions differ markedly across tasks (Figure 4), which foreshadow the character-type analysis in Section 4.3. Dense STVQA and Handwriting OCR are dominated by kanji (36.0% and 39.7%) and hiragana (22.6% and 24.8%), consistent with natural Japanese prose and placing heavy demands on mixed-script recognition. Receipt KIE contains a much larger share of ASCII digits (33.5%), reflecting prices, quantities, and dates; accurate digit reading is therefore a decisive factor for extraction performance in this task. Overall, JaWildText contains 2,866 unique kanji characters. Of the 2,136 Jōyō kanji (the Japanese government's daily-use list), the dataset covers 1,985 (92.9%), and it additionally includes 881 kanji beyond the Jōyō set, meaning that models must generalize beyond standard literacy inventories to handle real-world text.

**Task-specific properties.** Each task contributes a distinctive evaluation signal. In Dense STVQA, questions cover a balanced mix of reasoning types: compositional retrieval, counting, calculation, and spatial reasoning (Table 2). Questions are linked to minimal evidence regions, making it possible to separate recognition errors (failing to read individual regions) from reasoning errors (failing to combine correctly read evidence). In Receipt KIE, field fill rates vary sub-

stantially (Table 3): `store_name` and `date` are present in all receipts, whereas `store_address` has the lowest fill rate (48.9%), reflecting real-world omission patterns and testing whether models can handle missing fields without hallucinating content. In Handwriting OCR, 51 writers contributed data with controlled numbers of instances per writer, spanning multiple writing media and directions (Table 4), ensuring that recognition performance is evaluated across a range of handwriting variability rather than being biased toward a single style.

## 4 Experiments

### 4.1 Experimental Setup

**Inference.** We evaluate 14 open-weight VLMs from five model families. We include four recent high-performing families: Qwen3-VL [3], InternVL3.5 [41], Gemma3 [11], and Phi-4-Multimodal [1]. For families offering multiple sizes, we select variants with 1B to 38B parameters (see Table 5 for the complete list). We additionally include Sarashina2.2-Vision [37], a model built on a Japanese-centric LLM backbone and trained with Japanese document and OCR data. This selection lets us examine how model scale and language-specific training influence performance on JaWildText. We set the temperature to 0 and the maximum token length to 2,048. Each instance uses a single image at its original resolution; resizing or tiling is applied according to the default preprocessing.

**Evaluation.** To reliably extract the final answer from model output, we enforce machine-parseable output formats using a fixed prompt template for each task. Models must enclose the final answer in `\boxed{...}` for Dense STVQA, return a single JSON object following the predefined schema for Receipt KIE, and output plain text transcriptions for Handwriting OCR. Any output that cannot be parsed receives a score of 0. Before scoring, we apply Unicode NFKC normalization to both predictions and references to absorb superficial character-form differences. In the Dense STVQA task, answers are open-ended and may vary due to differences in units or paraphrasing. Thus, we adopt judge-based accuracy: an LLM verifier compares each prediction against the reference and returns a binary correctness label.[5] Combined with the evidence region annotations, this binary signal enables error analysis to attribute each failure to recognition or reasoning. In the Receipt KIE task, outputs are structured, and field boundaries are well-defined. We report overall F1 and field-level accuracy for major header fields. This indicates which field types are most challenging to extract. For Handwriting OCR, we compute character-level similarity as $\max(0,\ 1 - \mathrm{CER})$, where CER is defined as the Levenshtein distance between prediction and reference, divided by reference length. Character-level scoring is suitable for Japanese because it lacks explicit word boundaries. We further report script-type breakdown (e.g., kanji vs. hiragana) in CER. We compute the overall score as the unweighted average of the three task scores.

[5] We employ `openai/gpt-5.1-2025-11-13` via the Azure OpenAI API as the judge model. We will release the verifier prompt to reproduce scoring.

Table 5: Results on the JaWildText benchmark.

| Model | Params | Overall | Dense STVQA | Receipt KIE | Handwriting OCR |
|---|---|---|---|---|---|
| | | | Accuracy | F1 | 1-CER |
| Qwen3-VL-8B | 8B | **0.64** | **0.62** | **0.53** | **0.79** |
| Qwen3-VL-4B | 4B | 0.60 | 0.52 | 0.50 | 0.77 |
| Qwen3-VL-2B | 2B | 0.52 | 0.31 | 0.48 | 0.76 |
| InternVL3.5-38B | 38B | 0.55 | 0.44 | 0.50 | 0.71 |
| InternVL3.5-14B | 14B | 0.49 | 0.30 | 0.47 | 0.71 |
| InternVL3.5-8B | 8B | 0.53 | 0.39 | 0.48 | 0.72 |
| InternVL3.5-4B | 4B | 0.48 | 0.31 | 0.44 | 0.70 |
| InternVL3.5-2B | 2B | 0.44 | 0.23 | 0.42 | 0.66 |
| InternVL3.5-1B | 1B | 0.37 | 0.11 | 0.37 | 0.61 |
| Sarashina2.2-Vision-3B | 3B | 0.50 | 0.44 | 0.40 | 0.68 |
| Gemma3-27B-IT | 27B | 0.43 | 0.37 | 0.39 | 0.53 |
| Gemma3-12B-IT | 12B | 0.40 | 0.32 | 0.38 | 0.51 |
| Gemma3-4B-IT | 4B | 0.19 | 0.12 | 0.23 | 0.20 |
| Phi-4-Multimodal | 14B | 0.18 | 0.008 | 0.23 | 0.29 |


Fig. 5: Performance scaling trends with model size across benchmark tasks. Larger models improve overall performance, but gains differ by task family.

### 4.2 Main Results

Table 5 summarizes the performance of all evaluated models on JaWildText. The best model, Qwen3-VL-8B, achieves an overall score of only 0.64, indicating that Japanese scene text understanding remains a substantial challenge for current open-weight VLMs. Dense STVQA exhibits the widest performance spread across models (0.008–0.62), suggesting that this task effectively differentiates models with varying levels of Japanese scene text capability. In Handwriting OCR, 10 of 14 models score at least 0.60, indicating that most models already possess a baseline handwriting recognition ability, though a ceiling around 0.80 persists even for the strongest models. Performance differences across model families are significant: Qwen3-VL consistently outperforms InternVL3.5 at similar parameter scales, surpassing InternVL3.5 by 0.11 Overall (0.64 vs. 0.53) at 8B parameters. Gemma3 trails both families despite having up to $\sim 3.4\times$ more pa-

Table 6: Receipt KIE field-level accuracy on JaWildText.

| Model | Params | Store Name | Address | Receipt ID | Date | Time | Total | Tax |
|---|---|---|---|---|---|---|---|---|
| Qwen3-VL-8B | 8B | 0.14 | 0.52 | 0.13 | **0.81** | 0.91 | 0.81 | 0.77 |
| Qwen3-VL-4B | 4B | **0.16** | **0.55** | 0.11 | 0.28 | 0.92 | **0.84** | 0.80 |
| Qwen3-VL-2B | 2B | 0.11 | 0.43 | 0.08 | 0.52 | 0.88 | 0.82 | **0.82** |
| InternVL3.5-38B | 38B | 0.15 | 0.34 | 0.18 | 0.52 | **0.92** | 0.83 | 0.82 |
| InternVL3.5-14B | 14B | 0.15 | 0.35 | 0.13 | 0.26 | 0.89 | 0.83 | 0.79 |
| InternVL3.5-8B | 8B | 0.11 | 0.31 | 0.13 | 0.50 | 0.89 | 0.79 | 0.74 |
| InternVL3.5-4B | 4B | 0.14 | 0.29 | **0.21** | 0.26 | 0.90 | 0.77 | 0.73 |
| InternVL3.5-2B | 2B | 0.13 | 0.30 | 0.11 | 0.24 | 0.89 | 0.77 | 0.75 |
| InternVL3.5-1B | 1B | 0.09 | 0.23 | 0.10 | 0.33 | 0.80 | 0.69 | 0.70 |
| Sarashina2.2-Vision-3B | 3B | 0.11 | 0.28 | 0.18 | 0.28 | 0.75 | 0.72 | 0.67 |
| Gemma3-27B-IT | 27B | 0.11 | 0.20 | 0.08 | 0.52 | 0.71 | 0.72 | 0.64 |
| Gemma3-12B-IT | 12B | 0.06 | 0.21 | 0.11 | 0.39 | 0.67 | 0.70 | 0.60 |
| Gemma3-4B-IT | 4B | 0.03 | 0.10 | 0.05 | 0.20 | 0.59 | 0.59 | 0.35 |
| Phi-4-Multimodal | 14B | 0.01 | 0.004 | 0.07 | 0.13 | 0.51 | 0.51 | 0.48 |

rameters than the best-performing model. At the lower end, Phi-4-Multimodal attains near-zero accuracy on Dense STVQA (0.008), struggling to follow the required output format. Notably, raw parameter count does not fully explain these gaps. Sarashina2.2-Vision-3B achieves 0.44 accuracy on Dense STVQA, matching InternVL3.5-38B despite fewer parameters. However, this advantage does not generalize to Receipt KIE or Handwriting OCR, where Sarashina2.2-Vision-3B is comparable to InternVL3.5-2B. This contrast suggests that the benefit of Japanese-centric training data may be task-dependent.

To examine whether each task captures scaling behavior effectively, we plot performance trends within model families as parameter count (Figure 5). Within each family, all three tasks show improvement with scale, but their trajectories differ. Dense STVQA and Receipt KIE continue to improve across the parameter range we evaluate, with no clear saturation point. Handwriting OCR, by contrast, plateaus beyond a certain scale within each model family: InternVL3.5 saturates around 0.70 from 4B onward, and Qwen3-VL shows a marginal gain from 2B to 8B (0.76→0.79). Across families, however, scale is not decisive: Qwen3-VL-8B surpasses InternVL3.5-38B by 0.09 Overall, indicating that architecture and training data composition can matter more than parameter count alone.

The header-field accuracies in Table 6 exhibit sharp variation across field types in Receipt KIE. Format-constrained fields such as `time` are relatively well handled, with several top models achieving accuracy above 0.90. In contrast, `store_name` and `store_address` remain difficult, with best accuracies of only 0.16 and 0.55, respectively; these fields often require aggregating non-contiguous text spans across the receipt layout rather than copying a single contiguous line. This gap indicates that the primary bottleneck in Receipt KIE is not character recognition alone but spatial reasoning over document layout.

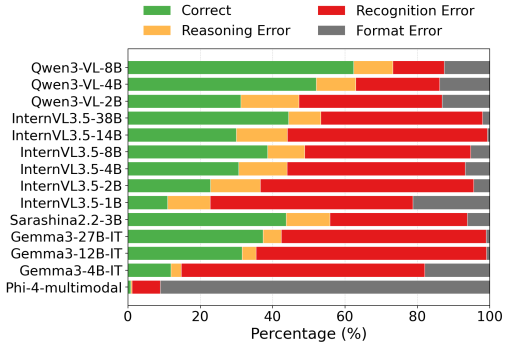



Fig. 6: Error taxonomy on Dense STVQA. Each bar decomposes instances into Correct, Reasoning Error, Recognition Error, and Format Error.

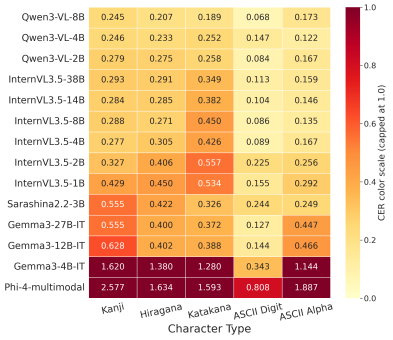



Fig. 7: Character-type CER on Handwriting OCR. Color scale is capped at 1.0; values exceeding 1.0 indicate hallucination-dominant outputs.

### 4.3 Analysis

**Error taxonomy for Dense STVQA.** To disentangle recognition failures from reasoning failures on Dense STVQA, we define an error taxonomy with four categories. For each Dense STVQA image, we separately prompt the model to transcribe all visible text, independently of the QA task. We then compare the resulting transcript against the ground-truth transcriptions of each question's annotated evidence regions: an evidence region is considered "read" if its whole ground-truth string appears as an exact substring in the transcript. Based on this comparison, each instance is assigned one of four outcomes. **Recognition Error**: the answer is not recoverable from the recognized text alone because at least one required evidence region is missing from the transcript. **Reasoning Error**: all evidence regions are present, but the final answer is incorrect. **Format Error**: the output cannot be parsed under the prescribed `\boxed{}` format. **Correct**: the parsed answer matches the reference.

Figure 6 illustrates that Recognition Error is the largest category for most models, indicating that Japanese scene text recognition remains the primary bottleneck. Qwen3-VL-8B, which achieves the highest Correct rate (62.3%), still exhibits a 14.2% Recognition Error rate, showing that even the best-performing model has not fully overcome this bottleneck. Sarashina2.2-Vision-3B shows a relatively low Recognition Error rate (38.0%) compared to other models, such as InternVL3.5-8B (45.9%) and Gemma3-27B-IT (56.7%), suggesting that Japanese-centric training may partially improve scene text recognition capability. InternVL3.5 and Gemma3 remain heavily dominated by Recognition Errors (44.7–58.9% and 56.7–67.3%, respectively). Phi-4-Multimodal is instead dominated by Format Errors, failing to follow the prescribed output format. This

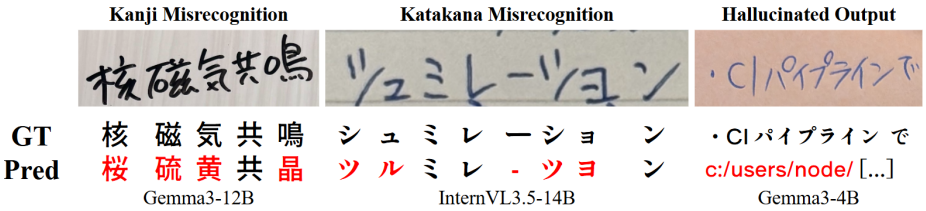



Fig. 8: Representative failure cases on Handwriting OCR. (Left) Gemma3-12B misrecognizes kanji characters, substituting visually dissimilar characters. (Center) InternVL3.5-14B confuses visually similar katakana characters. (Right) Gemma3-4B produces hallucinated output entirely unrelated to the input image. Red bold text indicates erroneous characters in the model predictions.

taxonomy makes visible the failure stage that aggregate accuracy alone cannot reveal, enabling targeted diagnosis of each model's bottleneck.

**Script-category analysis on Handwriting OCR.** To examine how error rates differ across script categories, we decompose CER by script category (kanji, hiragana, katakana, ASCII digits, and ASCII letters). For each instance, we obtain a character-level alignment between prediction and reference via minimum edit distance backtracing, then compute CER separately for each category.

Figure 7 presents per-category CER for each model, showing that CER varies substantially across script categories. ASCII digits achieve the lowest CER across models, consistent with their small and visually distinct character set. Kanji exhibits the highest error rate, which we attribute primarily to the large character inventory: models must disambiguate among thousands of classes, many with limited per-class training exposure. Since kanji accounts for 39.7% of all reference characters (Figure 4), its high CER is the dominant contributor to overall scores.

On the other hand, InternVL3.5 models exhibit elevated katakana CER, whereas Gemma3 shows high CER on both kanji and ASCII letters. Figure 8 illustrates these contrasts: InternVL3.5-14B confuses visually similar katakana pairs, while Gemma3-12B misrecognizes kanji with other unrelated kanji characters. These differences likely reflect variation in Japanese script coverage during pretraining. Among the weakest models, Gemma3-4B-IT and Phi-4-Multimodal produce CER values exceeding 1.0, indicating that their edit distances exceed the reference length. As with the Dense STVQA error taxonomy, stratifying evaluation by linguistically meaningful categories reveals distinct failure profiles that aggregate scoring would obscure.

**Comparison with OCR-specialized models.** To situate VLM performance on the recognition-dominant Handwriting OCR task, we compare against three OCR-specialized models: DeepSeek-OCR [42], PaddleOCR-VL [8], and olmOCR-2-7B [35], evaluated under the same conditions (Section 4.1). As Table 7 shows, the best OCR-specialized model (olmOCR-2-7B, 0.74) falls below the best general-purpose VLM (Qwen3-VL-8B, 0.79) at a comparable parameter scale. OCR-

Table 7: Handwriting OCR results for OCR-specialized models.

| Model | Params | Handwriting OCR (1-CER) |
|---|---|---|
| olmOCR-2-7B | 7B | **0.74** |
| PaddleOCR-VL | 0.9B | 0.61 |
| DeepSeek-OCR | 3B | 0.53 |

specialized models are often positioned as strong baselines for document OCR and document parsing. Still, they perform similarly or worse than general-purpose VLMs when recognizing handwritten text in real-world environments, where diverse writing media, writing instruments, and imaging conditions differ substantially from scanned documents. This result underscores that robust recognition of handwritten scene text remains an open challenge that cannot be addressed by OCR-specific training alone.

## 5 Conclusion

We introduced JaWildText, a diagnostic benchmark for evaluating VLMs on Japanese scene text understanding across three complementary tasks: Dense STVQA, Receipt KIE, and Handwriting OCR. Benchmarking 14 open-weight VLMs shows that the best model achieves only 0.64 on our unified score, confirming that robust Japanese text understanding in the wild remains far from solved. Stratifying errors by type and script category reveals that recognition remains the dominant bottleneck, with kanji posing a particular challenge. Closing the remaining gap will require targeted interventions at each stage, informed by the kind of fine-grained diagnosis that JaWildText provides. We hope that JaWildText will encourage diagnostic scene text evaluation in other typologically diverse languages.

## References

1. Abouelenin, A., et al.: Phi-4-Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv:2503.01743 (2025)
2. Baek, J., et al.: MangaVQA and MangaLMM: A Benchmark and Specialized Model for Multimodal Manga Understanding. arXiv:2505.20298 (2025)
3. Bai, S., et al.: Qwen3-VL Technical Report. arXiv:2511.21631 (2025)
4. Biten, A.F., et al.: Scene Text Visual Question Answering. In: IEEE/CVF International Conference on Computer Vision. pp. 4291–4301 (2019). https://doi.org/10.1109/ICCV.2019.00439
5. Chen, J., et al.: Benchmarking Chinese Text Recognition: Datasets, Baselines, and an Empirical Study. arXiv:2112.15093 (2022)
6. Clanuwat, T., et al.: Deep Learning for Classical Japanese Literature. arXiv:1812.01718 (2018)

7. Crosilla, G., et al.: Benchmarking Large Language Models for Handwritten Text Recognition. Journal of Documentation **81**(7), 334–360 (2025). https://doi.org/10.1108/JD-01-2025-0020
8. Cui, C., et al.: PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model. arXiv:2510.14528 (2025)
9. Fu, L., et al.: OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. arXiv:2501.00321 (2025)
10. Fujitake, M.: JaPOC: Japanese Post-OCR Correction Benchmark using Vouchers. arXiv:2409.19948 (2024)
11. Gemma Team: Gemma 3 Technical Report. arXiv:2503.19786 (2025)
12. Goto, H.: JPSC1400 – Japanese Scene Character Dataset. Dataset (Rev.20201218) (2020), https://www.imglab.org/db/
13. Huang, Z., et al.: ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. arXiv:2103.10213 (2021)
14. Hwang, T., et al.: KRETA: A Benchmark for Korean Reading and Reasoning in Text-Rich VQA Attuned to Diverse Visual Contexts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 33421–33432 (2025). https://doi.org/10.18653/v1/2025.emnlp-main.1696
15. Ishida, T., et al.: ICDAR 2019 Robust Reading Challenge on Omnidirectional Video. In: International Conference on Document Analysis and Recognition. pp. 1488–1493 (2019). https://doi.org/10.1109/ICDAR.2019.00240
16. Iwamura, M., et al.: Downtown Osaka Scene Text Dataset. In: European Conference on Computer Vision Workshops (ECCVW). Lecture Notes in Computer Science, vol. 9913, pp. 440–455 (2016). https://doi.org/10.1007/978-3-319-46604-0_32
17. Jaume, G., et al.: FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. In: International Conference on Document Analysis and Recognition Workshops (ICDARW). pp. 1–6 (2019). https://doi.org/10.1109/ICDARW.2019.10029
18. Landeghem, J.V., et al.: Document Understanding Dataset and Evaluation (DUDE). In: IEEE/CVF International Conference on Computer Vision. pp. 19528–19540 (2023). https://doi.org/10.1109/ICCV51070.2023.01789
19. Liu, Y., et al.: OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. Science China Information Sciences **67**(12), 220102 (2024). https://doi.org/10.1007/s11432-024-4235-6
20. Long, S., et al.: Scene Text Detection and Recognition: The Deep Learning Era. International Journal of Computer Vision **129**, 161–184 (2021). https://doi.org/10.1007/s11263-020-01369-0
21. Marti, U.V., Bunke, H.: The IAM-database: An English Sentence Database for Offline Handwriting Recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002). https://doi.org/10.1007/s100320200071
22. Mathew, M., et al.: DocVQA: A Dataset for VQA on Document Images. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2199–2208 (2021). https://doi.org/10.1109/WACV48630.2021.00225
23. Mathew, M., et al.: InfographicVQA. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022). https://doi.org/10.1109/WACV51458.2022.00264
24. Matsumoto, K., et al.: Collection and Analysis of On-line Handwritten Japanese Character Patterns. In: International Conference on Document Analysis and Recognition. pp. 496–500 (2001). https://doi.org/10.1109/ICDAR.2001.953839

25. Matsushita, T., et al.: A Database of On-Line Handwritten Mixed Objects Named Kondate. In: International Conference on Frontiers in Handwriting Recognition. pp. 369–374 (2014). https://doi.org/10.1109/ICFHR.2014.68
26. Nakagawa, M., et al.: Collection of on-line handwritten Japanese character pattern databases and their analyses. International Journal on Document Analysis and Recognition **7**(1), 69–81 (2004). https://doi.org/10.1007/s10032-004-0125-4
27. Nathan, S.: Japanese-Mobile-Receipt-OCR-1.3K: A Comprehensive Dataset Analysis and Fine-tuned Vision-Language Model for Structured Receipt Data Extraction. TechRxiv (preprint) (2025). https://doi.org/10.36227/techrxiv.175616889.90325672/v1
28. National Institute of Advanced Industrial Science and Technology (AIST): ETL Character Database. Online database (2014), collected 1973–1984; accessed 2025-12-18.
29. Nayef, N., et al.: ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification – RRC-MLT. In: International Conference on Document Analysis and Recognition. pp. 1454–1459 (2017). https://doi.org/10.1109/ICDAR.2017.237
30. Nayef, N., et al.: ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition – RRC-MLT-2019. In: International Conference on Document Analysis and Recognition. pp. 1582–1587 (2019). https://doi.org/10.1109/ICDAR.2019.00254
31. Onami, E., et al.: JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation. pp. 9503–9514 (2024)
32. Onohara, S., et al.: JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In: Proceedings of NAACL-HLT 2025. pp. 932–950 (2025). https://doi.org/10.18653/v1/2025.naacl-long.43
33. OpenAI: GPT-4 Technical Report. arXiv:2303.08774 (2023)
34. Park, S., et al.: CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In: NeurIPS 2019 Workshop on Document Intelligence (2019)
35. Poznanski, J., et al.: olmOCR 2: Unit Test Rewards for Document OCR. arXiv:2510.19817 (2025)
36. Sasagawa, K., et al.: Evaluating Multimodal Large Language Models on Vertically Written Japanese Text. arXiv:2511.15059 (2025)
37. SBIntuitions: Sarashina2.2-Vision. https://huggingface.co/sbintuitions/sarashina2.2-vision-3b (2025)
38. Singh, A., et al.: Towards VQA Models That Can Read. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019). https://doi.org/10.1109/CVPR.2019.00851
39. Tang, J., et al.: MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. arXiv:2405.11985 (2024)
40. Veit, A., et al.: COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. arXiv:1601.07140 (2016)
41. Wang, W., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv:2508.18265 (2025)
42. Wei, H., et al.: DeepSeek-OCR: Contexts Optical Compression. arXiv:2510.18234 (2025)
43. Xu, Y., et al.: XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 3214–3224 (2022). https://doi.org/10.18653/v1/2022.findings-acl.253

44. Yang, Z., et al.: CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy. arXiv:2412.02210 (2024)
45. Yao, C., et al.: Incidental Scene Text Understanding: Recent Progresses on ICDAR 2015 Robust Reading Competition Challenge 4. arXiv:1511.09207 (2015)