ThaiOCRBench: A Task-Diverse Benchmark for Vision-Language Understanding in Thai

Anonymous ACL submission

Abstract

We present ThaiOCRBench, the first comprehensive benchmark for evaluating vision-language models (VLMs) on Thai text-rich visual understanding tasks. Despite recent progress in multimodal modeling, existing benchmarks predominantly focus on high-resource languages, leaving Thai underrepresented, especially in tasks requiring document structure understanding. ThaiOCRBench addresses this gap by offering a diverse, human-annotated dataset comprising 2,808 samples across 13 task categories. We evaluate a wide range of state-of-the-art VLMs in a zero-shot setting, spanning both proprietary and open-source systems. Results show a significant performance gap, with proprietary models (e.g., Gemini 2.5 Pro) outperforming open-source counterparts. Notably, fine-grained text recognition and handwritten content extraction exhibit the steepest performance drops among open-source models. Through detailed error analysis, we identify key challenges such as language bias, structural mismatch, and hallucinated content. ThaiOCRBench provides a standardized framework for assessing VLMs in low-resource, script-complex settings, and provides actionable insights for improving Thai-language document understanding.

1 Introduction

Vision-Language Models (VLMs) have demonstrated strong performance across a variety of multimodal tasks, including image captioning, visual question answering (VQA), and visual grounding. These advancements are primarily driven by transformer-based architectures and large-scale pre-training on image–text pairs. However, despite these gains, VLMs continue to face significant challenges when processing text-heavy images, particularly documents characterized by complex layouts, dense text, and multilingual scripts (Hu et al., 2024; Zhang et al., 2025). These limitations are especially evident in low-resource languages like Thai, where both linguistic and structural characteristics are underrepresented in current training corpora.

Most existing VLMs are trained and evaluated on English-centric datasets that fail to capture the unique features of Thai, such as the absence of inter-word spacing, the presence of stacked diacritics, and the diversity of document formats. While some multilingual VLMs nominally support Thai at the tokenization and inference levels, their performance on Thai-specific tasks has not been systematically assessed. Furthermore, the lack of standardized, human-annotated benchmarks for Thai text-rich vision tasks hinders rigorous evaluation and slows progress toward developing robust, language-inclusive VLMs.

In contrast, numerous benchmarks have been developed for high-resource languages, especially English. Early efforts focused on scene text recognition (e.g., IIIT5K (Mishra et al., 2012), SVT (Wang et al., 2011)), followed by more complex datasets such as TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), and ChartQA (Masry et al., 2022). Additional benchmarks such as FUNSD (Jaume et al., 2019) and SROIE (Huang et al., 2019) target structured document understanding through Key information extraction. In the Thai context, existing datasets such as the NECTEC Thai OCR corpus, BEST2019 (NECTEC, 2020), and BurapaTH (Onuean et al., 2022) primarily support low-level tasks such as character or handwritten extraction, offering limited coverage of higher-level reasoning. Some small-scale efforts address scene text (Suwanwiwat et al., 2021), but comprehensive benchmarks for tasks such as layout parsing, relation extraction, or document-level VQA remain unavailable.

Recent evaluation frameworks OCRBench (Liu et al., 2024), OCRBench v2 (Fu et al., 2024), and CC-OCR (Yang et al., 2024) cover a broad set of tasks across document understanding and visual

reasoning. However, these benchmarks overwhelmingly focus on high-resource languages, with Thai either underrepresented or excluded. Although recent multilingual efforts such as MTVQA (Tang et al., 2024) and PM4Bench (Gao et al., 2025) include Thai, they are limited in task diversity and primarily address basic VQA.

To address this gap, we propose **ThaiOCRBench**, the first comprehensive benchmark designed to evaluate VLMs on Thai language, text-rich visual tasks. ThaiOCRBench contains 2,808 human-annotated samples spanning 13 task categories and diverse domains, including Chart parsing, Table parsing, Document parsing, Fine-grained text recognition, Full-page OCR, Handwritten content extraction, Text recognition, Key information extraction, Key information mapping, Document classification, Diagram VQA, Cognition VQA, and Infographics VQA.

This benchmark enables a focused investigation of the following research questions:

- **RQ1**: *How well do current VLMs generalize to Thai-language text-rich visual tasks?*
- **RQ2**: *What are the common failure modes of open-source VLMs on these tasks, and how do they vary across tasks and model scales?*

To explore these questions, we conduct two complementary studies. For **RQ1**, we perform a systematic zero-shot evaluation of both proprietary and open-source VLMs on ThaiOCRBench. For **RQ2**, we carry out a qualitative error analysis of open-source models to identify prevalent failure modes and characterize performance gaps.

Our findings indicate that proprietary models particularly Gemini 2.5 Pro (Comanici et al., 2025) consistently outperform open-source counterparts. Among open-source models, Qwen2.5-VL 72B (Bai et al., 2025) achieves the highest overall performance, though a notable gap remains. Detailed analysis reveals three dominant error patterns in open-source models: (1) Language Bias and Code-switching, (2) Structural Mismatch, and (3) Incorrect content.

Contributions. Our work makes the following key contributions:

- We introduce **ThaiOCRBench**, the first multi-task benchmark tailored for Thai-language vision-language understanding, with 2,808 human-annotated samples covering 13 task types. The dataset will be publicly released under the permissive CC BY-SA license to facilitate future research and reproducibility.
- We establish zero-shot baselines for state-of-the-art VLMs, spanning both proprietary and open-source systems, enabling standardized evaluation for Thai-language document tasks.
- We conduct an error analysis of open-source models, highlighting common limitations and offering insights for future improvements in Thai-specific VLM capabilities.

2 Related Work

2.1 Vision-Language Models with Thai Support

Most vision-language models (VLMs) have been developed and benchmarked primarily on high-resource languages, particularly English and Chinese. Recent advancements include both open-source models such as Gemma3 (Team et al., 2025a), Qwen2.5-VL, and LLaMA3.2 Vision (AI@Meta, 2024) and proprietary systems such as GPT-4o (OpenAI et al., 2024), Gemini 2.5 Pro, and Claude Sonet 4 (Anthropic, 2025). These models demonstrate strong performance across various document understanding tasks and generally support multiple languages at the tokenization and inference levels.

However, their evaluations are typically restricted to multilingual benchmarks such as MTVQA, which primarily emphasize high-level tasks such as visual question answering (VQA). Systematic assessments of these models on Thai-specific tasks, especially those requiring fine-grained reasoning over structured and complex content such as tables, forms, and charts remain limited. Consequently, the extent to which current VLMs generalize to Thai-language, text-rich scenarios is still largely unexplored. This work addresses this gap by introducing a benchmark specifically designed to enable systematic evaluation of VLMs across a wide range of Thai-language vision tasks.

2.2 Benchmarks for Thai Text-Rich Vision Tasks

Benchmark resources for Thai-language vision tasks remain limited in both task diversity and complexity. Existing datasets focus predominantly on low-level recognition. For instance, the NECTEC

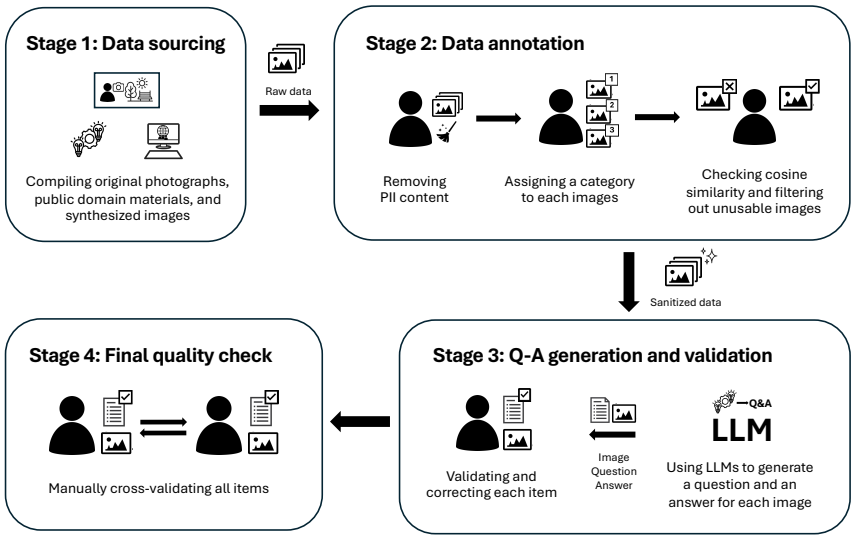


Figure 1: Overview of the ThaiOCRBench data collection and annotation pipeline.

Thai OCR corpus provides printed Thai text images for character-level optical character recognition (OCR), while BEST2019 offers annotated handwritten lines for offline handwritten extraction. Similarly, the Burapha-TH dataset targets isolated character and syllable recognition.

While these datasets are valuable for developing foundational OCR systems, they lack the structural and semantic annotations necessary to support higher-level tasks such as element parsing, relation extraction, or VQA. Moreover, no existing benchmark integrates a diverse set of Thai-language vision tasks within a unified framework. This limits comprehensive evaluation of models in realistic, document-centric scenarios.

3 ThaiOCRBench: Dataset Construction

The construction of ThaiOCRBench was a multi-stage process guided by clear design principles to ensure its cultural relevance, diversity, and overall quality. This section details these principles, the task definitions, the data sourcing and annotation pipeline, and the final dataset statistics.

3.1 Design Principles

The construction of ThaiOCRBench was guided by two core design principles: cultural specificity and data diversity.

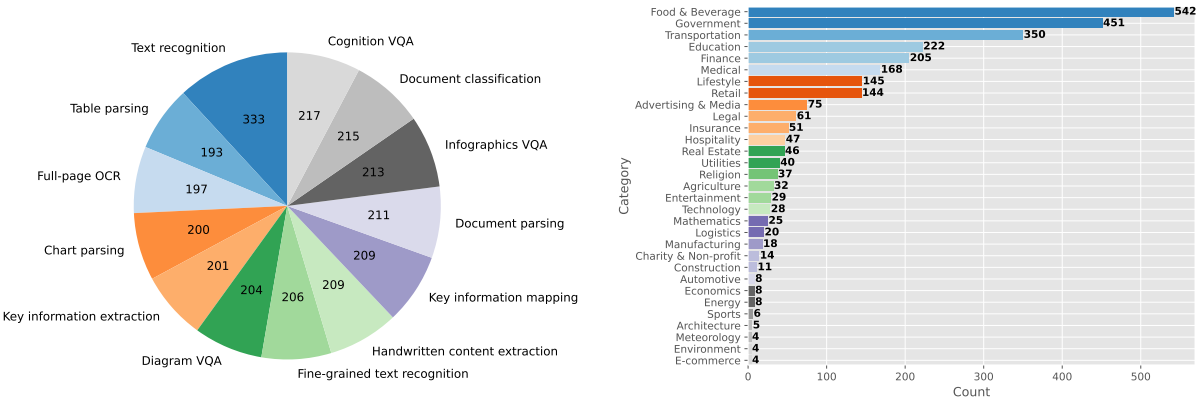
Cultural specificity emphasizes the inclusion of content that reflects linguistic, visual, and contextual elements unique to Thai cultures. This ensures that the benchmark evaluates model performance in authentically Thai scenarios rather than relying on generalized or translated content. Examples include visual elements requiring local cultural knowledge, such as Bangkok's color-coded public transportation signage, and culturally specific symbols, such as prohibition signs against durians in public areas. Additionally, the dataset incorporates linguistically complex content, such as Pali-Sanskrit chants written in Thai script, which feature rare characters and vocabulary that are typically absent from standard web-based corpora.

Data diversity ensures broad representation across domains, text modalities, and visual styles. The benchmark includes a variety of document types (e.g., government reports, restaurant menus, medical forms), text formats (e.g., machine-printed, handwritten, poetic verse), and typographic styles. This also includes both traditional Thai "headed" scripts and modern "headless" variants. The latter introduces significant recognition challenges due to their visual similarity to Latin characters. For example, the headless form of the Thai letter "Nor Nu" closely resembles the lowercase Latin letter "u", creating substantial ambiguity for OCR systems.

Representative examples illustrating these principles are provided in Appendix A.3, highlighting the benchmark's emphasis on linguistic complexity and real-world variability.

3.2 Task Categories Definition

ThaiOCRBench is a multi-task benchmark comprising 13 task types designed to evaluate the capabilities of vision-language models (VLMs) in processing Thai text-rich visual content. While OCRBench v2 provides a broader task set aggregated



(a) Distribution of the 13 task types in ThaiOCRBench.

(b) Domain-specific category distribution.

Figure 2: Distributional statistics of ThaiOCRBench, illustrating the coverage across task types and domain-specific categories.

from multiple datasets, we adopt it as a reference due to its comprehensive coverage and structured evaluation methodology.

In contrast to OCRBench v2, ThaiOCRBench emphasizes a focused set of linguistically and structurally challenging tasks tailored specifically to the Thai language. All images in the dataset are newly collected and manually annotated to reflect authentic layouts, localized formats, and language-specific phenomena. The task categories are defined in Appendix A.1

3.3 Data Sourcing and Annotation

As illustrated in Figure 1, the dataset was constructed through a four-stage pipeline designed to ensure data diversity, ethical compliance, and annotation quality.

Stage 1: Data Sourcing. Images were collected from a variety of sources, including original photographs taken in public spaces, publicly available materials, and licensed commercial datasets. For sensitive document types such as identification cards and legal certificates, synthetic documents were programmatically generated to avoid privacy concerns. All images underwent a sanitization process in which human annotators manually removed or obscured personally identifiable information (PII), such as faces, names, and identification numbers.

Stage 2: Data Annotation. Human annotators categorized each image based on content type and assigned relevant metadata, including source information, licensing details, and descriptive tags. To ensure data uniqueness and reduce redundancy, pairwise cosine similarity was computed across image embeddings. Instances of high similarity such as images with near-identical angles, fonts, or layouts in the same task category were flagged and reviewed, and duplicates or near-duplicates were removed accordingly. Detailed annotation guidelines are provided in Appendix A.2.

Stage 3: Question–Answer Generation and Validation. We employed multiple large language models (LLMs), including GPT-4o, Gemini 2.5 Pro, and Azure AI Services (Microsoft, 2025), to generate initial question–answer (QA) pairs for each image. Human annotators then reviewed these outputs, selecting or refining the most suitable pairs based on task-specific guidelines. As many generated QA pairs were found to be inaccurate or misaligned with the visual content, substantial manual revision or rewriting was conducted to ensure correctness and task relevance.

Stage 4: Final Quality Control. A separate team of annotators conducted a final review of all dataset entries. Each item comprising the image, associated question, and answer was assessed for coherence, accuracy, and alignment with the intended task definitions. Only items that met all quality standards were retained in the final benchmark.

3.4 Dataset Statistics

ThaiOCRBench consists of 2,808 images with human-annotated question-answer pairs. Figure 2a illustrates the distribution across the 13 task types, while Figure 2b shows the coverage of domain-specific categories. Token length statistics for questions and answers are provided in Appendix 5.

4 Experimental Design

We design and conduct two complementary studies using the ThaiOCRBench benchmark. To address **RQ1**, we perform zero-shot evaluations of state-of-the-art vision-language models (VLMs), encompassing both proprietary and open-source systems, to assess their effectiveness on Thai language, text-rich visual tasks (Section 5.1). To address **RQ2**, we conduct a qualitative error analysis of open-source models, categorizing failure cases into three primary types. This analysis aims to identify key limitations and inform strategies for narrowing the performance gap relative to proprietary models (Section 5.2).

4.1 Evaluation Models

We evaluate a range of state-of-the-art vision-language models (VLMs), encompassing both proprietary and open-source systems, to assess their zero-shot performance on ThaiOCRBench. The selected models vary in architectural design, training data, and parameter scale, allowing for a comprehensive comparison across model families.

All evaluations are conducted in a zero-shot setting using the `vLLM` inference engine [1] with greedy decoding. The models assessed include proprietary systems such as Gemini 2.5 Pro, Claude Sonnet 4, and GPT-4o , as well as open-source models including Qwen2.5-VL, Gemma3, LLaMA3.2 Vision, InternVL 3 (Zhu et al., 2025), Aya-Vision (Dash et al., 2025), Kimi-VL (Team et al., 2025b), SmolVLM (Marafioti et al., 2025), Pixtral (Agrawal et al., 2024), Phi-3 (Abdin et al., 2024a), Phi-4 (Abdin et al., 2024b), Skywork-R1V (Peng et al., 2025) and MiniCPM-o 2.6 (Yao et al., 2024).

4.2 Evaluation Metrics

We adopt evaluation metrics from OCRBench v2.

Structural Understanding Tasks. For tasks involving the reconstruction of document layout and hierarchical content such as *Table parsing*, *Chart parsing*, and *Document parsing*, we employ the Tree Edit Distance (TED) metric (Zhong et al., 2020), which quantifies structural similarity between predicted and reference outputs. TED is particularly suited to evaluating nested or hierarchical formats where layout consistency is critical.

Text Generation and Recognition Tasks. For tasks requiring the transcription or generation of text such as *Fine-grained text recognition*, *Full-page OCR*, and *Handwritten content extraction*, we report multiple complementary metrics to assess both character-level accuracy and linguistic fidelity. These include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), F1-score, and Normalized Levenshtein Similarity (NLS) (Biten et al., 2019). We average these into a single composite metric, referred to as BMFL.

Structured Prediction Tasks. For *Key information extraction* and *Key information mapping*, we use the F1-score to evaluate precision and recall in entity-level prediction. This metric is appropriate for scenarios where exact field alignment is required and partial matches are penalized.

Textual Understanding and Question Answering Tasks. For tasks involving semantic understanding and short-form textual generation such as *Text recognition*, *Document classification*, *Diagram VQA*, *Cognition VQA*, and *Infographics VQA*, we adopt the Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019). ANLS measures similarity by normalizing the edit distance between predicted and reference responses, allowing for partial credit when predictions are close but not exact.

Although each metric is calculated differently, all are designed such that higher scores indicate better performance. To facilitate model comparison across diverse tasks, we also report the average score as an overall performance indicator.

5 Experiment Results

5.1 Zero-Shot Performance Evaluation

To address **RQ1**, we evaluate state-of-the-art vision-language models (VLMs) on ThaiOCRBench under a zero-shot greedy decoding setting.

Results. Table 1 summarizes the performance of proprietary and open-source models. Proprietary models consistently outperform open-source counterparts across most tasks. Gemini 2.5 Pro achieves the highest overall average score (0.777), ranking first in 11 out of 13 tasks. It performs particularly well in Key information mapping (0.863), Full-page OCR (0.897), and Text recognition (0.910).

GPT-4o also demonstrates strong performance, leading in Document classification (0.973) and Cognition VQA (0.796), with an overall score of 0.645. Claude Sonnet 4 follows with an average score of 0.579.

Among open-source models, Qwen2.5-VL 72B achieves the highest average (0.615), closing the

[1] https://github.com/vllm-project/vllm

Model	TED			BMFL				F1		ANLS				
	Chart parsing	Table parsing	Doc. parsing	Fine-grained Rec.	Full-page OCR	Handwritten	Text recognition	Info. extraction	Info. mapping	Doc. classification	Diagram VQA	Cognition VQA	Infographics VQA	Average score
Proprietary model														
Gemini 2.5 Pro	0.812	**0.686**	**0.587**	**0.499**	**0.897**	**0.714**	**0.910**	**0.658**	**0.863**	0.943	**0.766**	**0.872**	**0.898**	**0.777**
Claude Sonnet 4	**0.817**	0.650	0.543	0.214	0.661	0.301	0.686	0.452	0.675	0.879	0.379	0.657	0.613	0.579
GPT-4o	0.766	0.571	0.515	0.254	0.610	0.489	0.778	0.546	0.734	**0.973**	0.562	0.796	0.791	0.645
Open-source model														
Gemma3 27B	0.783	0.519	0.350	0.144	0.608	0.280	0.561	0.389	0.574	0.831	0.309	0.514	0.552	0.493
Gemma3 12B	0.704	0.395	0.358	0.084	0.504	0.225	0.433	0.300	0.558	0.770	0.270	0.428	0.471	0.423
Gemma3 4B	0.635	0.322	0.355	0.089	0.363	0.143	0.233	0.225	0.493	0.683	0.129	0.349	0.343	0.336
Qwen2.5-VL 72B	0.801	0.549	0.454	0.147	0.720	0.393	0.749	0.497	0.719	0.914	0.519	0.746	0.782	0.615
Qwen2.5-VL 32B	0.765	0.483	0.334	0.139	0.553	0.280	0.635	0.394	0.708	0.860	0.409	0.650	0.681	0.530
Qwen2.5-VL 7B	0.712	0.509	0.308	0.218	0.631	0.314	0.597	0.354	0.623	0.862	0.416	0.702	0.763	0.539
Qwen2.5-VL 3B	0.650	0.431	0.338	0.130	0.430	0.210	0.475	0.284	0.481	0.821	0.308	0.532	0.550	0.434
InternVL3 78B	0.768	0.440	0.434	0.073	0.167	0.158	0.069	0.300	0.572	0.759	0.217	0.306	0.367	0.356
InternVL3 14B	0.760	0.399	0.405	0.059	0.184	0.140	0.038	0.334	0.534	0.712	0.170	0.321	0.352	0.339
InternVL3 8B	0.731	0.423	0.298	0.052	0.157	0.127	0.033	0.252	0.480	0.698	0.154	0.269	0.305	0.306
Aya-Vision 8B	0.567	0.229	0.322	0.027	0.080	0.075	0.005	0.056	0.187	0.466	0.058	0.115	0.123	0.178
Kimi-VL-A3B-Instruct	0.404	0.373	0.327	0.026	0.105	0.091	0.013	0.176	0.159	0.551	0.113	0.189	0.261	0.214
SmolVLM2 2.2B	0.015	0.042	0.134	0.030	0.049	0.048	0.000	0.003	0.000	0.135	0.010	0.017	0.030	0.039
Pixtral 12B	0.637	0.380	0.334	0.039	0.113	0.091	0.018	0.154	0.393	0.671	0.094	0.191	0.270	0.260
Phi-3 vision 4B	0.475	0.186	0.202	0.034	0.039	0.057	0.006	0.119	0.209	0.269	0.039	0.142	0.148	0.148
Skywork-R1V-38B	0.756	0.418	0.385	0.074	0.181	0.128	0.055	0.344	0.558	0.765	0.136	0.256	0.304	0.335
Phi-4 multimodal 5B	0.591	0.212	0.237	0.028	0.050	0.063	0.003	0.065	0.237	0.316	0.038	0.129	0.131	0.162
Llama 3.2-Vision 11B	0.222	0.326	0.252	0.051	0.207	0.145	0.237	0.097	0.485	0.769	0.163	0.368	0.424	0.288
MiniCPM-o 2.6 8B	0.497	0.181	0.170	0.046	0.082	0.075	0.008	0.050	0.256	0.628	0.106	0.206	0.241	0.196

Table 1: Performance comparison of proprietary and open-source models on ThaiOCRBench. Tasks are grouped by evaluation metric: **TED** (Chart, Table, Doc Parsing), **BMFL** (Generation and Recognition tasks), **F1** (Information extraction tasks), and **ANLS** (Understanding/VQA tasks). Bold values denote the best proprietary model; underlined values denote the best open-source model.

gap with proprietary systems in tasks such as Document classification (0.914), Full-page OCR (0.720), and Cognition VQA (0.615). Other Qwen variants (32B and 7B) also perform competitively across multiple tasks.

Discussion. Our results reveal clear disparities in model performance across different ThaiOCRBench tasks, highlighting the varied demands of each subtask and the influence of evaluation metrics. For example, as shown in Table 1, models such as InternVL3 78B perform well on Chart parsing (0.768) but poorly on Text recognition (0.069). This discrepancy rises from differences in metric sensitivity. Chart parsing is evaluated using structure-aware metrics such as Tree Edit Distance (TED), which are less sensitive to minor text errors as long as the overall structure remains accurate. In contrast, Text tecognition is assessed using normalized Levenshtein distance, a stricter metric that penalizes even small character-level mistakes, an especially challenging aspect when dealing with complex scripts such as Thai.

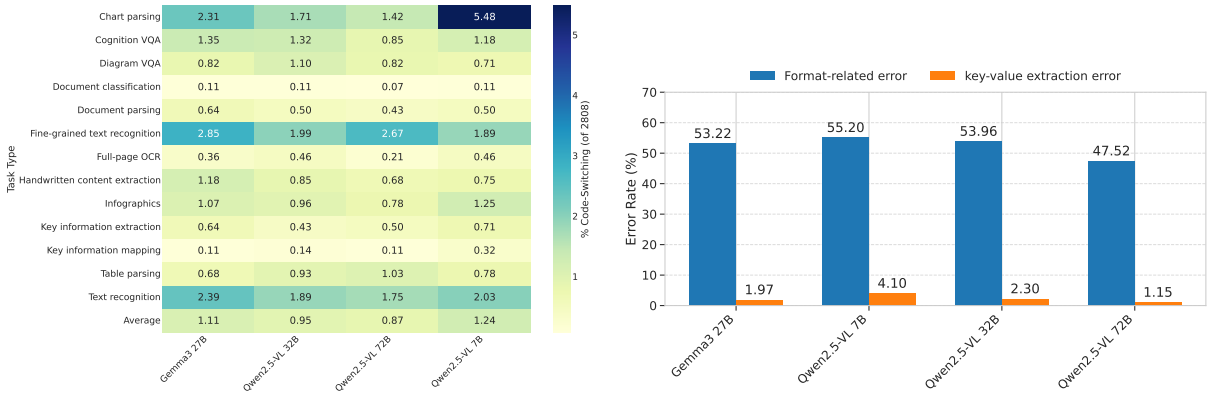
These results also reflect divergent model capabilities. InternVL3, for instance, may effectively capture document structure, benefiting layout-intensive tasks, but lacks robust Thai tokenization or pretraining, limiting its performance in script-level decoding.

We also observe that Document classification is comparatively less challenging. This task is framed as a constrained multiple-choice problem with seven predefined categories, reducing the need for complex reasoning. As a result, most models including smaller ones achieve strong performance. In contrast, Fine-grained text recognition remains the most difficult task, aligning with trends observed in English benchmarks such as OCRBench v2. It requires accurate localization and transcription of small text elements embedded in complex layouts capabilities that remain challenging even for large-scale models.

5.2 Qualitative Error Analysis

To address **RQ2**, we conduct a qualitative error analysis by categorizing common failure modes observed in model predictions. We focus on the top four open-source models based on their average performance on ThaiOCRBench (as reported in Table 1). Errors are classified into three primary categories: (1) *Language Bias and Code-Switching*, (2) *Structural Mismatch*, and (3) *Incorrect Content*.

Language Bias and Code-Switching. As



(a) Language Bias and Code-Switching behavior across models, measured by the proportion of data samples exhibiting language switching, normalized by the total number of samples.

(b) Structural Mismatch across models, measured by the proportion of data points exhibiting structural inconsistencies, normalized by the total number of error instances.

Figure 3: Qualitative error breakdown across Top-4 models.

	Substitutions ($\times 10^4$)				Deletions ($\times 10^4$)				Insertions ($\times 10^4$)				Correct ($\times 10^4$)			
Task	G27	Q32	Q72	Q7	G27	Q32	Q72	Q7	G27	Q32	Q72	Q7	G27	Q32	Q72	Q7
Fine-grained text	0.23	0.27	0.18	0.19	0.16	0.11	0.24	0.17	0.24	1.76	1.12	2.09	0.14	0.16	0.12	0.19
Infographics VQA	0.19	0.15	0.07	0.10	0.30	0.15	0.25	0.25	0.31	0.44	0.08	0.12	0.53	0.73	0.70	0.67
Chart parsing	4.89	3.33	3.34	3.72	3.49	3.42	2.69	3.87	5.44	7.00	6.92	16.96	9.10	10.74	11.45	9.89
Cognition VQA	0.12	0.09	0.06	0.09	0.21	0.17	0.19	0.17	0.29	0.13	0.05	1.62	0.45	0.52	0.54	0.53
Diagram VQA	0.15	0.15	0.10	0.12	0.22	0.16	0.23	0.22	0.20	5.14	0.07	0.22	0.23	0.30	0.27	0.26
Document parsing	7.64	11.74	6.02	9.02	10.05	5.09	5.33	8.47	**56.76**	**41.24**	**98.52**	**38.98**	18.61	19.46	24.94	18.80
Full-page OCR	4.69	6.25	2.97	3.44	6.07	2.47	2.25	4.51	5.55	59.66	8.85	21.65	17.87	19.91	23.40	20.67
Table parsing	6.46	7.84	5.46	6.05	**18.38**	**14.64**	**15.97**	**15.37**	6.32	30.04	12.71	18.07	28.17	30.54	31.60	31.60
Text recognition	0.94	0.79	0.45	0.51	0.88	0.54	0.55	1.66	2.21	7.91	2.88	7.51	4.73	5.21	5.55	4.37
Handwritten	0.87	0.97	0.72	0.73	0.48	0.25	0.49	0.63	19.53	12.39	3.18	10.20	1.23	1.36	1.37	1.22
Key info. extraction	0.49	0.56	0.34	0.69	1.22	0.68	0.73	1.12	1.41	6.25	7.53	6.18	6.33	6.79	6.96	6.22
Key info. mapping	0.56	0.51	0.53	0.55	0.91	0.89	0.91	0.89	0.37	0.48	0.30	1.98	4.02	4.11	4.06	4.06
Doc. classification	0.03	0.02	0.01	0.02	0.01	0.00	0.00	0.00	0.03	0.04	0.02	0.06	0.24	0.25	0.26	0.25
Total	27.26	32.67	20.59	25.78	61.69	28.17	29.33	45.90	98.66	172.47	144.19	146.73	111.85	120.88	136.42	118.84

Table 2: Incorrect Content analysis across models. Character-level errors are reported based on the Character Error Rate (CER), including substitutions, deletions, insertions, and correct tokens, aggregated across tasks for the top-4 models. All values are scaled by 10^4 for readability. Model abbreviations: G27 = *Gemma3 27B*, Q32 = *Qwen2.5-VL 32B*, Q72 = *Qwen2.5-VL 72B*, Q7 = *Qwen2.5-VL 7B*.

shown in Figure 3a, we analyze discrepancies between model predictions and ground-truth references using few-shot prompting with GPT-4o. This evaluation reveals two prevalent categories of linguistic errors: (i) *language bias*, wherein models systematically default to non-Thai outputs despite receiving Thai inputs, and (ii) *code-switching*, characterized by the inappropriate intermixing of Thai and non-Thai language elements within a single prediction. These phenomena underscore persistent challenges related to multilingual generalization and script fidelity in current model architectures.

Structural Mismatch. Structural errors fall into two categories, as illustrated in Figure 3b:

- *Key–Value Extraction Errors*, found in tasks such as chart parsing, key information extraction, and key information mapping, where models fail to align fields with corresponding keys detected via rule-based validation.
- *Format-Related Errors*, common in document and table parsing, where predicted structural formats deviate from reference outputs (e.g., tag mismatches or missing components).

Incorrect Content. To assess transcription fidelity, we calculate the Character Error Rate (CER), decomposed into standard edit operations: substitutions, deletions, insertions, and correct matches. This token-level analysis enables fine-grained assessment of recognition quality in tasks involving text transcription, particularly Full-page OCR and Fine-grained text recognition. The breakdown of these error components is provided in Table 2.

Results. As illustrated in Figure 3a, we observe a consistent trend in which larger models exhibit lower code-switching rates, indicating improved linguistic stability. In particular, Qwen2.5-VL 72B achieves the lowest average code-switching rate at 0.87%, while Qwen2.5-VL 7B records the highest

at 1.24%, with a task-specific peak of 5.48% in Chart parsing. These findings suggest that code-switching errors are more prevalent in tasks involving structured reasoning and complex visual layouts, such as tables and diagrams. Such content often contains multilingual or ambiguously formatted text, which appears to challenge smaller models with weaker cross-modal instruction following and limited multilingual grounding.

Figure 3b further highlights that format-related errors remain a major source of failure across all models, with error rates ranging from 47.52% (Qwen2.5-VL 72B) to 55.20% (Qwen2.5-VL 7B). These results underscore the persistent difficulty that VLMs face in layout-intensive tasks, particularly when structural fidelity is essential. In contrast, key–value extraction errors show greater sensitivity to the model scale, with error rates declining from 4.10% in Qwen2.5-VL 7B to 1.15% in Qwen2.5-VL 72B. This suggests that larger models are better able to capture semantic relationships and maintain accurate alignment between visual cues and entity representations.

As shown in Table 2, Gemma3 27B (G27) yields the lowest overall deletion count (28.17), with particularly strong performance in structurally demanding tasks such as Table parsing (18.38) and Document parsing (5.09). These results suggest that G27 exhibits more stable and conservative generation behavior, minimizing omission-related errors in complex text-structured scenarios.

In contrast, Qwen2.5-VL-32B (Q32) produces the highest number of insertions (172.47), driven largely by errors in Table parsing (30.04), Document parsing (41.24), and Full-page OCR (59.66). Other Qwen variants, such as Qwen2.5-VL 72B (Q72) and Qwen2.5-VL 7B (Q7), follow similar patterns, though with reduced insertion counts. In terms of correctly generated characters, Qwen2.5-VL 72B achieves the highest count (136.42), followed by Q32 (120.88), reflecting the models' ability to produce a large amount of valid content despite higher insertion rates.

These trends reflect a core trade-off: Gemma favors conservative output, minimizing hallucinations, while Qwen prioritizes recall, risking more insertions.

Discussion. Our experimental results reveal clear trade-offs between model size, generation behavior, and task structure in multimodal document understanding. Larger models, such as Qwen2.5-VL 72B and Gemma3 27B, consistently outperform smaller counterparts in both language alignment and structural accuracy. While Qwen models tend to favor recall, often producing more insertions, Gemma exhibits more conservative generation with fewer deletions, particularly in high-complexity tasks such as document parsing.

These patterns suggest that model scaling alone does not uniformly reduce all error types; rather, model-family-specific tendencies (e.g., overgeneration vs. omission) influence downstream performance. Moreover, code-switching behavior appears strongly tied to parameter count, with larger models demonstrating improved language consistency. Overall, these findings emphasize the importance of balancing precision and recall depending on task demands, and highlight the need for further refinement in hallucination control for generative vision-language models.

6 Conclusion

In this work, we present ThaiOCRBench, the first comprehensive multi-task benchmark for Thai-language vision-language understanding. The benchmark consists of 2,808 human-annotated samples spanning 13 diverse tasks, designed to evaluate VLM performance on text-rich visual content in a low-resource language context.

Our zero-shot evaluation of state-of-the-art models demonstrates a clear performance disparity between proprietary and open-source systems. Gemini 2.5 Pro achieves the highest overall performance, consistently leading across most tasks. Among open-source models, Qwen2.5-VL 72B emerges as the strongest performer, though a notable performance gap remains.

Through qualitative error analysis, we identify three prominent failure modes in open-source models: language bias and code-switching, structural mismatch, and hallucinated or incorrect content. These insights underscore the need for further research and model adaptation to better support Thai-language document understanding.

ThaiOCRBench provides a standardized, task-diverse evaluation framework and establishes strong baselines for future research. We hope this benchmark will catalyze progress in developing more robust, inclusive, and linguistically aware vision-language models for Thai.

Limitations

While ThaiOCRBench represents a significant step toward comprehensive evaluation of Thai vision-language tasks, several limitations remain. Although the dataset was explicitly designed to capture cultural specificity and long-tail document characteristics, the scale of 2,808 annotated samples may still fall short of exhaustively representing the full diversity of real-world document types encountered in practice. Expanding coverage across additional document genres and regional variations remains an important direction for future work.

Second, our evaluation relies primarily on conventional metrics such as BLEU, ANLS, and Tree Edit Distance. While these provide standardized measures, they may not fully reflect semantic or contextual correctness. Recent approaches employing large language models as evaluators ("LLM-as-Judges") offer a promising alternative for more nuanced, task-aware assessments, particularly in generative or open-ended settings.

Third, our experiments are conducted exclusively under a zero-shot setting to assess the out-of-the-box generalization capabilities of current VLMs. However, this does not account for performance gains achievable through fine-tuning or instruction tuning, which may significantly alter outcomes, especially for models adapted to Thai-specific data.

Lastly, ThaiOCRBench is currently monolingual, focusing solely on Thai-language content. Extending the benchmark to include multilingual and code-switched documents would enable more comprehensive evaluation of VLMs in linguistically diverse and real-world cross-lingual scenarios.

Despite these limitations, ThaiOCRBench provides a robust foundation for standardized and culturally informed evaluation of vision-language models in low-resource language settings.

Ethical Statement

All annotation work was conducted by part-time linguists based in Thailand, hired and compensated in accordance with local labor standards. The dataset will be publicly released under the Creative Commons Attribution-ShareAlike (CC BY-SA) license, allowing for both academic and commercial use with appropriate credit and license compatibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024b. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2025. Claude sonnet 4. `https://claude.ai`. Large Language Model.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *Preprint*, arXiv:1905.13648.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3289 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context,

and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Junyuan Gao, Jiahe Song, Jiang Wu, Runchuan Zhu, Guanlin Shen, Shasha Wang, Xingjian Wei, Haote Yang, Songyang Zhang, Weijia Li, Bin Wang, Dahua Lin, Lijun Wu, and Conghui He. 2025. Pm4bench: A parallel multilingual multi-modal multi-task benchmark for large vision language model. *Preprint*, arXiv:2503.18484.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, Miami, Florida, USA. Association for Computational Linguistics.

Zhaoyi Huang, Lianwen Jin, Yongpan Ding, Shuye Mao, Zenghui Zhou, Cheng Liu, Zhi Cui, Yuliang Wei, Hong Wang, and Xinyu Li. 2019. Icdar 2019 competition on scanned receipt ocr and information extraction. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDAR-W)*, pages 1–6. IEEE.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Microsoft. 2025. Azure ai document intelligence. `https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/`. Accessed July 2025.

Anand Mishra, Karteek Alahari, and C. V. Jawahar. 2012. Scene text recognition using higher order language priors. In *BMVC*.

NECTEC. 2020. Best-handwritten (best2019) thai handwriting recognition dataset. NECTEC AI For Thai Corpus Repository. Accessed via AI For Thai platform.

Athita Onuean, Uraiwan Buatoom, Thatsanee Charoenporn, Taehong Kim, and Hanmin Jung. 2022. Burapha-th: A multi-purpose character, digit, and syllable handwriting dataset. *Applied Sciences*, 12(8).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. 2025. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *Preprint*, arXiv:2504.05599.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Hemmaphan Suwanwiwat, Abhijit Das, Muhammad Saqib, Umapada Pal, Hemmaphan Suwanwiwat, A Das, and M Saqib. 2021. Benchmarked multi-script thai scene text dataset and its multi-class detection solution. *Multimedia Tools and Applications*, 80.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 73 others. 2025b. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. 2025. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *Preprint*, arXiv:2410.21169.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. *Preprint*, arXiv:1911.10683.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Appendix

A.1 Task Categories Definition

Table parsing. Extract tabular content from images and convert it into structured formats (e.g., Markdown or HTML), preserving both semantic relationships and layout fidelity.

Chart parsing. Interpret visual charts (e.g., bar, line, pie) and generate structured JSON representations using predefined entity keys that reflect underlying data semantics.

Document parsing. Parse full-page documents comprising heterogeneous content. Plain text is transcribed in Markdown, tables in HTML, formulas in LaTeX, and visual elements (e.g., figures and charts) are annotated using descriptive tags, enabling holistic structural understanding.

Full-page OCR. Transcribe all textual content present in a document image, without task-specific constraints.

Fine-grained text recognition. Extract targeted textual segments from specified regions within an image, emphasizing localized reading and semantic precision.

Text recognition. Transcribe general textual content from images, serving as a core OCR evaluation task.

Document classification. Assign a given document image to one of seven predefined categories based on both visual layout and textual content.

Diagram VQA. Answer questions grounded in diagrammatic visuals, assessing the model's ability to jointly reason over spatial and textual components.

Cognition VQA. Answer image-grounded questions where the correct answer is explicitly embedded within the image, testing reading comprehension and factual grounding.

Infographics VQA. This task focuses on more visually complex, infographic-style inputs. Unlike Cognition VQA, it often involves dense, multimodal layouts that require parsing of both structured and unstructured visual information.

Key information extraction. Extract values corresponding to provided entity keys (e.g., "Name," "Date," "Amount") from densely populated documents and output them in a structured JSON format.

Key information mapping. Given a set of entity keys and candidate values, match each key to its corresponding value and group them into semantically meaningful clusters.

Handwritten content extraction. Extract and transcribe handwritten Thai text, addressing challenges associated with handwriting variability in real-world documents.

A.2 Annotation Guideline

This appendix outlines the step-by-step procedures and criteria followed by human annotators throughout the dataset creation process. These guidelines ensured consistency, quality, and compliance with ethical standards.

Stage 1: Data Sourcing

Objective: Collect images and ensure that they are legally and ethically usable.

- **Compiling Images:** Annotators collected images either through self-captured photographs or from publicly available sources, ensuring that each image was accompanied by appropriate documentation of its license type and usage rights.
- **Synthetic Document Generation (for PII-sensitive categories):** For sensitive categories that may involve personally identifiable information (PII) such as identification cards or certificates, images were generated using predefined templates and controlled data scripts. Annotators then verified that they did not contain real names, photographs, or identification numbers.

Stage 2: Data Annotation

Objective: Accurately categorize images and remove redundant data.

- **PII Sanitization:** Annotators manually reviewed each image and obscured or removed any visible faces, license plates, personal names, ID numbers or any PII information, using masking or blurring tools. To ensure accuracy and compliance, a second annotator independently reviewed the sanitization of each image.
- **Category and Metadata Tagging:** Annotators manually assigned the category to each image based on the nature of its content. In addition, annotators recorded the image source, its license type (public, licensed, synthetic, or self-taken) and optional domain tags.
- **Similarity Review:** Cosine similarity was computed between image embeddings to identify redundant or highly similar content. If the similarity score between two or more images exceeded a threshold of 0.95, annotators visually inspected the images. In cases where the images were found to be duplicates or near-duplicates, only the most representative sample was retained. Exceptions were made for essential images such as ID cards and official certificates, given that their high cosine similarity was necessary and due to their template-based nature.

Stage 3: Question–Answer Generation and Validation

Objective: Produce question-answer pairs aligned with the image content and task objectives.

- **LLM Output Triage:** Annotators manually reviewed two or three question–answer (QA) pairs generated by different large language models for each image. They then selected the pair that was most relevant, prioritizing clarity, accuracy, and alignment with the intended task objectives.
- **Manual Refinement:** In cases where all outputs were highly unclear, incorrect, or hallucinated, annotators manually edited the QA pairs. They were also instructed to visually inspect the image to confirm that the information referenced in the QA pair is actually present and correctly interpreted.

Stage 4: Final Quality Control

Objective: Cross-validate the dataset's readiness for public release and research use.

- **Coherence and Compliance Check:** A distinct set of annotators verified the logical alignment between the image, the question, and the answer. They also ensured that the image is sanitized, the question is meaningful, and the corresponding answer is directly grounded in the visual content of the image.

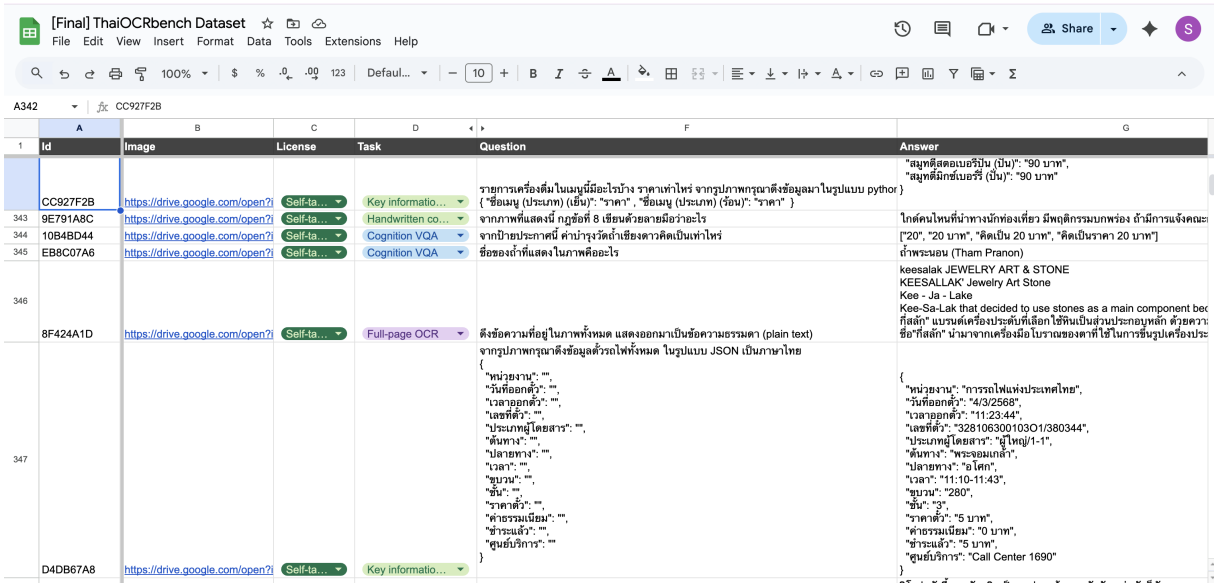


Figure 4: A screenshot of the annotation platform using Google Sheets

A.3 Dataset Samples

To illustrate the scope and diversity of our benchmark, we present representative samples for each task included in the dataset. Figure 6 showcases examples from the Infographics VQA and Key information mapping tasks. Figure 7 and Figure 8 present samples of Table parsing and Chart parsing, respectively. Figure 9 includes examples of Document parsing and Full-page OCR. Figure 10 highlights Fine-grained Text recognition and Text recognition tasks. Figure 11 illustrates Document classification and Diagram VQA. Figure 12 depicts a sample from the Key information extraction task. Lastly, Figure 13 presents examples of Handwritten content extraction and Cognition VQA.

These figures collectively demonstrate the wide range of document understanding tasks covered by our dataset, emphasizing its richness and complexity across both layout and semantic dimensions.

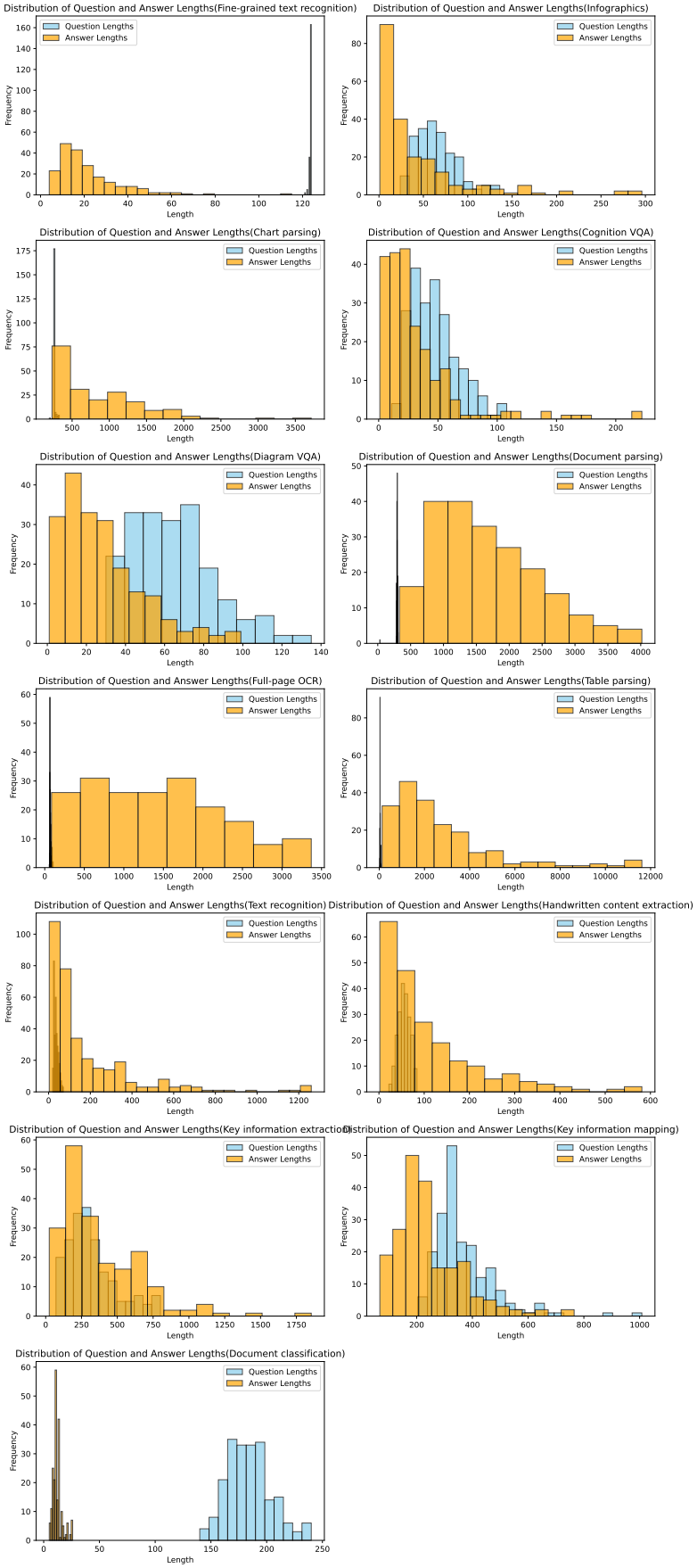


Figure 5: Comprehensive character-level statistics for both questions and answers in the ThaiOCRBench dataset.

Infographics VQA

Question

หากจ่ายค่าบริการด้วยธนบัตร 100 บาท จะได้ระยะเวลาในการนวดเท่าไหร่

Answer

75 นาที

Key information mapping

Question

ตามข้อมูลในภาพ กรุณาจับคู่และค่าที่เกี่ยวข้องดังต่อไปนี้: 'ปริมาณพลังงาน', 'คาร์โบไฮเดรต', 'ใยอาหาร', 'วิตามินซี' ,'ไขมันรวม' และ 'โซเดียม' กับค่าที่จะถูกจับคู่ได้แก่ '50 kcal', '0 mg.', '2 g.', '40%', '0 g.' และ '135 มก.' เลือกรายละเอียดจากเนื้อหาในเอกสารแล้วให้คำตอบกลับมาในรูปแบบ JSON โดยตรง:

Answer


```
{
    "ปริมาณพลังงาน": ["50 kcal"],
    "ไขมันรวม": ["0 g."],
    "คาร์โบไฮเดรต": ["0 mg."],
    "ใยอาหาร": ["2 g."],
    "วิตามินซี": ["40%"],
    "โซเดียม": ["135 มก."]
}
```


Figure 6: Examples of ThaiOCRBench: (Top) Infographics VQA, (Bottom) Key information mapping

Table parsing

Question

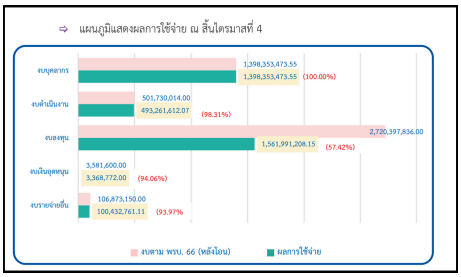
จงแปลงตารางทั้งหมดให้รูปภาพให้เป็นรูปแบบ HTML

Answer

<table> <tr> <td>สาขา-เขต
(Branch)</td> <td>ทะเบียนผู้ใช้น้ำ
(Account no.)</td> <td>เส้นทาง-ลำดับ
(Route)</td> <td>เลขที่แจ้งค่าน้ำ
(Invoice no.)</td> <td>วันที่แจ้งค่าน้ำ
(Invoice date)</td> </tr> <tr> <td>12-02</td> <td></td> <td>534-400</td> <td>017690-9</td> <td>03/01/68 09:42</td> </tr> <tr> <td>วันที่อ่านครั้งนี้
(Current date)</td> <td>เลขในมาตร
(Current reading)</td> <td>วันที่อ่านครั้งก่อน
(Previous date)</td> <td>เลขในมาตร
(Previous reading)</td> <td>จำนวนน้ำใช้
(Consumption)</td> </tr> <tr> <td>03/01/68</td> <td>2600</td> <td>03/12/67</td> <td>2574</td> <td>26</td> </tr> </table> <table> <tr> <td> <table> <tr> <td>ค่าน้ำดิบ (ลูกบาศก์เมตรละ...0.15...บาท)</td> <td>3.90</td> </tr> <tr> <td>ค่าน้ำประปา R1(00)</td> <td>221.00</td> </tr> <tr> <td>ส่วนลด</td> <td>0.00</td> </tr> <tr> <td>ค่าบริการรายเดือน</td> <td>25.00</td> </tr> <tr> <td>ยอดเงินก่อนคิดภาษี</td> <td>249.90</td> </tr> <tr> <td>ภาษีมูลค่าเพิ่ม 7%</td> <td>17.49</td> </tr> <tr> <td>รวมเงิน</td> <td>267.39</td> </tr> <tr> <td>ค่าน้ำเดือนก่อน.....0.....เดือน</td> <td>0.00</td> </tr> <tr> <td> รวมเงินที่ต้องชำระทั้งสิ้น Grand Total</td> <td> ******267.39</td> </tr> </table> </td> <!-- ส่วนประวัติและ QR Code (ขวา) --> <td valign="top"> <table> <tr> <th colspan="2">ประวัติการใช้น้ำประปา
(Consumption History) </th> </tr> <tr> <th>วันที่อ่านน้ำ
(Date)</th> <th>จำนวนน้ำใช้
(Consumption)</th> </tr> <tr> <td>03/10/67</td> <td>31</td> </tr> <tr> <td>03/11/67</td> <td>53</td> </tr> <tr> <td>03/12/67</td> <td>21</td> </tr> </table>
 [QR Code]
 QR Cross Bank </td> </tr> </table>

Figure 7: Examples of ThaiOCRBench: Table parsing

Chart parsing



Question

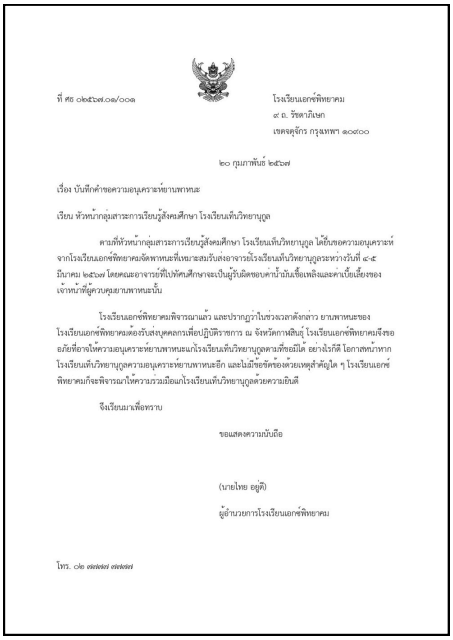
แปลงข้อมูลสำคัญในแผนภูมิเป็นโครงสร้าง Dict ในภาษา Python โดยต้องมีฟิลด์ 'title_th', 'title_en', 'chart_type', 'source' และ 'data' ให้ใส่ รายละเอียดทั้งหมดที่ได้จากแผนภูมิ (เช่น หมวดหมู่ จำนวน ร้อยละ) ลงใน 'data' หากไม่มีข้อมูลในฟิลด์ใด ให้ใส่เป็นสตริงว่าง

Answer


```
{
  "title_th": "แผนภูมิแสดงผลการใช้จ่าย ณ สิ้นไตรมาสที่ 4",
  "title_en":"",
  "source":"",
  "chart_type": "Bar Chart",
  "data": {
    "งบบุคลากร": {
      "งบตาม พรบ. 66 (หลังโอน)": 1398353473.55,
      "ผลการใช้จ่าย": 1398353473.55,
      "สัดส่วน": 100.00
    },
    "งบดำเนินงาน": {
      "งบตาม พรบ. 66 (หลังโอน)": 501730014.00,
      "ผลการใช้จ่าย": 493261612.07,
      "สัดส่วน": 98.31
    },
    "งบลงทุน": {
      "งบตาม พรบ. 66 (หลังโอน)": 2720397836.00,
      "ผลการใช้จ่าย": 1561991208.15,
      "สัดส่วน": 57.42
    },
    "งบเงินอุดหนุน": {
      "งบตาม พรบ. 66 (หลังโอน)": 3581600.00,
      "ผลการใช้จ่าย": 3368772.00,
      "สัดส่วน": 94.06
    },
    "งบรายจ่ายอื่น": {
      "งบตาม พรบ. 66 (หลังโอน)": 106873150.00,
      "ผลการใช้จ่าย": 100432761.11,
      "สัดส่วน": 93.97
    }
  }
}
```


Figure 8: Examples of ThaiOCRBench: Chart parsing

Document parsing



ที่ ศธ ๐๒๕๖๗.๐๑/๐๐๑

โรงเรียนเอกช์พิทยาคม
๙ ถ. รัชดาภิเษก
เขตจตุจักร กรุงเทพฯ ๑๐๙๐๐

๒๐ กุมภาพันธ์ ๒๕๖๗

เรื่อง บันทึกคำขอความอนุเคราะห์ยานพาหนะ

เรียน หัวหน้ากลุ่มสาระการเรียนรู้สังคมศึกษา โรงเรียนเท็นวิทยานุกูล

ตามที่หัวหน้ากลุ่มสาระการเรียนรู้สังคมศึกษา โรงเรียนเท็นวิทยานุกูล ได้ยื่นขอความอนุเคราะห์จากโรงเรียนเอกช์พิทยาคมจัดพาหนะที่เหมาะสมรับส่งอาจารย์โรงเรียนเท็นวิทยานุกูลระหว่างวันที่ ๔-๕ มีนาคม ๒๕๖๗ โดยคณะอาจารย์ที่ไปทัศนศึกษาจะเป็นผู้รับผิดชอบค่าน้ำมันเชื้อเพลิงและค่าเบี้ยเลี้ยงของเจ้าหน้าที่ผู้ควบคุมยานพาหนะนั้น

โรงเรียนเอกช์พิทยาคมพิจารณาแล้ว และปรากฏว่าในช่วงเวลาดังกล่าว ยานพาหนะของโรงเรียนเอกช์พิทยาคมต้องรับส่งบุคคลกรเพื่อปฏิบัติราชการ ณ จังหวัดกาฬสินธุ์ โรงเรียนเอกช์พิทยาคมจึงขออภัยที่อาจให้ความอนุเคราะห์ยานพาหนะแก่โรงเรียนเท็นวิทยานุกูลตามที่ขอมิได้ อย่างไรก็ดี โอกาสหน้าหากโรงเรียนเท็นวิทยานุกูลความอนุเคราะห์ยานพาหนะอีก และไม่มีข้อขัดข้องด้วยเหตุสำคัญใด ๆ โรงเรียนเอกช์พิทยาคมก็จะพิจารณาให้ความร่วมมือแก่โรงเรียนเท็นวิทยานุกูลด้วยความยินดี

จึงเรียนมาเพื่อทราบ

ขอแสดงความนับถือ

(นายไทย อยู่ดี)

ผู้อำนวยการโรงเรียนเอกช์พิทยาคม

โทร. ๐๒ ๗๗๗๗๗ ๗๗๗๗๗

Question

แปลงภาพเอกสารนี้ให้เป็น markdown

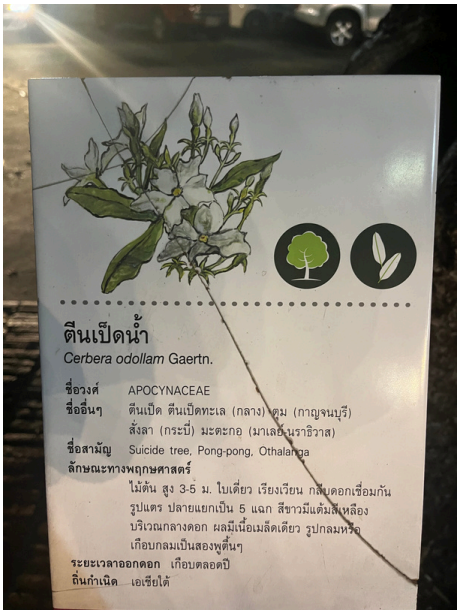
Answer

ที่ ศธ ๐๒๕๖๗.๐๑/๐๐๑

<figure>
</figure>

โรงเรียนเอกช์พิทยาคม
๙ ถ. รัชดาภิเษก
เขตจตุจักร กรุงเทพฯ ๑๐๙๐๐
๒๐ กุมภาพันธ์ ๒๕๖๗
เรื่อง บันทึกคำขอความอนุเคราะห์ยานพาหนะ
เรียน หัวหน้ากลุ่มสาระการเรียนรู้สังคมศึกษา โรงเรียนเท็นวิทยานุกูล
ตามที่หัวหน้ากลุ่มสาระการเรียนรู้สังคมศึกษา โรงเรียนเท็นวิทยานุกูล ได้ยื่นขอความอนุเคราะห์
จากโรงเรียนเอกช์พิทยาคมจัดพาหนะที่เหมาะสมรับส่งอาจารย์โรงเรียนเท็นวิทยานุกูลระหว่างวันที่ ๔-๕
มีนาคม ๒๕๖๗ โดยคณะอาจารย์ที่ไปทัศนศึกษาจะเป็นผู้รับผิดชอบค่าน้ำมันเชื้อเพลิงและค่าเบี้ยเลี้ยงของ
เจ้าหน้าที่ผู้ควบคุมยานพาหนะนั้น
โรงเรียนเอกช์พิทยาคมพิจารณาแล้ว และปรากฏว่าในช่วงเวลาดังกล่าว ยานพาหนะของ
โรงเรียนเอกช์พิทยาคมต้องรับส่งบุคคลกรเพื่อปฏิบัติราชการ ณ จังหวัดกาฬสินธุ์ โรงเรียนเอกช์พิทยาคมจึงขอ
อภัยที่อาจให้ความอนุเคราะห์ยานพาหนะแก่โรงเรียนเท็นวิทยานุกูลตามที่ขอมิได้ อย่างไรก็ดี โอกาสหน้าหาก
โรงเรียนเท็นวิทยานุกูลความอนุเคราะห์ยานพาหนะอีก และไม่มีข้อขัดข้องด้วยเหตุสำคัญใด ๆ โรงเรียนเอกช์
พิทยาคมก็จะพิจารณาให้ความร่วมมือแก่โรงเรียนเท็นวิทยานุกูลด้วยความยินดี
จึงเรียนมาเพื่อทราบ
ขอแสดงความนับถือ
(นายไทย อยู่ดี)
ผู้อำนวยการโรงเรียนเอกช์พิทยาคม
โทร. ๐๒ ๗๗๗๗๗ ๗๗๗๗๗

Full-page OCR



Question

แปลงข้อความทั้งหมดจากภาพนี้ แสดงออกมาเป็นข้อความธรรมดา (plain text)

Answer

ตีนเป็ดน้ำ Cerbera odollam Gaertn.
ชื่อวงศ์ APOCYNACEAE
ชื่ออื่นๆ ตีนเป็ด ตีนเป็ดทะเล (กลาง) ตุม (กาญจนบุรี) สั่งลา (กระบี่) มะตะกอ (มาเลย์-นราธิวาส)
ชื่อสามัญ Suicide tree, Pong-pong, Othalanga
ลักษณะทางพฤกษศาสตร์ ไม้ต้น สูง 3-5 ม. ใบเดี่ยว เรียงเวียน กลีบดอกเชื่อมกัน รูปแตร ปลายแยกเป็น 5 แฉก สีขาวมีแต้มสีเหลือง บริเวณกลางดอก ผลมีเนื้อเมล็ดเดียว รูปกลมหรือเกือบกลมเป็นสองพูตื้นๆ
ระยะเวลาออกดอก เกือบตลอดปี
ถิ่นกำเนิด เอเชียใต้

Figure 9: Examples of ThaiOCRBench: (Top) Document parsing; (Bottom) Full-page OCR

Fine-grained text recognition

Question

แบ่งความยาวและความสูงของรูปภาพออกเป็น 1000 ส่วน แล้วช่วยดึงข้อความที่อยู่ในพิกัด [221, 558, 518, 576] ของรูปภาพออกมาให้หน่อย

Answer

วิธีใช้: ทาบริเวณผิวหน้าและลำคอ ก่อนออกแดด 15 นาที เป็นประจำทุกวัน

Text recognition

Question

ในภาพนี้มีข้อความในป้ายเขียนไว้ว่าอะไร

Answer

Quick and Easy ซักแห้งง่ายๆ ไม่ต้องใช้น้ำ FREE! ผ้าเช็ด Microfiber

1. บีบโฟมใส่ผ้า 2 ครั้ง
2. เช็ดลงบนคราบสกปรก
3. เช็ดออกด้วยผ้าอีกด้าน

Figure 10: Examples of ThaiOCRBench: (Top) Fine-grained text recognition; (Bottom) Text recognition

Document classification

Question

เอกสารในภาพเป็นประเภทใด ระหว่าง ['เอกสารแบบฟอร์มราชการ', 'เอกสารราชการ', 'ทะเบียนสมรส', 'เอกสารการลงบันทึกประจำวันของตำรวจ', 'รายงานทางวิทยาศาสตร์', 'แบบฟอร์มภาษีเงินได้', 'ใบอนุญาตขับรถ', 'ไปรษณียบัตร']

Answer

แบบฟอร์มภาษีเงินได้

Diagram VQA

Question

จากภาพ สถานีหมายเลขใดที่เชื่อมต่อกับสถานีหมายเลข YL01

Answer

["สถานีหมายเลข BL15", "หมายเลข BL15", "BL15", "สถานีหมายเลข YL02", "หมายเลข YL02", "YL02"]

Figure 11: Examples of ThaiOCRBench: (Top) Document classification; (Bottom) Diagram VQA

Key information extraction

Question

รายการเครื่องดื่มในเมนูนี้มีอะไรบ้าง ราคาเท่าไหร่ จากรูปภาพกรุณาดึงข้อมูลมาในรูปแบบ JSON Dictionary ตัวอย่างเช่น:

{"ชื่อ": "ราคา", ...}

Answer


```
{
  "เอสเพรสโซ (Single)": "55 บาท",
  "เอสเพรสโซ (Double)": "65 บาท",
  "มอคคาซิโน่ (ร้อน)": "75 บาท",
  "มอคคาซิโน่ (เย็น)": "85 บาท",
  "คาปูชิโน่ (ร้อน)": "75 บาท",
  "คาปูชิโน่ (เย็น)": "85 บาท",
  "คาราเมล มัคคิอาโต้ (ร้อน)": "85 บาท",
  "คาราเมล มัคคิอาโต้ (เย็น)": "95 บาท",
  "ลาเต้ (ร้อน)": "75 บาท",
  "ลาเต้ (เย็น)": "85 บาท",
  "ยกล้อ (เย็น)": "85 บาท",
  "อเมริกาโน่ (ร้อน)": "75 บาท",
  "อเมริกาโน่ (เย็น)": "85 บาท",
  "โอเลี้ยง (เย็น)": "85 บาท",
  "ชาเขียวมัทฉะลาเต้ร้อน": "75 บาท",
  "ชาเขียวมัทฉะลาเต้เย็น": "85 บาท",
  "ชาทไวนิ่งส์": "65 บาท",
  "ชาเย็น (เย็น)": "85 บาท",
  "ชาดำเย็น (เย็น)": "85 บาท",
  "ไอซ์เลม่อนที (เย็น)": "85 บาท",
  "ไอซ์เอิร์ลเกรย์โรส (เย็น)": "85 บาท",
  "ป้อน สเปเชียล พันช์ (เย็น)": "115 บาท",
  "เวอจิ้น โมฮิโต้ (เย็น)": "95 บาท",
  "คลาสสิค ช็อกโกแลต (ร้อน)": "75 บาท",
  "คลาสสิค ช็อกโกแลต (เย็น)": "85 บาท",
  "คลาสสิค ช็อกโกแลต (ปั่น)": "95 บาท",
  "โรส-ลิ้นจี่โซดา (เย็น)": "85 บาท",
  "แพสชั่นฟรุ๊ตโซดา (เย็น)": "95 บาท",
  "น้ำผึ้งมะนาวโซดา (เย็น)": "85 บาท",
  "เฟรช ออเรนจ์ จูส (เย็น)": "85 บาท",
  "สเปเชียลพลัมกรีนแอปเปิ้ล (เย็น)": "85 บาท",
  "มะพร้าวลูก (เย็น)": "115 บาท",
  "มะพร้าวปั่น (ปั่น)": "135 บาท",
  "สตรอเบอร์รี่สมูทตี้ (ปั่น)": "95 บาท",
  "สโนว์โรสลิ้นจี่ (ปั่น)": "95 บาท",
  "แตงโมสดปั่น (ปั่น)": "95 บาท",
  "เป๊ปซี่": "25 บาท",
  "เป๊ปซี่แมกซ์": "25 บาท",
  "น้ำดื่ม": "20 บาท",
  "เบียร์ไฮเนเก้น": "110 บาท",
  "ช้างคลาสสิก": "100 บาท",
  "เบียร์สิงห์": "100 บาท"
}
```


Figure 12: Examples of ThaiOCRBench: Key information extraction

Handwritten content extraction

Question

ในภาพนี้ มีข้อความภาษาไทยที่ปรากฏอยู่บนกระดาน ข้อความเหล่านั้นคืออะไร

Answer

คนฮินดูในอินเดียมักดื่ม/อาบแม่น้ำคงคา แม้จะพิสูจน์ไม่ได้ว่าสะอาด แต่พวกเขาเชื่อว่าบริสุทธิ์ ...ถ้าคุณมีโอกาส...จะลองดื่ม/อาบหรือไม่?
ดร.อัมเบดการ์ไม่สนับสนุนระบบ-วรรณะ จะได้ไม่เกิดความเหลื่อมล้ำ เขาจึงเปลี่ยนไปนับถือและสนับสนุนศาสนาพุทธ ถ้าเป็นคุณ...จะนับถือหรือเปลี่ยน?

Cognition VQA

Question

ราคาของตู้เย็น Beko RDNT252I50HFK ปกติกี่บาท

Answer

["9,490 บาท", "9,490"]

Figure 13: Examples of ThaiOCRBench: (Top) Handwritten content extraction; (Bottom) Cognition VQA