BaFCo: A Document Understanding Benchmark for Complex Bangla Form Comprehension

Abu Tyeb Azad[1,⋆(✉)], Ishita Sur Apan[2,⋆], Fahim Ahmed[2,⋆], Sumaiya Karim Katha[2], Ezharuddin Jubaer[2], Armun Alam[2], Pranjal Kumar Nandi[3], Amin Ahsan Ali[2], Aman Chadha[4,‡], Md Mofijul Islam[4,†,‡], and AKM Mahbubur Rahman[2,†]

[1] Wichita State University, USA
[2] Center for Computational & Data Sciences, Bangladesh
[3] University of Dhaka, Bangladesh
[4] Amazon GenAI, USA

Abstract. Document comprehension is a challenging yet impactful task for Multimodal Large Language Models, especially as these systems see growing adoption in real-world, human-centric applications. However, this adoption is limited for low-resource languages such as Bangla due to the scarcity of high-quality annotated data. To address this gap, we introduce **BaFCo**, a benchmark dataset for Bangla form comprehension with a focus on Document Layout Analysis (DLA) and Key Information Extraction (KIE). BaFCo curates 200 multi-page complex Bangladeshi government forms, sourced from across diverse sectors including agriculture, education, banking, and land management. To accurately capture the structural and contextual complexity of these forms, we define a fine-grained annotation schema comprising 26 types of form entities, along with a separate coarse form entity set consisting of 5 types. We evaluate the latest MLLMs from the ChatGPT, Gemini, Claude, Qwen, and Kimi series using zero-shot and chain-of-thought prompts under both low and high reasoning setups. Our results reveal limitations in current MLLMs' ability in comprehending Bangla forms, particularly in accurately localizing highly granular form entities. Our dataset and code is available at: https://huggingface.co/datasets/Mausul/bafco

Keywords: Document Comprehension, Document Layout Analysis, Key Information Extraction, Low-resource NLP, Multimodal LLM

1 Introduction

Information retrieval from documents such as forms underpins many real-world systems including banking, education, and public administration [28]. **Document Layout Analysis** (DLA) and **Key Information Extraction** (KIE) [1, 8, 25] are tasks at the core of Document Understanding. DLA identifies the

⋆Equal contribution. †Equal supervision. ‡Work done outside role at Amazon.
(✉) Corresponding author: mausulazad495@gmail.com

structural elements of a page and the relationships between them. KIE involves locating and extracting values from form fields filled digitally or by hand. Both DLA and KIE are foundational precursors to downstream document tasks such as document question answering, entity linking, and summarization.

Bangla remains a low-resource language in document understanding despite being the world's 7th most spoken language, with 284 million speakers [24]. Government forms are also underrepresented in research, despite their semantic diversity and practical importance in public services. Benchmarks such as FUNSD [16] and XFUND [38] have advanced form understanding in English and other languages, but no comparable benchmark has been available for Bangla forms. Consequently, the lack of high-quality datasets and benchmarks continues to limit the development of Bangla document understanding systems.

To address this gap, we introduce **BaFCo**, a curated benchmark for multi-page Bangla form comprehension, focusing on government forms and the tasks of DLA and KIE. BaFCo provides fine-grained annotations spanning 26 entity types and labeled relationships between related fields. For DLA, the dataset contains 16,382 entities and 8,771 relationships across 200 forms (316 pages, with 1–5 pages per form). For KIE, it further includes 1,926 key-value pairs spanning 156 forms (186 pages). To maximize diversity, we prioritize complex and varied layouts over simpler forms. All annotations are created by trained annotators and reviewed by experts to ensure structural and semantic quality. In addition, we provide an end-to-end evaluation pipeline, standard document-comprehension metrics, and a coarse label set with five entity types, enabling systematic evaluation of MLLMs on Bangla DLA and KIE tasks.

With the advent of LLMs and MLLMs, document understanding workflows have increasingly shifted toward generative approaches, driven in part by the popularity of conversational interfaces. MLLMs are now widely used for document understanding tasks [3–5], replacing earlier OCR-based methods [33]. Although document-specialized models are available [10, 13, 30, 39], off-the-shelf flagship MLLMs offer a compelling alternative due to their lower development overhead through publicly accessible APIs, the scarcity of high-quality domain-specific training data, and the substantial computational resources required to train custom models [6, 37].

Consequently, we evaluate flagship MLLMs on BaFCo using tuning-free prompt-based methods to assess their layout-grounding capabilities and their understanding of Bangla-specific form elements. In DLA, `Gemini 3 Pro` performs the best across all experimental setup with average mAP scores of 0.1177 and 0.2646 for granular and coarse entity set, respectively. For KIE, `Gemini 3 Pro` again outperforms other models for Bangla forms, but for English ones `GPT-5.2` is surpassing others. We also evaluate the models on a set of English forms from the same domains to examine the effects of language. In case of DLA (particularly for granular entity set), the effect is minimal (performance difference ≤ 0.02), whereas for KIE `GPT-5.2` and `Claude Opus 4.6` performs better on English forms and `Gemini 3 Pro` on Bangla forms.

Our main contributions are as follows:

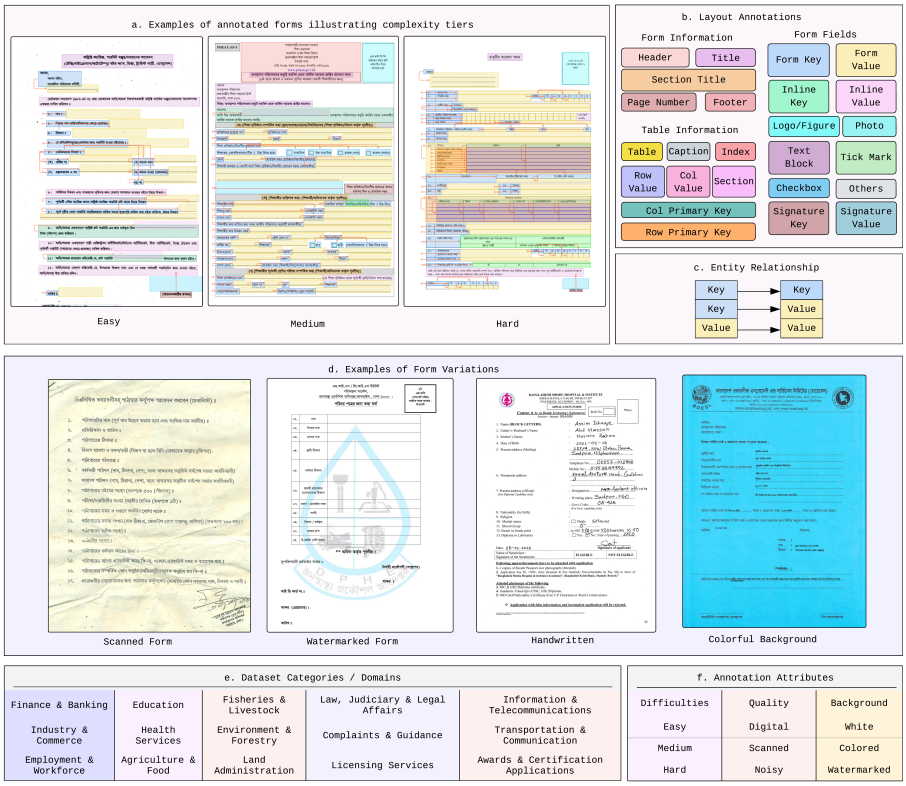

Fig. 1: Overview of BaFCo data diversity. (a) Examples of annotated forms across three difficulty levels: easy, medium, and hard (see Sec. 3.2); (b) BaFCo supports detailed layout annotations with 26 entity types, including titles, form key-value pairs, and tables; (c) It contains links between related entities through three relationship types: key-to-key, key-to-value, and value-to-value; (d) Examples of form variations, including scanned, watermarked, and handwritten forms; (e) BaFCo includes government forms from 15 domains; (f) Each form is annotated with nine page-level attributes based on annotation difficulty, image quality, and form background.

1. We introduce BaFCo, a high-quality dataset of Bangladeshi government forms with 26 entity types and annotated relationships between related fields, validated through human annotation and expert review.
2. We present the first benchmark for Document Layout Analysis (DLA) and Key Information Extraction (KIE) on Bangla forms, addressing a critical gap in low-resource document understanding.
3. We provide an empirical analysis of the performance of off-the-shelf flagship MLLMs on DLA and KIE tasks for Bangla documents, revealing limitations of the current models, particularly in handling fine-grained layouts.

2 Related Works

Document Layout Analysis (DLA) aims to identify and localize structural elements, such as text blocks, tables, images, and form fields. Early benchmarks spurred research from rule-based systems [11, 17] to machine learning [2, 36] for segmenting and classifying document regions. However, they struggle with complex real-world layouts. Hence, recent approaches predominantly rely on deep learning models [9, 13, 14, 19, 40] that combine visual, textual, and structural cues, including CNN–Transformer hybrids and multimodal architectures, which require high-quality annotations to generalize across varied domains and document types. Early foundational datasets like PubLayNet [40] and DocBank [18] scaled through leveraging scientific articles from PubMed Central and arXiv, respectively. However, these born-digital scientific datasets lack real-world layout variability. DocLayNet [23] addressed this gap by providing data from diverse sources; including financial reports, manuals, and legal documents.

Key Information Extraction (KIE) focuses on identifying and extracting form key-value pairs. Among generative KIE approaches, GenKIE [5] uses an encoder-decoder architecture with zero-shot prompt based evaluation. In BROS [12] uses a multimodal, layout-aware encoder. LiLT [32] and OmniDocBench [21] are multilingual, nonetheless Bangla is absent from their pool. Apart from GPT4o in OmniDocBench, no work evaluates recent flagship MLLMs' on KIE.

Form Understanding is an important subfield of document intelligence, focusing on extracting, structuring, and linking information from forms. FUNSD [16] was among the first public benchmarks in this area, providing annotations for questions, answers, headers, and other text entities, along with question–answer relationships. It enabled research on semantic entity recognition and relation extraction, but its annotation schema is largely limited to flat question–answer structures. Subsequent datasets have targeted specific domains and document types, including government business documents [31], financial forms [7], receipts [15, 22, 29], and legal and financial documents [27]. While these datasets contain challenging layouts and long-form content, they generally lack fine-grained text-entity annotations and spatial relationships required for detailed form structure understanding, particularly in low-resource and domain-specific settings.

Multilingual and Low-Resource Document Datasets Most document layout analysis benchmarks focus on high-resource languages, leaving low-resource settings underexplored. XFUND [38] introduced a multilingual benchmark covering seven languages. However, many low-resource languages remain absent. Bangla is one such example: despite its large speaker base, public benchmarks for Bangla document understanding have been scarce. A first step was BaDLAD [26], which introduced a multi-domain dataset spanning books, newspapers, government documents, and historical records. Although it provides 710k polygon annotations, its labels are limited to coarse layout elements such as text regions, paragraphs, tables, and images. However, it does not support fine-grained form understanding with form-specific entities, key–value relationships, or government-form structures.

Table 1: Comparison of **BaFCo** with existing document understanding benchmarks. **Domain** denotes the no. of document domains (e.g., finance, education, healthcare). **Form** indicates support for form-style documents; **LRL**, low-resource languages; **Multipage**, multi-page annotations; and **Human**, human-created annotations. **BBox** and **Text** denote bounding-box and text annotations, respectively. **Tab**, **Img**, **Sig**, **Cb**, and **Photo** indicate annotations for tables, images (e.g., figures, logos, diagrams), signatures, checkboxes (including single-select, multi-select, and radio buttons), and photographs. **EL** denotes entity linking between related fields. **Tasks** lists supported tasks: Document Layout Analysis (DLA) and Key Information Extraction (KIE).

Benchmark	**Data**			**Annotation Attributes and Types**										**Tasks**	
	Domain	Form	LRL	Multipage	Human	BBox	Text	Tab	Img	Sig	Cb	Photo	EL	DLA	KIE
DLA Benchmarks															
PubLayNet [40]	2	–	–	–	–	✓	✓	✓	✓	–	–	–	–	✓	–
DocBank [18]	1	–	–	✓	–	✓	✓	✓	✓	–	–	–	–	✓	–
DocLayNet [23]	5	–	–	–	✓	✓	✓	✓	✓	–	–	–	–	✓	–
OmniDocBench [21]	9	–	–	✓	✓	✓	✓	✓	✓	–	–	–	–	✓	–
Form Benchmarks															
BuDDIE [31]	3	✓	–	–	✓	✓	✓	–	–	–	–	–	–	–	✓
FUNSD [16]	1	✓	–	–	✓	✓	✓	–	–	–	–	–	✓	✓	✓
XFUND [38]	1	✓	–	–	✓	✓	✓	–	–	–	–	–	✓	–	✓
FormNLU [7]	1	✓	–	–	✓	✓	✓	–	–	–	–	–	✓	✓	✓
Low-Resource Language Benchmark															
BaDLAD [26]	4	–	✓	–	✓	✓	✓	✓	✓	–	–	–	–	✓	–
BaFCo (ours)	15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Our work addresses this gap with BaFCo, the first publicly available benchmark for Bangla form comprehension and layout analysis. BaFCo provides fine-grained annotations spanning 26 entity types and labeled relationships between related entities (e.g., linking field labels to values), a capability not previously available for Bangla documents.

3 Dataset

3.1 Data Collection and Curation

We adopt a quality-over-quantity strategy, prioritizing careful selection, diversity, and annotation fidelity over scale. The dataset is composed of publicly available Bangladeshi government forms[5], ensuring authenticity and real-world relevance. Government forms were chosen due to their structured yet highly variable layouts, dense semantic content, and practical importance.

Forms were filtered to remove duplicates, incomplete scans, and low-quality images that could introduce annotation noise. The selected forms span multiple administrative domains, including taxation, healthcare, education, and civil services. They exhibit varying degrees of layout complexity, including multi-column structures, nested fields, tables, and handwritten or stamped regions. We further

[5] https://forms.portal.gov.bd/

categorize forms into three difficulty levels—easy, medium, and difficult—based on layout density, visual clutter, and semantic coupling between fields.

Unlike large-scale multilingual benchmarks such as OmniDocBench [21] or BuDDIE [31], which emphasize breadth across document types and languages, our dataset is purpose-built for Bangla forms. Accordingly, comparisons in this work focus on Bangla-specific and multilingual datasets that include Bangla content. The dataset is intended as a high-quality benchmark for low-resource document layout analysis rather than a general-purpose corpus.

3.2 Dataset Description

Form Selection Criteria and Difficulty Levels: Forms containing 1–5 pages were first shortlisted. The selected forms were then grouped into difficulty levels based on layout complexity and component diversity. *Easy* forms contained basic elements such as form key–value pairs (including one-to-many and key-to-key relationships), inline and signature key–value pairs, headers, titles, section headings, text blocks, photo fields, and footers. However, they did not include tables, checkboxes, or tick marks. *Medium* forms introduced additional structural complexity through checkboxes, tick marks, and tables consisting only of columns, while explicitly excluding tables with both rows and columns, sub-columns, or embedded checkboxes and tick marks. *Hard* forms represented the highest level of complexity and included tables with rows, columns, and sub-columns; tables containing checkboxes or tick marks; and other unique, dense, or unconventional layout components.

DLA Dataset Composition: Following the difficulty-based grouping, BaFCo encompasses a diverse range of form structures, including application and non-application forms, sparse and dense layouts (ranging from forms with few fields to those containing crowded or heavily tabular regions), and both guided forms (with explicit bounding boxes) and unguided forms (without such cues). These variations are representative of real-world government forms. BaFCo also exhibits a deliberately skewed entity-class distribution, preserving the natural long-tail distribution of entities found in real-world documents rather than imposing artificial class balance. Common elements (e.g., form keys and values) dominate the dataset, whereas specialized elements (e.g., signatures and tabular sub-fields) are relatively rare, reflecting the inherent long-tailed nature of government forms. Overall, the dataset contains 16,382 entities and 8,771 relationships across 200 forms (316 pages), with each form comprising between 1 and 5 pages.

KIE Dataset Composition: For KIE, we evaluate 156 forms (186 pages and 1,926 key–value pairs) spanning the same Easy, Medium, and Hard difficulty levels. English forms are included to facilitate comparisons between a high-resource language and a low-resource language. The Bangla-to-English ratio is 1.14:1 by form count and 1.24:1 by key–value pair count.

Table 2: BaFCo dataset statistics for the DLA and KIE tasks, broken down by difficulty level. Under our difficulty rubric, easy forms do not contain tables or selection elements (checkboxes/tick marks); consequently, these fields are absent from the Easy category and are denoted by '–'.

	Easy	Medium	Hard	Total
Dataset Size				
DLA Forms	74	66	60	200
DLA Pages	90	108	118	316
DLA Tables	–	59	143	202
KIE Forms	86	42	28	156
KIE Pages	95	57	34	186
Structure				
DLA Avg. entities/form	55.76	77.76	118.73	81.91
DLA Avg. relations/form	30.01	37.74	67.65	43.85
KIE Avg. key-value pairs/form	11.2	13.5	15.0	12.3

	Easy	Medium	Hard	Total
Annotation Types				
DLA Form values	1,418	1,179	1,079	3,676
DLA Inline values	155	198	216	569
DLA Signatures	187	189	144	520
DLA Checkboxes	–	245	166	411
DLA Tick marks	–	65	51	116
DLA Table rows	–	313	622	935
DLA Table columns	–	262	963	1,225
DLA Total annotations	1,760	2,451	3,241	7,452
KIE Filled key-value pairs	955	568	391	1,926

3.3 Semantic Form Entity Definition

Based on our analysis of Bangladeshi government forms, we defined a fine-grained and comprehensive taxonomy of semantic form entities. The dataset contains 26 entity types covering structural components (e.g., headers, sections, and tables), functional fields (e.g., key–value pairs, checkboxes, and signatures), and auxiliary elements (e.g., instructions, seals, and stamps) across both single-page and multi-page forms (see Tab. 3). To examine the effect of entity granularity on model performance, we additionally group these 26 entity types into five coarse-grained categories.

In addition to entity labels, we annotate relationships between semantically linked fields, enabling the use of BaFCo for downstream document understanding tasks. Tab. 2 summarizes the total number of annotated entities and relationships across the different difficulty levels. The taxonomy for each entity type was iteratively refined through pilot annotations and expert feedback to ensure coverage, consistency, and applicability across diverse form layouts. This detailed taxonomy distinguishes our dataset from existing Bangla document resources, which typically employ generic or task-specific entity label sets.

Table 3: Semantic Form Entities. Both coarse and granular form entities are enlisted.

Coarse	Granular
Headers	Header, Title, Footer, Section Title
Table	Table, Table Caption, Table Section, Table Index, Row Primary Key, Row Value, Column Primary Key, Column Value
Image	Diagram / Logo / Figure
Fields	Form Key, Form Value, Inline Key, Inline Value, Signature Key, Signature Value, Tick Mark, Checkbox, Image Field
Others	Text Block, Page Number, Gibberish, Others

3.4 Annotation Guidelines

DLA: To ensure high annotation quality and reproducibility, we developed a detailed annotation guideline document that specifies entity definitions, bounding box rules, relationship constraints, and edge cases. Particular emphasis is placed on resolving ambiguities common in Bangla forms, such as overlapping text regions, visually implicit field boundaries, and mixed printed–handwritten content. The guideline is refined through multiple rounds of annotator feedback and validation, resulting in a consistent annotation protocol that minimizes subjective interpretation while retaining flexibility for complex layouts.

KIE: For KIE, we annotate bounding boxes and define explicit relationships between regions. Two labels — *key* and *value* — were used. Annotators filled a text field for each region: key fields copied the annotated text, while value fields recorded the corresponding entry. These fields served as ground truth. To evaluate models, values were overlaid on form images, and LLMs were prompted to extract them given the associated keys. To assess model performance, the values were superimposed onto the form images, and the LLMs were prompted to extract the values given the corresponding keys.

3.5 Annotation Procedure

17 annotators underwent two days of training using guideline documents, practice forms, and tutorial videos covering dataset nuances. All were proficient in Bangla and experienced in document labeling. Each document was independently annotated in *Label Studio*[6] by a trained annotator and reviewed by an expert reviewer. Disagreements or uncertainties in labels or bounding boxes were resolved via group discussion and majority voting. To quantify reliability, we measured agreement between the trained annotator and the expert reviewer on the pre-majority-vote annotations, obtaining a Cohen's κ of 0.974.

3.6 Difficulties Faced During Annotation:

Most annotation errors stemmed from layout complexity and fell into three categories. *Semantic confusions* arose when distinguishing closely related entity types: in fill-in-the-blank forms, standard key–value pairs were frequently mislabeled as inline or signature pairs, and slashes ("/") were occasionally mistaken as separators between Tick Mark entities. *Structural ambiguities* were most common in guided layouts, where form keys were hard to separate from table cells, and where a few multi-cell tables were annotated cell-by-cell rather than as a single structure. *Technical inconsistencies* included misclassification, duplicate boxes, and overly broad inline annotations, along with occasional omissions of essential elements (e.g., header logos, signature links) and erroneous inclusion of irrelevant footer text. All these issues were identified and corrected by expert reviewers prior to finalization.

[6] https://labelstud.io/

4 Experiments

4.1 Experimental Setup

Models: We evaluate five flagship MLLMs spanning both proprietary and open-source ecosystems: `GPT-5.2`, `Gemini 3 Pro`, and `Claude Opus 4.6` on the proprietary side, and `Qwen 3.6 Plus` and `Kimi K2.5` among open-source models. We focus exclusively on MLLMs because they represent the most promising paradigm for general-purpose document understanding. Unlike OCR-based pipelines and encoder architectures, which are typically designed for predefined tasks with fixed output schemas, MLLMs operate in an open-ended, instruction-driven manner that better aligns with emerging document AI applications. Consequently, direct comparisons with OCR or encoder-based systems are less informative for our objectives. We therefore center our evaluation on MLLMs, which are also widely accessible through API-based deployment. For inference, we use native batch endpoints for proprietary models to improve cost efficiency and *OpenRouter*[7] endpoints for open-source models.

Prompts: To utilize MLLMs' inference-time task-completion capabilities, we used zero-shot [34] and Chain-of-Thought (CoT) prompting [35]. We included task descriptions, form entity specifications, and output-structure instructions in the prompts. For KIE, we followed the prompt structure of GenKIE [5], where KIE is formulated as a visual question-answering task.

Reasoning Effort: Besides CoT prompting, for DLA experiments we also use the built-in reasoning of the flagship models. We run every setup at two reasoning levels, *low* and *high*, set via the *reasoning_effort* API parameter. We use this as our main control because it is the only setting shared across providers for adjusting models' reasoning level. Since black-box APIs do not reveal each model's internals, it is the fairest basis for comparison. To cap runaway generation, we limit *max_output_tokens* at 16,000 for low reasoning and 64,000 for high reasoning. For consistency, we use only the *low* and *high* levels, as these settings were available across all providers (`Gemini 3 Pro`'s API did not offer a *medium* reasoning effort level at the time of writing).

DLA Evaluation Metrics: We evaluate model predictions using standard detection and classification metrics following prior form understanding and object detection benchmarks [7, 20, 21]. As in [7, 21], we perform greedy matching between predicted and ground-truth bounding boxes with two Intersection-over-Union (IoU) thresholds $\tau \in 0.3, 0.5$ to ensure order-invariant assignment.

A predicted box is counted as a true positive (TP) if matched to a ground-truth instance with IoU $\geq \tau$. Unmatched predictions are false positives (FP), and ground-truth instances without corresponding predictions are false negatives

[7] https://openrouter.ai/

(FN). Using these counts, we compute precision, recall, F1 score, and mean average precision (mAP). For a fixed IoU threshold τ, mAP is obtained by averaging class-wise Average Precision (AP) across all form field categories.

KIE Evaluation Metrics: Following prior work [5, 7, 21], we use precision, recall, F1, and Normalized Edit Similarity (NES) for KIE evaluation.

NES is defined as $NES = 1 - NED$, where $NED = \frac{d(s,t)}{\max(|s|,|t|)}$ and $d(s,t)$ is the Levenshtein distance between the predicted string s and ground-truth t.

4.2 Evaluation Setup

To evaluate MLLMs' performance for each task $t \in T$ where $T = \{DLA,\ KIE\}$, image of page j of form i, $\mathbf{b}_{i,j}$ is selected from BaFCo dataset B. Then along with prompt p ($p \in P_t$) the page image $\mathbf{b}_{i,j}$ is passed to each MLLM, M_θ from a pool of models (see Sec. 4.1),to generate raw prediction $\tilde{y}^{p}_{i,j}$.

$$\tilde{y}^{p}_{i,j} = M_\theta(\mathbf{b}_{i,j},\, p) \tag{1}$$

The raw prediction $\tilde{y}^{p_t}_{i,j}$ is subsequently processed by a validator function $V(\cdot)$ that checks for adherence of predictions to a predefined output schema customized for DLA and KIE. Invalid inferences are excluded from evaluation.

$$\hat{y}^{p_t}_{i,j} = V\big(\tilde{y}^{p_t}_{i,j}\big) \tag{2}$$

The validated prediction is then compared against the ground-truth annotation $y_{i,j}$ to obtain page-level performance measurements aggregated across pages and forms. For each predicted form entity, the output also contains a brief label justification and a confidence score for the bounding box coordinates.

For DLA, we define a prompt set as $P_{DLA} = \{p_{\text{zs}},\ p_{\text{cot}}\}$, where p_{zs} and p_{cot} denote DLA-specific zero-shot and chain-of-thought prompts, respectively. For KIE the prompt set P_{KIE} contains a zero-shot evaluation prompt. To observe effects of entity set size, in addition to the granular form entity set containing 26 entity categories, we developed a minimal set containing 5 coarse entity categories and mapped all original 26 categories to coarse categories (See Tab. 3).

5 Results

5.1 Document Layout Analysis (DLA)

Tab. 4 reports DLA performance across models, prompting strategies, reasoning effort levels, and entity granularities. Overall, `Gemini 3 Pro` achieves the best results in most configurations, followed by `GPT-5.2`, while `Claude Opus 4.6` is consistently the weakest. Performance improves noticeably from the granular to the coarse entity set. For instance, under high reasoning effort with zero-shot prompting, `Gemini 3 Pro` increases from 0.1177 mAP on the granular entity set to 0.2646 on the coarse set at IoU@0.3, with similar trends observed across

Table 4: Layout analysis performance for both Granular and Coarse label sets. **Bold** indicates the best overall result per column within each entity set, and Blue indicates the best result within each combination of prompt variant (Zero-Shot: ZS, Chain-of-Thoughts: CoT) and reasoning effort (Low, High). μIoU is used to show overall average IoU regions for all predicted bounding boxes. μIoU@[0.3, 0.5] is used to show average of bounding boxes that are over the thresholds 0.3 and 0.5, the thresholds are selected based on works from Form-NLU [7] and OmniDocBench [21]. Across all prompt, reasoning effort, and entity set granularity level combination, `Gemini 3 Pro` outperforms `GPT-5.2` and `Claude Opus 4.6`. For each metric higher (↑) is better.

Reasoning Effort	Prompt	Model	IoU@0.3				IoU@0.5			
			mAP	F1	μIoU	μIoU@0.3	mAP	F1	μIoU	μIoU@0.5
Granular Entity Set										
Low	ZS	GPT-5.2	0.0530	0.1158	0.1242	0.4556	0.0155	0.0444	0.1259	0.6175
		Claude Opus 4.6	0.0041	0.0161	0.0316	0.4334	0.0010	0.0049	0.0323	0.5998
		Gemini-3 Pro	0.0900	0.1953	**0.2430**	**0.5647**	0.0483	0.1222	**0.2449**	**0.6787**
		Kimi K2.5	0.0156	0.0566	0.0676	0.4787	0.0048	0.0196	0.0695	0.6087
		Qwen 3.6-Plus	0.0097	0.0422	0.0691	0.4184	0.0008	0.0088	0.0705	0.6048
	CoT	GPT-5.2	0.0643	0.1273	0.1343	0.4782	0.0266	0.0618	0.1349	0.5991
		Claude Opus 4.6	0.0037	0.0174	0.0355	0.4206	0.0007	0.0051	0.0363	0.5907
		Gemini-3 Pro	0.0853	0.1943	0.2420	0.5390	0.0452	0.1207	0.2433	0.6594
		Kimi K2.5	0.0105	0.0480	0.0690	0.4530	0.0027	0.0182	0.0707	0.5989
		Qwen 3.6-Plus	0.0083	0.0371	0.0650	0.4101	0.0010	0.0072	0.0717	0.6025
High	ZS	GPT-5.2	0.0511	0.1192	0.1148	0.4658	0.0157	0.0481	0.1157	0.6134
		Claude Opus 4.6	0.0074	0.0321	0.0460	0.4408	0.0012	0.0075	0.0465	0.5771
		Gemini-3 Pro	**0.1177**	**0.2282**	0.2243	0.5312	0.0641	0.1411	0.2257	0.6419
		Kimi K2.5	0.0453	0.1117	0.1132	0.4617	0.0150	0.0441	0.1140	0.6123
		Qwen 3.6-Plus	0.0175	0.0606	0.0661	0.4501	0.0040	0.0234	0.0668	0.5910
	CoT	GPT-5.2	0.0615	0.1291	0.1234	0.4723	0.0197	0.0581	0.1247	0.5985
		Claude Opus 4.6	0.0065	0.0310	0.0501	0.4350	0.0016	0.0104	0.0505	0.5683
		Gemini-3 Pro	0.1134	0.2206	0.2193	0.5438	**0.0643**	**0.1435**	0.2205	0.6586
		Kimi K2.5	0.0476	0.1133	0.1162	0.4559	0.0152	0.0435	0.1175	0.6130
		Qwen 3.6-Plus	0.0166	0.0609	0.0657	0.4329	0.0052	0.0216	0.0665	0.5959
Coarse Entity Set										
Low	ZS	GPT-5.2	0.1800	0.2974	0.2742	0.5177	0.0852	0.1654	0.2767	0.6495
		Claude Opus 4.6	0.0168	0.0555	0.0697	0.4081	0.0042	0.0138	0.0722	0.5875
		Gemini-3 Pro	0.2578	0.4020	0.3602	0.6066	**0.1606**	0.2847	0.3616	0.6986
		Kimi K2.5	0.0469	0.1213	0.1306	0.4529	0.0131	0.0392	0.1344	0.6433
		Qwen 3.6-Plus	0.0330	0.1273	0.1511	0.4111	0.0021	0.0296	0.1584	0.5865
	CoT	GPT-5.2	0.1671	0.2861	0.2659	0.5174	0.0781	0.1571	0.2682	0.6409
		Claude Opus 4.6	0.0131	0.0578	0.0731	0.4164	0.0022	0.0132	0.0751	0.5884
		Gemini-3 Pro	0.2513	0.4051	0.3682	0.6085	0.1440	0.2827	0.3698	0.7002
		Kimi K2.5	0.0426	0.1293	0.1342	0.4385	0.0057	0.0359	0.1398	0.6033
		Qwen 3.6-Plus	0.0377	0.1199	0.1512	0.4123	0.0021	0.0250	0.1577	0.5981
High	ZS	GPT-5.2	0.1434	0.2866	0.2561	0.5046	0.0532	0.1391	0.2592	0.6472
		Claude Opus 4.6	0.0440	0.1168	0.1181	0.4283	0.0136	0.0454	0.1199	0.5932
		Gemini-3 Pro	**0.2646**	**0.4116**	**0.3729**	**0.6125**	0.1480	**0.2915**	**0.3745**	**0.7003**
		Kimi K2.5	0.1336	0.2638	0.2259	0.4748	0.0466	0.1190	0.2283	0.6205
		Qwen 3.6-Plus	0.0589	0.1692	0.1583	0.4740	0.0204	0.0714	0.1604	0.6143
	CoT	GPT-5.2	0.1354	0.2776	0.2549	0.5158	0.0663	0.1546	0.2575	0.6340
		Claude Opus 4.6	0.0358	0.1012	0.1112	0.4518	0.0091	0.0385	0.1132	0.5829
		Gemini-3 Pro	0.2444	0.3928	0.3499	0.6071	0.1426	0.2751	0.3511	0.6975
		Kimi K2.5	0.1361	0.2602	0.2245	0.4874	0.0472	0.1217	0.2272	0.6281
		Qwen 3.6-Plus	0.0693	0.1749	0.1615	0.4766	0.0200	0.0733	0.1637	0.6218

other models. Among the open-source models, `Kimi K2.5` and `Qwen 3.6-Plus` trail the proprietary leaders under low reasoning effort. However, `Kimi K2.5` benefits substantially from increased reasoning effort: its coarse mAP@0.3 under zero-shot prompting rises from 0.0469 to 0.1336, approaching `GPT-5.2` (0.1434), whereas `Qwen 3.6-Plus` improves only modestly, from 0.0330 to 0.0589.

Increasing reasoning effort has mixed effects, yielding small gains in some settings but slight decreases in Avg. IoU in others. CoT prompting likewise provides limited benefit, often performing comparably to zero-shot prompting.

The effect of language on MLLM performance is minimal. Tab. 5 compares Bangla and English performance for the two strongest models, `Gemini 3 Pro` and `GPT-5.2`. For both models, the two languages yield similar results across most configurations, with ΔmAP typically within 0.02 on the granular entity

Table 5: Bangla and English Form Comparison for DLA (IoU@0.3) under different experimental setups. LR and HR stand for Low Reasoning and High Reasoning respectively. ZS and CoT denote Zero-Shot and Chain-of-Thought prompts. Shapes indicate higher performance for Bangla (●), English (■), or neither (⧫) language. For granular entity set, the effect of language is minimal ($\leq$ 0.02). For coarse entity set, effect of language is slightly more with maxΔ of 0.12. For each metric higher (↑) is better.

Variant	Model	Bangla		English		Difference (Bn-En)	
		mAP	F1	mAP	F1	Δ mAP	Δ F1
Granular Entity Set							
LR + ZS	GPT-5.2	0.09	0.16	0.09	0.14	⧫ 0.00	● 0.02
	Gemini 3 Pro	0.16	0.24	0.15	0.24	● 0.01	⧫ 0.00
LR + CoT	GPT-5.2	0.12	0.18	0.11	0.15	● 0.01	● 0.03
	Gemini 3 Pro	0.15	0.21	0.14	0.23	● 0.01	■ 0.02
HR + ZS	GPT-5.2	0.09	0.16	0.11	0.15	■ 0.02	● 0.01
	Gemini 3 Pro	0.16	0.24	0.16	0.23	⧫ 0.00	● 0.01
HR + CoT	GPT-5.2	0.12	0.18	0.13	0.17	■ 0.01	● 0.01
	Gemini 3 Pro	0.15	0.21	0.16	0.24	■ 0.01	■ 0.03
Coarse Entity Set							
LR + ZS	GPT-5.2	0.34	0.44	0.22	0.32	● 0.12	● 0.12
	Gemini 3 Pro	0.38	0.51	0.35	0.42	● 0.03	● 0.09
LR + CoT	GPT-5.2	0.29	0.44	0.28	0.40	● 0.01	● 0.04
	Gemini 3 Pro	0.35	0.47	0.25	0.40	● 0.09	● 0.07
HR + ZS	GPT-5.2	0.34	0.44	0.27	0.35	● 0.07	● 0.09
	Gemini 3 Pro	0.38	0.51	0.36	0.47	● 0.02	● 0.04
HR + CoT	GPT-5.2	0.29	0.44	0.27	0.38	● 0.02	● 0.07
	Gemini 3 Pro	0.35	0.47	0.42	0.52	■ 0.07	■ 0.05

set. Overall, DLA performance remains largely consistent across both languages and relatively low for granular entities.

5.2 Key Information Extraction (KIE)

Table 6: KIE evaluation results on Bangla (BN) and English (EN) forms. Normalized Edit Similarity (NES) is reported along with Precision, Recall, and Macro-F1. English forms are selected from the same domain as Bangla. `Gemini 3 Pro` performs the best for Bangla forms, whereas for English `GPT-5.2` performs better.

Model	Precision		Recall		Macro-F1		NES	
	BN	EN	BN	EN	BN	EN	BN	EN
Proprietary								
GPT-5.2	0.784	**0.852**	0.780	**0.844**	0.781	**0.847**	0.796	**0.851**
Claude Opus 4.6	0.675	0.839	0.685	0.832	0.678	0.835	0.704	0.840
Gemini 3 Pro	**0.844**	0.832	**0.853**	0.827	**0.848**	0.828	**0.866**	0.833
Open-Source								
Kimi K2.5	0.684	0.835	0.681	0.825	0.681	0.828	0.700	0.832
Qwen 3.6-Plus	0.798	0.839	0.793	0.834	0.794	0.835	0.804	0.844

Tab. 6 reports KIE performance using Precision, Recall, F1, and Normalized Edit Similarity (NES). All models score substantially higher here than on DLA. Among proprietary models, `Gemini 3 Pro` leads on Bangla (F1 0.848, NES 0.866), while `GPT-5.2` leads on English (F1 0.847, NES 0.851). `Claude`

Opus 4.6 trails both but remains competitive, unlike in DLA, where it was the weakest performer. Among open-source models, Qwen 3.6-Plus performs best on both languages (Bangla F1 0.794, English F1 0.835), ahead of Kimi K2.5. It ranks second overall on Bangla, surpassing GPT-5.2 (F1 0.781) and Claude Opus 4.6 (F1 0.678), and matches Claude Opus 4.6 on English (F1 0.835), narrowing the gap between open-source and proprietary models.

Unlike DLA, where language differences were minimal, KIE varies more clearly by language. Most models score higher on English—the gap is largest for Claude Opus 4.6 (0.678 vs. 0.835 F1) and Kimi K2.5 (0.681 vs. 0.828), while only Gemini 3 Pro favors Bangla (0.848 vs. 0.828). Consequently, the best model is language-dependent: Gemini 3 Pro on Bangla and GPT-5.2 on English. So cross-lingual differences are stronger in KIE compared to layout detection.

6 Discussion

Granularity of form entities strongly affects DLA performance. Across all models, performance improves markedly when coarse entity set is used. Reducing the number of entity types substantially increases both mAP and F1 across most configurations. For example, mAP score for Gemini-3 Pro improves from 0.1177 to 0.2646, for granular to coarse entity set shift for IoU@0.3, more than doubling performance. This indicates that much of the challenge in DLA lies in distinguishing fine-grained entity types; when grouped into broader categories, models localize layout regions more reliably.

Increasing reasoning effort does not consistently improve geometric precision. Across both entity sets and prompting strategies, higher reasoning effort yields only small and inconsistent performance changes. In several cases, it slightly reduces localization quality, particularly Avg. IoU. For example, Gemini-3 Pro shows lower Avg. IoU when moving from low to high reasoning effort in the granular setting, while GPT-5.2 shows small improvements in some configurations. Overall, this mixed behavior suggests that additional reasoning does not reliably improve geometric alignment; layout prediction appears to depend more on visual–spatial pattern recognition than extended reasoning.

Chain-of-thought prompting provides limited gains for DLA. Across most models and settings, CoT prompting does not clearly outperform zero-shot prompting; performance often remains similar or slightly declines in mAP and F1. For example, Gemini-3 Pro achieves slightly lower mAP with CoT than with zero-shot in multiple settings across both entity sets. This suggests that encouraging multi-step reasoning offers limited benefit for DLA, which appears to rely more on visual–spatial understanding than explicit reasoning chains.

Language has limited impact on DLA performance. Across most experimental variants, performance differences between Bangla and English forms are small, with negligible gaps in mAP and F1 (see Tab. 5). In the granular entity setting, differences are near zero (e.g., ΔmAP ≤ 0.02 across most setups). Even in the coarse setting, improvements are inconsistent and modest relative to

overall performance. This suggests that DLA difficulty for MLLMs stems mainly from geometric localization and structural understanding rather than language.

MLLMs perform substantially better on KIE than on DLA. Across all models, KIE results are considerably stronger than document layout analysis performance. The best-performing model achieves F1 scores above 0.84 on Bangla forms and 0.80 on English forms (see Tab. 6), while DLA results remain much lower across comparable configurations. This indicates that MLLMs are more effective at textual understanding and semantic extraction than at precise geometric localization of layout regions.

Language differences are more pronounced in KIE than in DLA. Unlike DLA experiments—where Bangla and English forms yield nearly identical performance. KIE results show clearer language-dependent variation across models. For example, `Gemini 3 Pro` achieves the best performance on Bangla forms (F1 = 0.848), while `GPT-5.2` performs best on English forms (F1 = 0.800) (see Tab. 6). This suggests that language-specific factors affect text-heavy extraction tasks, while layout detection remains largely language-agnostic.

7 Conclusion

We propose **BaFCo**, the first Bangla document understanding dataset featuring well-curated multi-domain Bangladeshi government forms. It includes an expanded, fine-grained set of 26 form entities designed for better Bangla document comprehension. We evaluated flagship MLLMs on DLA and KIE using both granular and coarse entity sets, revealing significant performance gaps. DLA performance is inconsistent across models and sensitive to reasoning effort and prompting strategies, whereas KIE results are generally stronger, though some lexical mismatches remain.

8 Limitations and Future Works

The scope of generative model–based Bangla form comprehension can be expanded along several directions:

(a) Our experiments focus on DLA and KIE, key building blocks for downstream tasks like summarization and ontology generation. Future work will explore frontier models on broader document comprehension tasks.

(b) Low-resource languages face greater challenges in curating high-quality domain-specific documents due to limited data access, lack of experts, and other constraints. Although BaFCo contains fewer annotated forms than high-resource datasets like [21, 38, 40], it aims to stimulate research on Bangla form comprehension. Future work will expand both the quantity and diversity of forms.

(c) To address model limitations and scarcity of large-scale Bangla data, lightweight post-training methods and agentic AI approaches can be explored.

Acknowledgement

This work was partially funded by Independent University, Bangladesh (IUB).

References

1. Binmakhashen, G.M., Mahmoud, S.A.: Document layout analysis: A comprehensive survey. ACM Comput. Surv. **52**(6) (Oct 2019). https://doi.org/10.1145/3355610
2. Bukhari, S.S., Al Azawi, M.I.A., Shafait, F., Breuel, T.M.: Document image segmentation using discriminative learning over connected components. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. p. 183–190. DAS '10, Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1815330.1815354
3. Cao, H., Li, X., Ma, J., Jiang, D., Guo, A., Hu, Y., Liu, H., Liu, Y., Ren, B.: Query-driven generative network for document information extraction in the wild. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 4261–4271. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3547877
4. Cao, H., Ma, J., Guo, A., Hu, Y., Liu, H., Jiang, D., Liu, Y., Ren, B.: GMN: Generative multi-modal network for practical document information extraction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3768–3778. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.276
5. Cao, P., Wang, Y., Zhang, Q., Meng, Z.: GenKIE: Robust generative multimodal document key information extraction. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 14702–14713. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.979
6. Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., Owen, D.: The rising costs of training frontier ai models. arXiv preprint arXiv:2405.21015 (2024)
7. Ding, Y., Long, S., Huang, J., Ren, K., Luo, X., Chung, H., Han, S.C.: Form-nlu: Dataset for the form natural language understanding. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). pp. 2807–2816. Association for Computing Machinery, Taipei, Taiwan (2023). https://doi.org/10.1145/3539618.3591886
8. Ding, Y., Luo, S., Dai, Y., Jiang, Y., Li, Z., Martin, G., Peng, Y.: A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. arXiv preprint arXiv:2507.09861 (2025)
9. Duan, Y., Chen, Z., Hu, Y., Wang, W., Ye, S., Shi, B., Lu, L., Hou, Q., Lu, T., Li, H., Dai, J., Wang, W.: Docopilot: Improving multimodal models for document-level understanding. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4026–4037 (2025). https://doi.org/10.1109/CVPR52734.2025.00381
10. Fujitake, M.: LayoutLLM: Large language model instruction tuning for visually rich document understanding. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 10219–10224. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.892/

11. Ha, J., Haralick, R., Phillips, I.: Recursive x-y cut using bounding boxes of connected components. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. vol. 2, pp. 952–955 vol.2 (1995). https://doi.org/10.1109/ICDAR.1995.602059
12. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10767–10775 (2022). https://doi.org/https://doi.org/10.1609/aaai.v36i10.21322
13. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3096–3120. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.175
14. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 4083–4091. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548112
15. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520 (2019). https://doi.org/10.1109/ICDAR.2019.00244
16. Jaume, G., Kemal Ekenel, H., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6 (2019). https://doi.org/10.1109/ICDARW.2019.10029
17. Journet, N., Eglin, V., Ramel, J., Mullot, R.: Text/graphic labelling of ancient printed documents. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 1010–1014 Vol. 2 (2005). https://doi.org/10.1109/ICDAR.2005.235
18. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank: A benchmark dataset for document layout analysis. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 949–960. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.82
19. Liao, W., Wang, J., Li, H., Wang, C., Huang, J., Jin, L.: Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4038–4049 (June 2025)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
21. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., Shi, J., Wu, F., Chu, P., Liu, M., Li, Z., Xu, C., Zhang, B., Shi, B., Tu, Z., He, C.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24838–24848 (2025). https://doi.org/10.1109/CVPR52734.2025.02313

22. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: Cord: a consolidated receipt dataset for post-ocr parsing. In: Workshop on Document Intelligence at NeurIPS 2019 (2019)
23. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: Doclaynet: A large human-annotated dataset for document-layout segmentation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 3743–3751. KDD '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539043
24. Poria, S., Huang, X.: Bhaasha, bhāṣā, zaban: A survey for low-resourced languages in South Asia – current stage and challenges. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 1386–1406. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.73
25. Sassioui, A., Benouini, R., El Ouargui, Y., El Kamili, M., Chergui, M., Ouzzif, M.: Visually-rich document understanding: Concepts, taxonomy and challenges. In: 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM). pp. 1–7 (2023). https://doi.org/10.1109/WINCOM59760.2023.10322990
26. Shihab, M.I.H., Hasan, M.R., Emon, M.R., Hossen, S.M., Ansary, M.N., Ahmed, I., Rakib, F.R., Dhruvo, S.E., Dip, S.S., Pavel, A.H., Meghla, M.H., Haque, M.R., Chowdhury, S.S., Sadeque, F., Reasat, T., Humayun, A.I., Sushmit, A.: Badlad: A large multi-domain bengali document layout analysis dataset. In: Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part I. p. 326–341. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-41676-7_19
27. Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., Biecek, P.: Kleister: Key information extraction datasets involving long documents with complex layouts. In: Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I. p. 564–579. Springer-Verlag, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-86549-8_36
28. United Nations Department of Economic and Social Affairs: United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development - With the addendum on Artificial Intelligence. United Nations (2024), https://www.un-ilibrary.org/content/books/9789211067286#overview
29. Šimsa, v., Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Skalický, M., Matas, J., Doucet, A., Coustaty, M., Karatzas, D.: Docile benchmark for document information localization and extraction. In: Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part II. p. 147–166. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-41679-8_9
30. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., Liu, X.: DocLLM: A layout-aware generative language model for multimodal document understanding. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8529–8548. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.463
31. Wang, D., Zmigrod, R., Sibue, M.J., Pei, Y., Babkin, P., Brugere, I., Liu, X., Navarro, N., Papadimitriou, A., Watson, W., Ma, Z., Nourbakhsh, A., Shah, S.:

BuDDIE: A business document dataset for multi-task information extraction. In: Chen, C.C., Moreno-Sandoval, A., Huang, J., Xie, Q., Ananiadou, S., Chen, H.H. (eds.) Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal). pp. 35–47. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.finnlp-1.3/

32. Wang, J., Jin, L., Ding, K.: LiLT: A simple yet effective language-independent layout transformer for structured document understanding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (May 2022). https://doi.org/10.18653/v1/2022.acl-long.534
33. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 23318–23340. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/wang22al.html
34. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
35. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
36. Wu, C.C., Chou, C.H., Chang, F.: A machine-learning approach for analyzing document layout structures with two reading orders. Pattern Recogn. **41**(10), 3200–3213 (Oct 2008). https://doi.org/10.1016/j.patcog.2008.03.014
37. Xia, Y., Kim, J., Chen, Y., Ye, H., Kundu, S., Hao, C.C., Talati, N.: Understanding the performance and estimating the cost of llm fine-tuning. In: 2024 IEEE International Symposium on Workload Characterization (IISWC). pp. 210–223 (2024)
38. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: XFUND: A benchmark dataset for multilingual visually rich form understanding. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 3214–3224. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.253
39. Zhang, J., Yang, W., Lai, S., Xie, Z., Jin, L.: Dockylin: a large multimodal model for visual document understanding with efficient visual slimming. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press (2025). https://doi.org/10.1609/aaai.v39i9.33076
40. Zhong, X., Tang, J., Jimeno Yepes, A.: Publaynet: Largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022 (2019). https://doi.org/10.1109/ICDAR.2019.00166