BuDDIE: A Business Document Dataset for Multi-task Information Extraction

Dongsheng Wang*, **Ran Zmigrod***, **Mathieu Sibue***, **Yulong Pei**,
Petr Babkin, Ivan Brugere, Xiaomo Liu, Nacho Navarro, Antony Papadimitriou,
William Watson, Zhiqiang Ma, Armineh Nourbakhsh, Sameena Shah
JPMorgan AI Research
first.last@jpmchase.com

Abstract

The field of visually rich document understanding (VRDU) aims to solve a multitude of well-researched NLP tasks in the multi-modal domain. Several datasets exist for research on specific tasks of VRDU, such as document classification (DC), key entity extraction (KEE), entity linking, visual question answering (VQA), *inter alia*. These datasets cover documents like invoices and receipts with sparse annotations such that they support one or two co-related tasks (e.g., entity extraction and entity linking). Unfortunately, only focusing on a single specific type of documents or task is not representative of how documents often need to be processed in the wild – where variety in style and requirements is expected. In this paper, we introduce **BuDDIE** (**Bu**siness **D**ocument **D**ataset for **I**nformation **E**xtraction),[1] the first multi-task dataset of 1,665 real-world business documents that contains rich and dense annotations for DC, KEE, and VQA. Our dataset consists of publicly available business entity documents from US state government websites. The documents are structured and vary in their style and layout across states and types (e.g., forms, certificates, reports, etc.). We provide data variety and quality metrics for BuDDIE as well as a series of baselines for each task. Our baselines cover traditional textual, multi-modal, and large language model approaches to VRDU.

Dataset	Types	Tasks	Docs	Labels	OCR
CORD	Receipts	$\mathcal{K}$	1,000	30	✓
DeepForm	Receipts	$\mathcal{K}$	1,100	5	✓
DocILE	Receipts	$\mathcal{K}$	7,000	55	✓
DocVQA	Varied	$\mathcal{Q}$	12,767	–	✓
DUDE	Varied	$\mathcal{Q}$	4,973	–	✓
FUNSD	Forms	$\mathcal{K}, \mathcal{L}$	199	4	✓
Kleister Char.	Reports	$\mathcal{K}$	540	8	✓
Kleister NDA	Legal	$\mathcal{K}$	2,778	4	✓
NAF	Forms	$\mathcal{K}, \mathcal{L}$	860	14	✓
RVL-CDIP	Varied	$\mathcal{C}$	400,000	16	✗
SROIE	Receipts	$\mathcal{K}$	1,000	4	✓
VRDU Ad-buy	Receipts	$\mathcal{K}$	641	10	✓
VRDU Reg.	Forms	$\mathcal{K}$	1,915	6	✓
BuDDIE	Varied	$\mathcal{C}, \mathcal{K}, \mathcal{Q}$	1,665	69	✓

Table 1: Existing VRDU dataset information. Tasks Legend: DC ($\mathcal{C}$), Entity linking ($\mathcal{L}$), KEE ($\mathcal{K}$), VQA ($\mathcal{Q}$). Note that OCR is not available for the original versions of DeepForm, Kleister Charity, and Kleister NDA. However, Borchmann et al. (2021) provides OCR for these datasets.

1 Introduction

Document images are ubiquitous in the real world, especially in the financial industry. Reports, receipts, forms, certificates, *inter alia*, are integral throughout the business pipeline. For example, during the Know Your Customer (KYC) process in banking, officers must conduct due diligence by reviewing documents such as government registration forms, financial reports, organizational charts, and other relevant materials to verify the customers' identities. This kind of process is usually conducted manually, which is extremely challenging due to massive data volumes and widely varying data formats. Modern systems thus need to efficiently and accurately capture and understand information from digital or scanned documents. As a result, computer vision, machine learning, and NLP researchers have focused on creating models for VRDU (Xu et al., 2020; Appalaraju et al., 2021; Davis et al., 2021; Xu et al., 2021; Zhang et al., 2022). With rising interest in the field, the necessity for publicly available, large, and robust datasets is becoming ever-more evident.

Numerous datasets have been created to support the modeling of individual document understanding tasks such as document classification (DC), key entity extraction (KEE), entity linking, and visual question answering (VQA) (Jaume et al., 2019; Park et al., 2019; Stanisławek et al., 2021; Mathew

*Equal contribution.

[1]Full dataset available for non-commercial use upon request at airdata.requests@jpmorgan.com

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 35–47
January 19–20, 2025.

et al., 2021). Datasets often contain ground-truth annotations, based on optical character recognition (OCR), that support a single or two co-related document understanding tasks. For example, RVL-CDIP (Harley et al., 2015) contains annotations for DC, and FUNSD (Jaume et al., 2019) provides annotations for KEE and entity linking. The majority of VRDU datasets, specifically those targeting forms and receipts, are designed for KEE. (Davis et al., 2019; Huang et al., 2019; Park et al., 2019; Simsa et al., 2023; Wang et al., 2023).

In this paper, we introduce **BuDDIE**, a new dataset comprised of 1,665 publicly available structured business documents from US state government websites. Our dataset is unique in that it tackles multiple distinct VRDU tasks over the same documents: DC, KEE, and VQA. Such a dataset is particularly beneficial to assess document processing in the wild, where requirements may necessitate models to perform several tasks on the same input. We created a hierarchical ontology of 69 key entity classes over seven super categories that can be augmented with even more entity types in the future. These provide a semantically rich and annotation-dense KEE dataset which enables us to construct a varied VQA set. While similarly sized or larger VRDU datasets exist (Stanisławek et al., 2021; Simsa et al., 2023; Wang et al., 2023), they tend to focus on a single sort of document (e.g., receipts) and task. This may be insufficient for general purpose models that may be required in industry to accurately infer on a plethora of document types. Therefore, BuDDIE contributes a new large and varied dataset to the field.

Our contributions are summarized below:

- We present BuDDIE, a new annotated dataset consisting of 1,665 structured business documents. BuDDIE is the first VRDU dataset that supports three distinct tasks: DC, KEE, and VQA. Furthermore, it can be extended to facilitate multi-turn VQA, instruction tuning, and other downstream tasks with minimal additional effort. BuDDIE is publicly available for non-commercial use.
- We provide six baselines for each task in BuDDIE: two traditional text-only language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2020); two multi-modal language models, LayoutLM (Xu et al., 2020) and LayoutLMv3 (Huang et al., 2022); and finally, two large language models (LLMs), GPT4 and DocLLM (Wang et al., 2024), where DocLLM incorporates multi-modal information into the language model. The best baseline across all tasks, DocLLM, achieves a DC F1 of 99.15, KEE F1 score of 89.97, and VQA ANLS score of 89.58.

2 Related Work

In this section, we describe past datasets from the VRDU community as well as models.

2.1 Datasets

The RVL-CDIP dataset (Harley et al., 2015), which consists of 400,000 business related documents annotated for DC, is one of the first VRDU datasets that was released. It solves an important but somewhat coarse-grained task, and RVL-CDIP is now mainly used to pre-train models. Most modern VRDU datasets target information and entity extractions, which were first introduced in 2019 when FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), and CORD (Park et al., 2019) were released. While the latter two focused on receipt documents, FUNSD (**F**orm **U**nderstanding in **N**oisy **S**canned **D**ocuments) introduced the tasks of entity extraction and entity linking over forms. It provided annotations of 199 form documents from the RVL-CDIP dataset. FUNSD annotates entities as *question*, *answer*, *header*, or *other*. FUNSD is, however, more targeted at form structure extraction as its entities have structural rather than semantic meaning and are connected via entity linking. FUNSD later received a revision that corrected annotation errors found in the original version (Vu and Nguyen, 2020), and has also been adapted for the task of form parsing (Zmigrod et al., 2024). While FUNSD is commonly used for VRDU fine tuning and evaluation, its small size means it may be unreliable for comparing larger models (Borchmann et al., 2021).

CORD (**Co**nsolidated **R**eceipt **D**ataset) and SROIE (**S**canned **R**eceipt **O**CR and **I**nformation **E**xtraction) are KEE datasets for receipts. SROIE provides 1,000 documents with four semantic key entity labels that are commonly found in receipts, along with text localisation and OCR output. CORD contains a richer key entity label set. It consists of 1,000 receipt documents that contain 30 unique key entities subsumed by four super categories.[2] Inspired by the CORD label ontology,

[2] 1,000 documents out of the 11,000 claimed in the CORD

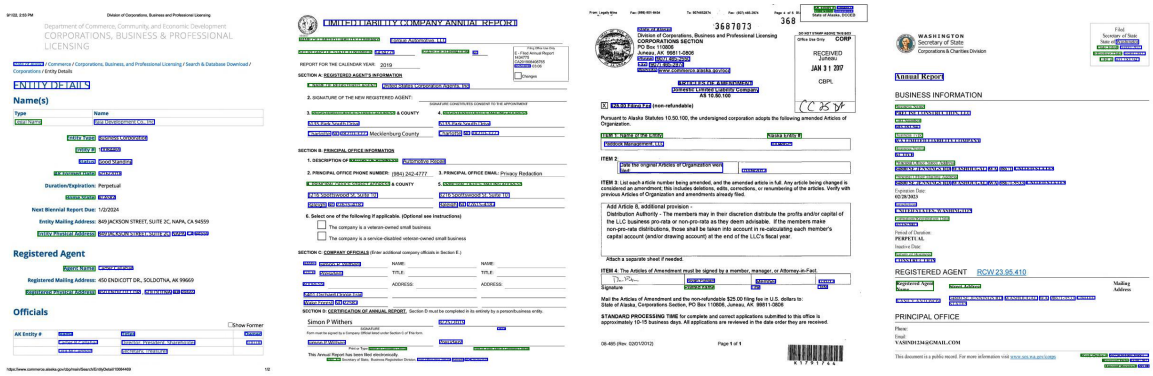

Figure 1: Examples of varied document styles in BuDDIE with KEE annotations (entity labels are omitted for document format clarity).

we designed our own key entity label ontology in Section 3.3. More recently, DocILE (Simsa et al., 2023), a large dataset containing 7,000 real-world receipts and 100,000 synthetically generated receipts, annotated for KEE, has been introduced and used in the literature. Their KEE task contains 55 fine-grained labels. Other KEE datasets cover additional document styles such as registration forms, NDAs, advertisements, *inter alia* (Stanisławek et al., 2021; Wang et al., 2023).

DocVQA (Mathew et al., 2021) introduced the task of VQA to the VRDU community. The dataset comprises 12,767 document images (6,071 total documents) from a wide variety of document types (e.g., forms, letters, and reports) with a total of 50,000 questions. Recently, a new document VQA dataset, DUDE (Landeghem et al., 2023), has been proposed to offer a more varied VQA dataset. Though non-English VRDU datasets also exist (Qi et al., 2022; Xu et al., 2022), this work only considers English datasets. We provide a detailed comparison of the datasets described above with BuDDIE in Table 1.

2.2 Models

Early VRDU models incorporated textual and visual features in parallel and then merged them together. Most commonly, a pre-trained transformer was used to embed spatially localized text and a pre-trained CNN-based model was used to encode the visual features (Denk and Reisswig, 2019; Wang et al., 2020; Xu et al., 2020; Garncarek et al., 2021; Lin et al., 2021; Zhang et al., 2020). Subsequent models enabled richer interactions between text, spatial, and visual features by using a single multi-modal Transformer (Appalaraju et al., 2021; Powalski et al., 2021; Xu et al., 2021; Peng et al., 2022; Huang et al., 2022; Tang et al., 2023).

Other VRDU model architectures also exist in the literature. For example, (Davis et al., 2021; Zhang et al., 2022; Lee et al., 2023) opted for graph-based approaches. While graph-based methods still use the full multi-modal pipeline, some works have also discarded certain elements. Li et al. (2021); Hong et al. (2022) abandoned visual features and instead solely relied on text and bounding box information. On the other hand, a few recent models have experimented with vision-only approaches to reduce the need of OCR (Davis et al., 2022; Kim et al., 2022). Recently, LLMs have been increasingly used for VRDU tasks. LLM architectures such as DocLLM (Wang et al., 2024) make use of text and layout features, while models such as mPLUG-DocOwl (Ye et al., 2023) leverage both text and general vision.

3 The Business Document Dataset for Information Extraction

In this paper, we introduce a new dataset for VRDU, **BuDDIE**, which consists of 1,665 publicly available business documents. In particular, we searched documents from US state websites (or their department of business website) which were under one of five document classes of interest shown in Table 2. We obtained documents for Puerto Rico and all but eight of the 50 states. Documents from Illinois, Indiana, Louisiana, Maine, Mississippi, Texas, Colorado, and Michigan are either blocked by a paywall or restricted for distribution, so they are not included in our dataset. Table 5 provides a breakdown of the number of

paper were made public. The original version of CORD featured 54 unique entity labels over eight super categories (Park et al., 2019), but some labels and super categories were removed since.

Class	Examples	Total	Train	Val	Test
Amendment Document	Article of Amend., Change of Address, Statement of Change	85	60	9	16
Application or Article	Application for Corporation, Article of Org., Name Reservation	153	111	12	30
Business Entity Details	Business Search Results, State Registry	815	570	81	164
Certificate or Statement	Certificate of Reinstatement, Statement of Good Standing	90	64	9	17
Periodical Report	Annual Report, Biennial Report	522	367	50	105
Total		1,665	1,172	161	332

Table 2: Example document titles and number of occurrences for BuDDIE document classes.

documents collected per US state. The documents of BuDDIE are partially structured, i.e., documents fall into styles such as forms, certificates, etc. Examples of the varied structures and formats in the dataset are given in Figure 1. BuDDIE targets three prominent tasks in VRDU: DC, KEE, and VQA. To the best of our knowledge, no current VRDU dataset tackles all three of these tasks simultaneously, and no dataset of this size exists for KEE over multiple document types. Furthermore, due to the rich and multi-task annotation scheme, our dataset has the potential to be extended to support multi-turn VQA, instruction tuning, as well as other downstream VRDU tasks (e.g., entity linking), with minimal additional effort. This could be of particular interest when considering multi-modal LLMs (Ye et al., 2023; Wang et al., 2024).

In the remainder of this section, we describe the data collection, annotation, and processing steps for each of the three tasks. Our annotation instructions are provided in App. A.

3.1 Document Processing

The initial collection for raw data yielded 1,890 documents. Many documents contained multiple pages, however, we only used the first page of each document in order to reduce annotation cost. We also observed that the first page tends to be the most complex in terms of layout and style in many US state filings. We used OCR to extract the text elements of each document. More precisely, our annotation tool uses PDFPlumber to extract the OCR tokens and decide on a reading order. Throughout the annotation process described in Section 3.2 and Section 3.3, 150 documents were discarded due to poor OCR quality, lack of entities (fewer than five), or incompatibility with the document classes defined in Table 2. A further 75 documents were discarded due to copyright issues. After the annotation process, we created a train, validation, and test split of 70%, 20%, and 10% respectively. The split was done using stratified sampling on the document classes. In future, we plan to release train, validation, and test splits based on states, i.e., some states will be held out for the validation and test sets. This will work towards assessing generalization to unseen document styles.

3.2 Document Classification

Document classification is the task of assigning a label to a document to denote its semantic or structural content. For example, RVL-CDIP categorises documents based on their style (e.g., form, letter, resume). In BuDDIE, document classes have a semantic meaning. The classes defined in Table 2 contain an underlying structural separation as well as semantic differences. For instance, *Business Entity Details* and *Periodical Report* documents typically present a form-based format, while *Certificate or Statement* documents tend to be more closely linked to letters. There may exist semantic ambiguity and overlap between our classes; for example, an *Article of Amendment* could be classified as *Amendment Document* or *Article or Application*. Therefore, we constructed a list of ordered annotation rules for annotators to follow; we provide these rules in App. A. In our above *Article of Amendment* example, we rank the amendment documents higher than the other article documents, and so the document considered would fall into the *Amendment Document* category. We provide examples for each document class in Table 2.

The DC annotation task was split between five annotators who have prior experience in the VRDU field. There were two rounds of annotation: (1) an initial annotation task to assign each document a class, and (2) a validation task to verify the labels that resulted from the initial round. If there were repeated disagreements between an annotator and a validator, a third annotator would discuss discrepancies with both and decide on the final label based on the rules and discussions. Documents for which

Super Category	Label	Fine-grained Entity Examples	Total	Train	Val	Test
Business Entity	ENT	ENT_name, ENT_number, ENT_type	13,884	9,703	1,339	2,842
Entity Key Personnel	KP	KP_address_street, KP_name, KP_title	9,845	6,853	906	2,086
File Attribute	FILE	FILE_date, FILE_name, FILE_number,	4,028	2,840	410	778
Government Official	GO	GO_adress_city, GO_name, GO_title	3,046	2,197	280	569
Other	OTHER	OTHER_address, OTHER_date, OTHER_unknown	839	638	48	153
Registered Agent	AGT	AGT_address_city, AGT_address_state, AGT_name	6,072	4,248	582	1,242
Signature	SIG	SIG_KP_date, SIG_KP_printed_name, SIG_KP_title	1,192	850	93	249
Total			38,906	27,329	3,658	7,919

Table 3: BuDDIE key entity extraction super categories. For each super category, we provide the three most common fine-grained entity labels and the total number of occurrences of the super category.

Entity Label	Total	Train	Val	Test	Entity Label	Total	Train	Val	Test	Entity Label	Total	Train	Val	Test
AGT_adrs_city	1174	820	113	241	ENT_residency	152	106	19	27	GO_fax	39	28	2	9
AGT_adrs_country	240	162	24	54	ENT_shares_auth	50	43	3	4	GO_telephone	262	182	23	57
AGT_adrs_state	1150	806	112	232	ENT_shares_issued	50	33	4	13	GO_website	212	146	23	43
AGT_adrs_street	1146	802	109	235	ENT_status	806	552	85	169	GO_name	480	360	47	73
AGT_adrs_zipcode	1148	804	109	235	ENT_type	1041	727	103	211	GO_title	627	462	60	105
AGT_name	1214	854	115	245	FILE_adrs_city	70	50	9	11	KP_adrs_city	1413	972	144	297
ENT_NAICS	107	70	13	24	FILE_adrs_state	114	81	13	20	KP_adrs_country	490	350	37	103
ENT_adrs_city	1552	1083	142	327	FILE_adrs_street	71	50	9	12	KP_adrs_state	1374	953	130	291
ENT_adrs_country	377	253	44	80	FILE_adrs_zipcode	71	50	9	12	KP_adrs_street	1488	1026	141	321
ENT_adrs_state	1500	1046	140	314	FILE_date	907	633	90	184	KP_adrs_zipcode	1383	958	134	291
ENT_adrs_street	1485	1050	137	298	FILE_due_date	235	163	29	43	KP_name	1934	1337	171	426
ENT_adrs_zipcode	1450	1010	135	305	FILE_eff_date	155	115	15	25	KP_shares_owned	78	58	12	8
ENT_alt_name	29	23	1	5	FILE_exp_date	48	40	2	6	KP_title	1685	1199	137	349
ENT_am_adrs_city	21	19	1	1	FILE_fee	398	284	44	70	OTHER_unknown	522	404	24	94
ENT_am_adrs_state	21	19	1	1	FILE_name	927	660	77	190	OTHER_adrs	95	64	13	18
ENT_am_adrs_street	23	21	1	1	FILE_number	494	345	47	102	OTHER_date_time	185	142	9	34
ENT_am_adrs_zipcode	20	18	1	1	FILE_state	300	214	38	48	OTHER_name	37	28	2	7
ENT_am_name	16	13	2	1	FILE_type	238	155	28	55	SIG_GO_date	19	16	1	2
ENT_cob	295	200	30	65	GO_adrs_city	344	247	30	67	SIG_GO_name	29	23	1	5
ENT_formation_date	704	487	69	148	GO_adrs_state	343	245	30	68	SIG_GO_title	45	35	4	6
ENT_jurisdiction	863	600	84	179	GO_adrs_street	328	235	27	66	SIG_KP_date	317	227	22	68
ENT_name	1890	1330	184	376	GO_adrs_zipcode	342	245	30	67	SIG_KP_name	488	343	37	108
ENT_number	1432	1000	140	292	GO_email	69	47	8	14	SIG_KP_title	294	206	28	60

Table 4: Number of occurrences in the train, validation, and test splits of BuDDIE for each key entity label.

no agreement was reached or which did not fall into any of our five document classes were discarded from the dataset. In total, four documents were discarded due to the above reasons.

3.3 Key Entity Extraction

Key entity extraction is the most popular task in VRDU. The task is akin to a named entity recognition problem where each entity represents a key piece of information. As documents vary in their content, KEE label sets tend to be large. For example, CORD and DocILE, two similar datasets to ours, have label sets of 30 and 55 labels, respectively. We offer a larger set of 69 labels, since we focus on a wider domain (general business rather than receipts). Like CORD and DocILE, we create our label set using super categories and specific detailed types. In total, we consider six super categories: *business entity*, *entity key personnel*, *file attribute*, *government official*, *registered agent*, and *signature*. We additionally have an *other* super category. Under these seven super categories, we then have 69 fine-grained labels. We give frequency statistics for each of the super categories in Table 3 and a finer-grained analysis in Table 4 of all labels.

The KEE annotation task was performed similarly to the DC annotation task. The collection of documents was split between 12 annotators with past experience in VRDU who used the PAWLS annotation tool (Neumann et al., 2021) to draw bounding boxes around relevant key entities. Any OCR token that laid in the bounding box was then highlighted for the annotation.[3] After the initial annotation round, each document was then validated by a different annotator. If a validator found repeated inconsistencies with any annotations with regards to the annotation instructions, a third annotator would be consulted. Any annotation in question either reached agreement across the three annotators or was discarded. Annotators were instructed to only annotate an entity if they were confident in the specific annotation. Consequently,

[3]The annotation tool also enabled free-form bounding boxes that were not bound to OCR tokens. While annotators were allowed to make such annotations, they were not included in this version of the dataset as our models assume the existence of OCR tokens. A future version of this dataset may include free-form bounding boxes as well as OCR based bounding boxes.

State	State Abb.	Total	Train	Val	Test	State	State Abb.	Total	Train	Val	Test
Alabama	AL	40	30	4	6	New Hampshire	NH	40	29	4	7
Alaska	AK	68	47	8	13	New Jersey	NJ	36	25	0	11
Arizona	AZ	78	52	6	20	New Mexico	NM	19	12	3	4
Arkansas	AR	46	32	6	8	New York	NY	18	12	4	2
California	CA	25	19	2	4	North Carolina	NC	122	92	9	21
Connecticut	CT	18	17	1	0	North Dakota	ND	10	9	0	1
Delaware	DE	12	11	1	0	Ohio	OH	11	8	0	3
Florida	FL	34	24	1	9	Oklahoma	OK	30	22	2	6
Georgia	GA	58	41	6	11	Oregon	OR	29	21	2	6
Hawaii	HI	35	25	3	7	Pennsylvania	PA	53	35	5	13
Idaho	ID	30	19	3	8	Puerto Rico	PR	20	14	3	3
Iowa	IA	96	72	9	15	Rhode Island	RI	26	21	1	4
Kansas	KS	35	22	3	10	South Dakota	SD	88	61	7	20
Kentucky	KY	65	47	7	11	Tennessee	TN	24	14	4	6
Maryland	MD	23	19	1	3	Utah	UT	9	3	2	4
Massachusetts	MA	32	25	1	6	Vermont	VT	19	12	3	4
Minnesota	MN	20	11	4	5	Virginia	VA	35	23	6	6
Missouri	MO	50	33	9	8	Washington	WA	40	20	8	12
Montana	MT	20	15	1	4	West Virginia	WV	10	8	0	2
Nebraska	NE	20	12	4	4	Wisconsin	WI	20	15	4	1
Nevada	NV	40	24	4	12	Wyoming	WY	161	119	10	32

Table 5: Number of occurrences in the train, validation, and test splits of BuDDIE for US states.

Question Type	Total	Train	Val	Test
Boolean No	1,032	739	100	193
Boolean Yes	1,067	742	116	209
Span	6,571	4,580	674	1,317
Total	8,670	6,061	890	1,719

Table 6: Train, validation, and test splits of BuDDIE for each type of question for VQA.

our dataset may contain incomplete annotations as we put a stronger preference on the precision of our annotations. We do not anticipate this to greatly impact the quality of our dataset given the high agreement score for KEE described in Section 3.5.

3.4 Visual Question Answering

Question answering is a common NLP task where a model must provide a natural language response to a question given a passage (Yang et al., 2015; Rajpurkar et al., 2016; Joshi et al., 2017; Yang et al., 2018). This naturally extends to images and evolves into VQA (Antol et al., 2015). Document VQA is a mixture of these two tasks in which questions require understanding of both text and visual properties of a document (Mathew et al., 2021).

We consider two types of questions in BuDDIE. Firstly, **span** questions are phrased as "What is the *X*?", where *X* is a key entity and the actual entity is the answer. Secondly, **boolean** questions are phrased as "Is the *X Y*?", where *X* is a key entity as before and *Y* is a candidate answer. These questions have **yes** or **no** answers and help assess a model's ability to verify assertions about the content of a document, which KEE annotations alone could not permit. Each key entity has an associated phrase to use in the question templates. For example, questions for the the entity `AGT_address_zipcode` are phrased as "What is the zip code of the registered agent?" (for span questions) and "Is the zip code of the registered agent 12345?" (for boolean questions).

For each key entity observed in a document, we generate a question with a 30% likelihood. For the questions generated, 70% are span questions and 30% are boolean questions. Span questions are generated by inserting the key entity phrase into the question template. The answer is given as a list of key entity annotations, as it is possible to observe multiple key entities of the same type in a document.[4] Each key entity annotation corresponds to a set of OCR tokens. As for boolean questions, we create a "Yes" question or a "No" question with equal probability. In the case of a "Yes" question, the candidate answer is any of the annotations in the document with the specified entity label. In the case of a "No" answer, we derive a candidate list from two sources. Firstly, we consider other entities from the *entire* dataset with the same fine-grained label (but not the same value). Secondly, we consider key entities *within the document* that share the same key entity detailed type but not the same super category. The candidate answer is chosen randomly from these two pools. The total number of occurrences of each question

[4] Past question answering datasets have allowed multiple spans to be a valid answer (Yang et al., 2018).

Model	Model Size	Doc. Class. F1 ↑	Key Entity Extraction Prec. ↑	Rec. ↑	F1 ↑	Visual Question Ans. Acc. ↑	ANLS ↑	F1 ↑
$\text{BERT}_{\text{base}}$	110 M	94.43	80.94	85.85	83.32	83.49	86.54	75.52
$\text{RoBERTa}_{\text{base}}$	125 M	91.96	84.49	87.48	85.96	84.28	85.64	90.06
$\text{LayoutLM}_{\text{base}}$	160 M	96.01	83.62	88.16	85.83	54.95	86.52	75.32
$\text{LayoutLMv3}_{\text{base}}$	133 M	88.48	84.23	88.86	86.49	84.90	86.85	89.32
GPT4	–	83.54	77.76	80.36	77.76	63.83	80.05	75.42
DocLLM	7 B	**99.15**	**90.55**	**89.97**	**89.97**	**92.45**	**89.58**	**93.79**

Table 7: Baseline results on DC, KEE, and VQA for BuDDIE. VQA accuracy considers Boolean questions while ANLS and F1 consider span questions. Note that GPT4 was run in a zero-shot setting while DocLLM had been instruction-tuned using BuDDIE along with other VRDU datasets.

type in the dataset is given in Table 6.

3.5 Annotation Quality

Using a sample of 60 documents from BuDDIE, we measure the agreement between the original annotators and new quality validators on each annotation task (DC and KEE). Following previous studies (Artstein and Poesio, 2008; Jochim et al., 2018), we sampled from a wide variety of annotations to mitigate some of the bias that could be caused by the sample size. We observe a Cohen's κ of 0.976 for document classification and 0.889 for key entity extraction. Note that since a validation task was performed as a post-processing step to obtain the final BuDDIE annotations (see "two rounds of annotation" in Sections 3.2 and 3.3), the agreement was thus computed by assessing the quality of a sample of already-refined final annotations. While our calculations may consequently provide an upper bound on Cohen's κ for the original *first round* annotations, they yield a representative estimate of the quality of our *final* annotations. Importantly, the data quality validators of the 60 sampled documents had not previously seen the documents they reviewed during this quality assessment exercise.

4 Experiments

4.1 Baseline Models

We consider six baseline models for our tasks. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2020) are text-only models that solely rely on the OCR token sequence. LayoutLM (Xu et al., 2020) integrates additional spatial features into the transformer, and merges the transformer output with a vision CNN. LayoutLMv3 (Huang et al., 2022) incorporates vision features into the transformer architecture for each token. For the aforementioned baselines, we finetune the base version of the model on each of the three tasks individually. We leverage the default hyperparameters of each respective model; a base learning rate of 10^{-4} was used with the Adam optimizer (Kingma and Ba, 2015), and a batch size of four was selected. All experiments were run with up to eight NVIDIA T4 GPUs. Smaller models used fewer GPUs.

In addition to the previous traditional baseline models, we further include two LLM baselines: GPT4-0613 and DocLLM (Wang et al., 2024). GPT4 is the text-only variant of the OpenAI model, to which we feed a document's OCR along with a prompt to represent the task at hand – following the templates used in Wang et al. (2024). Lastly, DocLLM-7B (based on Llama2-7B (Touvron et al., 2023)) is given the document's OCR along with spatial bounding box information and the task prompt. GPT4 is used in a zero-shot setting while DocLLM has been instruction-tuned on the training split of BuDDIE as well as other VRDU datasets. Full details regarding the training setup of DocLLM are described in the original manuscript (Wang et al., 2024). Due to cost and API usage constraints, we do not benchmark GPT4o on BuDDIE. In addition, the discrepancy between the OCR tokens on which our annotations rely and GPT4o's proprietary image processor could potentially skew the scores of KEE and VQA token-level metrics.

4.2 Evaluation Metrics

We assess model performance on the three VRDU tasks of BuDDIE with different metrics. As our document classes are imbalanced in the dataset (see Table 2), we report a macro F1 score for DC. In other words, we take the mean F1 score across the five document classes. For KEE, we report the weighted average token-level recall, precision, and F1 scores. We also measure VQA performance using several metrics. We evaluate boolean question performance using accuracy, and span questions using the Average Normalized Levenshtein Simi-

larity (ANLS) and F1 scores. The ANLS metric is a character-level metric used in Mathew et al. (2021) whereas the F1 score gives the traditional token-level score. These two metrics are reported separately to capture different aspects of measurement and granularity.

4.3 Results

Table 7 reports the performance of our baselines on BuDDIE.[5] We note that the performance reported for GPT4 and DocLLM slightly differ from those in Wang et al. (2024). This is because the manuscript used accuracy rather than F1 for DC, included additional prompts for KEE that do not enable a fair comparison with non-LLM models, and aggregated results for VQA between span and boolean questions, which we separate in this paper.

With regards to DC, we observe strong performance from all models. This was expected as certain keywords can be highly characteristic of specific document categories. Furthermore, the imbalanced class distribution may further inflate performance even though we use macro F1. We plan to add more fine-grained document classes in future versions of BuDDIE as well as more documents to help alleviate the class imbalance.

For KEE, we observe that the spatially aware models (LayoutLM, LayoutLMv3, and DocLLM) tend to have a much better recall than their text-only counterparts. While GPT4 demonstrates the worst result, the spatially-aware LLM, DocLLM, outperforms any of the dedicated smaller models. Note that GPT4's scores are still considerably resilient given the zero-shot setting, as opposed to the fine-tuning setting used for the other models.

The VQA F1 scores exhibit high variability in the reported results. This can be attributed to the inherent fluctuation in token-level evaluation compared to the character-based ANLS metric. Specifically, we observe a large discrepancy between the VQA F1 scores of BERT, LayoutLM, and GPT4 with respect to the other models. We hypothesize that the performance of the first two is due to a difference in tokenizers used. Specifically, LayoutLM and BERT employ a word-piece tokenizer, whereas the other models employ a Byte-Pair Encoding (BPE) tokenizer. The BPE tokenizer is likely to capture tokens with greater accuracy, consequently leading to improved F1 scores. It is probable again that GPT4's relatively low performance across all VQA metrics can be attested to both its lack of input layout information and to the zero-shot inference setting (the model sometimes extracts less or more context than expected in the annotations). DocLLM once again outperforms the other models on VQA, specifically in terms of the boolean question accuracy.

[5]The experiments include 75 Colorado and Michigan documents, which will be omitted from the public version of BuDDIE due to distribution licenses.

5 Conclusion

In this paper, we introduced a new VRDU dataset for the finance domain, BuDDIE, consisting of 1,665 annotated documents. BuDDIE is unique in its varied document styles, sizes, and annotations for three distinct tasks. We use a variety of language models, multi-modal language models, and LLMs to provide comprehensive baselines for our dataset. While we note DocLLM's impressive performance across the tasks, VRDU model performance is still not comparable to human performance on the tasks of KEE and VQA as of the date of publication(Mathew et al., 2021), and zero-shot prompted LLMs still have room for improvement. We hope that our dataset can be a valuable resource to encourage the research community to seek more robust VRDU models that help on processes such as KYC, and will spur further research in this domain. Future work on BuDDIE will include multi-page annotations, multi-turn VQA, and instruction tuning benchmarks.

Disclaimer

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 973–983. IEEE.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Lukasz Borchmann, Michal Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michal Turski, Karolina Szyndler, and Filip Gralinski. 2021. DUE: end-to-end document understanding benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Brian L. Davis, Bryan S. Morse, Scott Cohen, Brian L. Price, and Chris Tensmeyer. 2019. Deep visual template-free form parsing. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 134–141. IEEE.

Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, and Curtis Wigington. 2021. Visual FUDGE: Form understanding via dynamic graph editing. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 416–431. Springer.

Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, Curtis Wigington, and Vlad I. Morariu. 2022. End-to-end document recognition and understanding with dessurt. In *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13804 of *Lecture Notes in Computer Science*, pages 280–296. Springer.

Timo I. Denk and Christian Reisswig. 2019. BERTgrid: Contextualized embedding for 2D document representation and understanding. *CoRR*, abs/1909.04948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. LAMBERT: Layout-aware language modeling for information extraction. In *Document Analysis and Recognition – ICDAR 2021*, pages 532–547, Cham. Springer International Publishing.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 991–995. IEEE Computer Society.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10767–10775. AAAI Press.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, 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 *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019*, pages 1–6. IEEE.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jordy Van Landeghem, Rubèn Tito, Lukasz Borchmann, Michal Pietruszka, Pawel Józiak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew B. Blaschko, Sien Moens, and Tomasz Stanislawek. 2023. Document understanding dataset and evaluation (DUDE). *CoRR*, abs/2305.08455.

Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolay Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. 2023. FormNetV2: Multimodal graph contrastive learning for form document information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9011–9026, Toronto, Canada. Association for Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural pre-training for form understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6309–6318, Online. Association for Computational Linguistics.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. ViBERTgrid: A jointly trained multi-modal 2D document representation for key information extraction from documents. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 548–563. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. PAWLS: PDF annotation with labels and structure. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 258–264, Online. Association for Computational Linguistics.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Yuhui Cao, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-Layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3744–3756, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going full-TILT boogie on document understanding with text-image-layout transformer. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 732–747. Springer.

Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. 2022. DuReader$_{\text{vis}}$: A Chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, Dublin, Ireland. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stepán Simsa, Milan Sulc, Michal Uricár, Yash Patel, Ahmed Hamdi, Matej Kocián, Matyás Skalický, Jirí Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. DocILE benchmark for document information localization and extraction. *CoRR*, abs/2302.05658.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *Document Analysis and Recognition – ICDAR 2021*, pages 564–579, Cham. Springer International Publishing.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19254–19264. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hieu M. Vu and Diep Thi-Ngoc Nguyen. 2020. Revising FUNSD dataset for key-value detection in document images. *CoRR*, abs/2010.05322.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. 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)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020. DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 898–908, Online. Association for Computational Linguistics.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. VRDU: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5184–5193, New York, NY, USA. Association for Computing Machinery.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *Preprint*, arXiv:2307.02499.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. TRIE: End-to-end text reading and information extraction for document understanding. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1413–1422. ACM.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022. Multimodal pre-training based on graph attention network for document understanding. *CoRR*, abs/2203.13530.

Ran Zmigrod, Zhiqiang Ma, Armineh Nourbakhsh, and Sameena Shah. 2024. TreeForm: End-to-end annotation and evaluation for form document parsing. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 1–11, St. Julians, Malta. Association for Computational Linguistics.

A BuDDIE Annotation Instructions

In this section, we provide a more detailed description of the instructions received by annotators for the DC and KEE annotation tasks. For both tasks, annotators first annotated their assigned documents using the instructions provided below. Then, a validator was assigned to check these annotations using the same instructions. Any major disagreements that the validator and annotator were not able to resolve with the help of a third annotator were discarded.

A.1 Document Classification

Annotators were instructed to pick a document class using these ordered instructions.

1. If the document title contains the word "detail", "business", "entity", or "search", classify the document as *Business Entity Details*.
2. If the document title contains the word "annual", "biennial", "periodic", etc., or contains a year (e.g., 2007), classify the document as *Periodic Report*.
3. If the document title contains the word "amend", "update", or "change", classify the document as *Amendment Document*.
4. If the document title contains the word "application", "article", or "reservation', classify the document as *Article or Application*.
5. If the document title contains the word "certificate", "statement", "affidavit", "report", "confirmation", "notice", or "receipt", classify the document as *Certificate or Statement*. Note that an "Application for a Certificate" should be classified as *Article or Application* by the previous instruction.
6. If there is no title, examine the format and content; if it seems descriptive of a business, classify the document as *Business Entity Details*.
7. If none of the above rules hold, do not label this document.

A.2 Key Entity Extraction

For the KEE task, annotators utilised an annotation tool that allowed them to create labelled bounding boxes where the labels available are given in Table 4 (an additional `is_key` label was annotated but not included in this version of BuDDIE). Annotators were asked to abide by the following annotation instructions.

1. For each meaningful value in the document, check whether the value relates to any of the super categories (given in Table 3). If no super category is identified but you are sure this is a meaningful value, select the `OTHER` category. Please see below for examples for some of the super categories.
    - Business Entity (`ENT`): Corporation, business, trade, etc.
    - Government Official (`GO`): State secretary, mayor, etc.
    - Key Personal (`KP`): Director, vice president, treasurer, etc.
2. Select from the fine-grained labels (given in Table 4) of the category the appropriate label for the value. If the value does not have an appropriate label, omit the annotation.
3. Create a bounding box around the value tokens. This will select all OCR tokens that are in or lay on the bounding. If this selection is not accurate, you may also turn off the OCR selection tool and draw a free form bounding box. *Note: For this version of the dataset, we only include bounding boxes that use the OCR selection tool.*
4. If the value has an associated key, select the `is_key` label and create a bounding box as in the previous step. *Note: For this version of the dataset, we did not include the* `is_key` *entities.*
5. Only create an annotation if you are sure that the value is meaningful and you have chosen the correct label.

All annotators first annotated ten practice documents for which they received feedback before they began annotating the dataset documents.