Towards a new research agenda for multimodal enterprise document understanding: What are we missing?

Armineh Nourbakhsh
Language Technologies Institute,
Carnegie Mellon University
J.P. Morgan
anourbak@cs.cmu.edu

Sameena Shah
J.P. Morgan
sameena.shah@jpmorgan.com

Carolyn Rosé
Language Technologies Institute,
Carnegie Mellon University
cprose@cs.cmu.edu

Abstract

The field of multimodal document understanding has produced a suite of models that have achieved stellar performance across several tasks, even coming close to human performance on certain benchmarks. Nevertheless, the application of these models to real-world enterprise datasets remains constrained by a number of limitations. In this position paper, we discuss these limitations in the context of three key aspects of research: dataset curation, model development, and evaluation on downstream tasks. By analyzing 14 datasets and 7 SotA models, we identify major gaps in their utility in the context of a real-world scenario. We demonstrate how each limitation impedes the widespread use of SotA models in enterprise settings, and present a set of research challenges that are motivated by these limitations. Lastly, we propose a research agenda that is aimed at driving the field towards higher impact in enterprise applications.

1 Introduction

Multimodal document understanding, also known as Visually rich Document Understanding (VrDU), and the tasks that it encompasses—including visual information extraction, visual question answering, and image document classification—constitute a major operational bottleneck in enterprise settings (Paycom, 2021; McKinsey, 2022). Due to their rich structure, length, domain-specific language, and spatio-visual complexity, enterprise documents such as reports, memos, invoices, forms, and contracts, are often processed using human supervision. This manual process is cumbersome and therefore error prone, so much so that some institutions are compelled to adopt a dual review protocol for the task of information extraction and data entry (CBS). This, coupled with the large volume of documents in many enterprise settings[1] has led to a substantial and growing demand for Document Intelligence services (ibml, 2024; Insights, 2024).

Against this backdrop, adoption of multimodal document understanding models remains constrained by certain limitations in SotA models and benchmarks. In this position paper, we discuss these limitation in the context of three main components of research: datasets, models, and evaluation strategies. We will contextualize our discussion with a common real-world example of a Document AI task in enterprise settings.

1.1 Real-world task example: authorized signatory identification

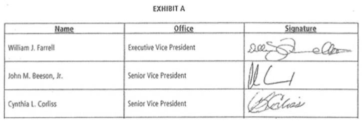
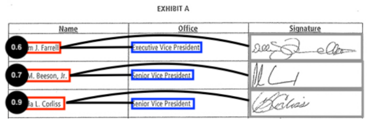
To demonstrate the challenges of operationalizing SotA VrDU models in enterprise settings, throughout this paper we will use a scenario that is inspired by a common real-world task. Suppose Alice is a knowledge worker at a financial institution who is in charge of reviewing client documents. Each institutional client provides a document with a list of authorized signatories, their titles, contact information, and signature samples. This list is later used to verify whether legally-binding agreements have been signed by authorized stakeholders. As part of a remediation project, Alice is tasked with reviewing 1,000 authorized signatory forms, extracting key information, and keying them into a new database.

To make the task more manageable, Alice would like to reduce her manual workload by 70%, either by reviewing only 30% of the documents, or by reviewing only 30% of each document's contents, and delegating the remainder of the work to an automated solution. This would require the automated solution to perform the following tasks:

1. Process each authorized signatory document and extract the following information: names,

[1] As an example, in 2024 J.P. Morgan Chase served nearly 80 consumers (Center, 2024). Considering each customer's records, filings, tax forms, identification documents, and other disclosures, the volume of documentation in the retail banking business alone could scale to hundreds of millions.

Findings of the Association for Computational Linguistics ACL 2024, pages 14610–14622
August 11-16, 2024

		**Ideal output:** Each signatory, title, and signature box is extracted and tagged within the document [red, blue, and gray boxes]. Each signatory is used as an anchor to group the metadata [black links]. Each entity (or grouping) is assigned a confidence score [black circles].

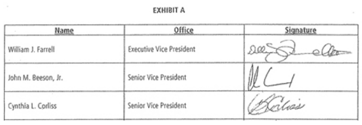
	`William J. Farrel: NAME0`
	`Executive Vice President: TITLE0`
	`bbox{30, 10, 50, 25}: SIG0`
	`[NAME0, TITLE0]: LINK0`
	`John M. Beeson, Jr.: NAME1`
	`Senior Vice President: TITLE1`
	`bbox{30, 30, 50, 45}: SIG1`
	`[NAME1, TITLE1]: LINK1`
	...
		Common output:
		Extractions may not be associated with bounding boxes.
		Different solutions may be needed for text vs. handwriting.
		Links may be unavailable or only partially available.
		Confidence scores are often not available (or log-probs are uncalibrated).

Table 1: Top row: The expected output of a VrDU model when processing an authorized signatory document. Bottom row: The output often generated by SotA approaches. Note that due to the confidentiality of authorized signatory forms, we have used a public example from a credit agreement (CS, 2013).

metadata (such as titles, addresses, and contact information), and signature samples.

2. Associate all attributes related to the same entity. For example the name, contact information, and signature sample of each individual should be grouped together.

3. Map the information about each entity into a schema that the new database recognizes.

4. Create a detailed trace of where each piece of information was extracted from, so that auditors can verify that proper protocol was followed.

The top row of Table 1 illustrates the output that Alice expects from an automated system. Ideally, each entity is to be tagged within the document, so that the extracted entity can be mapped to a bounding box within the page. The model also needs to be able to group each signatory's name and corresponding title and signature. Lastly, each extraction (or ideally, each grouping, each page, or each document) needs to have a confidence score that Alice can use to decide whether she needs to review the model's output for accuracy.

Note that the above requirements are not limited to our particular scenario, and generalize to most information extraction tasks in enterprise settings. Some requirements (such as traceability of output) extend to tasks beyond information extraction such as question answering over documents and summarization.

In the remainder of this paper, we demonstrate where and how current SotA approaches fall short of the above-mentioned requirements. In order to ground our arguments in current research, we use a set of SotA VrDU models and datasets as the backdrop to our analysis.

2 SotA models and datasets

VrDU spans a suite of tasks, which, together comprise a complete picture of a given document. These include Document Classification (CLS), Page Segmentation (SEG), Key Entity Extraction (KIE), Tabular Reasoning (TR), and Visual Question Answering (VQA). Note that we exclude the Optical Character Recognition (OCR) task, given the increasing availability of open-source and commercial solutions that can process enterprise documents.[2]

SotA VrDU models fall into three categories based on their core architecture: multimodal transformer-based models (TR), graph-based models (GR), and multimodal large language models (MLLM). Table 2 lists the SotA models from each category, along with information about their licensing, availability of code and pre-trained weights, and model size. We have defined SotA as top within-category performance across a variety of VrDU tasks, including CLS, SEG, TR, KIE, and

[2]Open-source solutions include TrOCR (Li et al., 2023), PaddleOCR, Surya, and Tesseract-ocr, and commercial solutions include ABBYY, IBM DataCap, Google Cloud Vision, Amazon Textract, and Microsoft Azure OCR services, among others.

VQA.[3] We exclude the family of MLLMs known as generalist models, such as GPT-4V (OpenAI, 2023). This is because these models have not been systematically benchmarked on enterprise VrDU tasks, and can occasionally struggle with modeling a given task's context.[4]

Table 3 lists the datasets that the SotA models from Table 2 have used for pre-training, fine-tuning, or instruction-tuning. Since our focus is on enterprise applications, we have excluded datasets that cover web-based documents (e.g VisualMRC (Tanaka et al., 2021), TabFact (Chen et al., 2019), WTQ (Pasupat and Liang, 2015), InfographicVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022)), research publications (e.g. PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020)), or other non-enterprise collections (e.g. OCR-VQA(Mishra et al., 2019) and AI2D (Kembhavi et al., 2016)).

In the following sections, we will discuss the limitations of these datasets, models, and evaluation methods used within the VrDU literature, and how they impact adoption in the enterprise domain. We will end each section with one or more **Research Challenges (RCs)** that arise from the discussion, opening up new opportunities for VrDU researchers and practitioners to engage in high-impact research.

3 Data limitations

3.1 Limited publishers

Image documents, especially in the enterprise domain, are often owned by specific legal entities and are therefore not always available for redistribution, even when they do not include confidential content. This makes open-domain collections of public documents difficult to source compared to unimodal corpora such as Wikipedia, Common Crawl, and social media posts. As a result, research in the domain of multimodal document understanding is dominated by a limited set of corpora, as demonstrated in Table 3.

Unsurprisingly, due to licensing restrictions, the US government is the leading supplier of VrDU datasets.[5] Regulatory disclosure protocols and legal settlements have resulted in collections such as IIT-CDIP (Lewis et al., 2006), Kleister (Stanisławek et al., 2021), and DeepForm (Svetlichnaya, 2020), which cover documents from regulated industries such as tobacco manufacturing and political advertising. This poses two challenges to VrDU researchers and practitioners. First, these public collections are not always representative of the variety of enterprise documents. As an example, the authorized signatory forms used in our scenario would not be represented in any of the datasets due to their confidentiality. Second, permissive licenses such as Fair Use do not necessarily permit the use of these datasets in commercial applications. In our example, even though Alice does not intend to use the VrDU technology for commercial purposes, the fact that she is employed by a for-profit entity may subject her to more restrictive terms (Gordon-Murnane, 2010). These limitations motivate the following Research Challenge:

RC1: The licensing challenges mentioned above have led to the overuse of few datasets from narrow domains, which can lead to poor OOD performance. Can these datasets be combined, augmented, synthesized, or diversified to create better benchmarks for generalizability and robustness of VrDU models? Can synthetic datasets, designed to prevent the leakage of proprietary or confidential information, help expose SotA models to enterprise corpora?

3.2 Under-representation of associative tasks

Page Segmentation (SEG) and Key Information Extraction (KIE) are two of the core VrDU tasks that are both extractive, but both have complementary associative counterparts, i.e. Relation Extraction (RE). Models that perform SEG can identify various components on a page (e.g. heading, paragraph, figure, etc.), but they do not identify the relationship between components. That would require a mapping of the full hierarchy of the document's structure. In a similar vein, models that perform KIE can identify entities within a page, but to identify the relationship between entities, the models need to perform the additional task of RE.

In many real-world settings, KIE and RE need to be performed in tandem to enable end-to-end automation. In our scenario, a model that is solely trained on KIE would be able to identify each authorized signatory, phone number, and address, but would not be able to group them together or map

[3] Note that GR models often focus on the KIE task alone, and are therefore represented by one model that currently reports top performance for that task, i.e. FormNetV2 (Lee et al., 2023).

[4] See Appendix A for an example of how GPT-4V can struggle with interpreting a VrDU task in the enterprise setting.

[5] See Appendix B for an illustration of the lineage of the datasets listed in Table 3.

Model	**Citation**	**# Params**	**Architecture**	**License**	**Commercial Affiliate**	**OSS Status**	**Generative/ Grounded**	**VrDU tasks**
LayoutLMv3$_{\text{LARGE}}$	(Huang et al., 2022)	368M	TR	CC BY-NC-SA 4.0	Microsoft	PW FW FC	N/Y	CLS SEG KIE VQA
UDOP	(Tang et al., 2023)	794M	TR	MIT	Microsoft	PW FC	Y/N	CLS KIE TR VQA
FormNetV2	(Lee et al., 2023)	204M	GR	N/A	Google	None	N/Y	KIE
UReader	(Ye et al., 2023a)	86M*	MLLM	Apache 2.0	Alibaba	PW PC	Y/N	KIE TR VQA
DocLLM	(Wang et al., 2023a)	1B, 7B	MLLM	N/A	JP Morgan	None	Y/N	CLS KIE TR VQA
Qwen-VL-MAX	(Bai et al., 2023)	Unknown	MLLM	Custom	Alibaba	FC	Y/N	VQA
SMoLA-PALI-X	(Wu et al., 2023)	Unknown	MLLM	N/A	Google	None	Y/N	VQA

Table 2: Models with SotA performance on a variety of VrDU tasks as of Jan 31, 2024. Architecture legend: `TR`: Transormer-based. `GR`: Graph-based. `MLLM`: Multi-modal LLM. OSS Status legend: `PC`: Pre-training code. `PW`: Pre-trained weights. `FC`: Fine-tuning code. `FW`: Fine-tuning weights. VrDU task legend: `CLS`: Document classification, `SEG`: Page segmentation, `KIE`: Key information extraction, `TR`: Tabular reasoning, `VQA`: Visual question answering. *UReader reports its number of trainable parameters, but the model is created by applying LoRA (Hu et al., 2022) to mPLUG-Owl (Ye et al., 2023b), which has around 7B parameters.

Dataset	**Citation**	**Training size**	**License**	**Upstream publisher**	**VrDU tasks**
IIT-CDIP	(Lewis et al., 2006)	6,910,192 docs	Fair Use	US Gov.	None
RVL-CDIP	(Harley et al., 2015)	400,000 pages	Fair Use	US Gov.	CLS
DocLayNey	(Pfitzmann et al., 2022)	80,863 pages	CDLA-Permissive	Unknown/varied	SEG
DocILE	(Šimsa et al., 2023)	106,680 docs 108,715 pages	Fair Use	US Gov.	KIE LIR
DocVQA	(Mathew et al., 2021)	12,767 pages	Fair Use	US Gov.	VQA
DUDE	(Van Landeghem et al., 2023)	5,019 docs	Unspecified	Unknown/varied	VQA
BuDDIE	(Zmigrod et al., 2024b)	1,665 pages	Proprietary	US State Govs.	KIE
FUNSD	(Jaume et al., 2019)	199 pages	Custom	US Gov.	KIE RE
CORD	(Park et al., 2019)	2,000 pages	CC-BY-4.0	Businesses	KIE
SROIE	(Huang et al., 2019)	1,000 pages	CCA 4.0	Businesses	KIE
DeepForm	(Svetlichnaya, 2020)	~20,000 docs	MIT	US Gov.	KIE
Kleister	(Stanisławek et al., 2021)	3,318 docs 64,872 pages	OGL	US & UK Govs.	KIE
VRDU	(Wang et al., 2023b)	2,556 docs	Fair Use	US Gov.	KIE
Payment	(Lee et al., 2023)	~10,000 docs	Proprietary	Google	KIE

Table 3: 14 popular datasets in the VrDU literature. The number of used by the models in Table 2. VrDU task legend: `CLS`: Document classification. `SEG`: Page segmentation. `KIE`: Key information extraction. `VQA`: Visual question answering. `LIR`: Line item recognition. Note: The table excludes `OCR` datasets as well as those focused on historical document understanding.

them to the relational schema of a database.

Despite their relevance to real-world applications, associative tasks are often ignored in VrDU datasets, possibly due to the high cost of annotating documents for multiple tasks. Of the 14 datasets listed in Table 3, 9 cover the task of Key Information Extraction (KIE). Of these, only 1 combines this task with Relation Extraction (RE).[6] This has led to an under-representation of RE in research publications, e.g. none of the SotA models listed in Table 2 report their performance on RE. This motivates the following Research Challenge:

RC2: How well do current SotA models perform on the RE task? Can relational architectures such as graph-based models improve SotA performance on the RE task?

3.3 Limited grounding

As was noted in our scenario, real-world enterprise applications of VrDU models often require the model to ground its output within the input document by mapping each token to its location on the page, establishing a clear evidentiary trace for possible audits. While most KIE datasets are annotated to support such grounding, VQA datasets usually lack the grounding annotations. Given that VQA datasets often include a mix of extractive and abstractive questions, absence of grounding poses two fundamental challenges to end users:

First, lack of grounding poses an immediate challenge to the verification and evaluation of extractive models. The top row of Table 4 illustrates this with an example. Given an authorized signatory form and the question "What is the title of Cynthia L. Carliss?" two hypothetical models are shown to provide ungrounded answers. The models are evaluated using Average Normalized Levenshtein Similarity, popularized by DocVQA (Mathew et al., 2021). Model A produces the correct output with a perfect score, but without any grounding information, it is unclear whether the output refers to the proper bounding box (blue) or is based on the incorrect context (red). Model B produces an incorrect response, referring to the title of another signatory. Nevertheless, the ANLS metric is calculated at 0.625 due to a partial match with the gold answer.

In 8.5% of the training samples in DocVQA, there are two or more instances of the gold answer within the input page, making it difficult to properly contextualize the answers in a post-processing step. Partial matches only complicate this problem further.

The second challenge arises from the lack of grounding for abstractive questions, despite the requirement in many enterprise settings that every abstractive decision needs to be explicitly evidenced. Consider the second row of Table 4 that shows an example of an abstractive Yes/No question. To determine whether "Cynthia L. Carliss" is a senior executive, a model would need to follow a reasoning path, first locating her title on the page, and then mapping it to a collection of possible roles that qualify as senior executive titles.[7] In the absence of any grounding or explanation, it would be unclear whether a model is providing the correct answer ("No") or the incorrect answer ("Yes") based on a simple match with the keywords "Senior" and "Executive", respectively. While grounded reasoning datasets exist in the unimodal literature (Chen et al., 2021; Zhang et al., 2021) and in open-domain VQA (Zhu et al., 2016; Pont-Tuset et al., 2020), such datasets are yet to be popularized in multimodal document understanding. This gives rise to the following Research Challenge:

RC3: Having explicitly-encoded knowledge in the form of KBs or taxonomies may enable researchers to design models that can perform grounded abstractive QA by linking the evidence between the documents and the external knowledge. Can VrDU literature contribute to grounded question answering by generating deeply annotated and knowledge-augmented datasets for extractive as well as abstractive questions? Can these datasets enhance the grounding and explainability of VrDU models?

4 Model limitations

4.1 Calibration

Let's once again consider our scenario. As stated in Section 1.1, Alice would like to reduce her workload by 70%. Let's suppose that she is able to find a SotA model that has an F1 score of 0.99 across all KIE benchmarks. For simplicity, we will assume that this means the model makes one mistake per 100 extractions. If Alice applies the model to the signatory forms, there will likely be errors, given

[6]Some of the datasets such as CORD, DocILE, and BuDDIE include shallow relation annotations, but they do not include complete hierarchical relations or key-value pairings.

[7]In many enterprise settings, such knowledge bases and taxonomies are available as part of training material, policy documents, business rulesets, or structured databases.

Document	Question: `What is the title of Cynthia L. Carliss?`	
Name — Office William J. Farrell — Executive Vice President John M. Beeson, Jr. — Senior Vice President Cynthia L. Corliss — Senior Vice President	**Model A** Answer: `Senior Vice President` ANLS: 1.0	**Model B** Answer: `Executive Vice President` ANLS: 0.625
Document	Question: `Is Cynthia L. Carliss a senior executive?`	
Name — Office William J. Farrell — Executive Vice President John M. Beeson, Jr. — Senior Vice President Cynthia L. Corliss — Senior Vice President	**Model A** Answer: `No` ANLS: 1.0	**Model B** Answer: `Yes` ANLS: 0.0

Table 4: An example illustrating how lack of grounding can lead to misleading assessments of a model's performance. Top-row: Extractive QA. Bottom row: Abstractive QA. The image is excerpted from (CS, 2013).

that 1,000 forms are likely to include more than 100 signatories. If Alice is not able to locate the possible errors, she will have to review every one of the model's extractions to verify its accuracy. Assuming that Alice can perform the verification task faster than the extraction task, we will estimate her time-saving as 50%.[8] This will still not meet her target of 70% of documents (or 70% of fields) being processed in a "straight-through" fashion without a manual touchpoint. In order for Alice to reach her target, she would need a model that is well calibrated, and can indicate which documents or which contexts are likely to include to errors.

Despite the recent attention that calibration research has attracted with regards to the detection of hallucinations in LLM outputs, the VrDU literature has largely remained focused on performance without much regard for calibration. This topic is especially worthy of attention in the VrDU literature, given the fact that most SotA models (6 out of 7 models in Table 2) do not follow a generative objective with a causal decoder, resulting in output probabilities that are essentially not well calibrated.

RC4: How can calibration methods proposed in the unimodal literature be adapted to VrDU models? What can the VrDU field offer to calibration research with regards to multimodal semantics?

[8]If Alice is following a dual review process (i.e. a Maker-Checker process), then the 50% time-saving estimate is consistent with removing the Maker from the process, allowing Alice to act as the Checker. Having said that, the estimate is still likely to be very generous, because ungrounded models do not contextualize their extractions, and Alice would need to manually locate each extracted output in the original document before verifying it.

4.2 Model licensing and availability

As discussed in Section 3.1, SotA models inherit the licensing challenges of the datasets that they have been trained on. Additionally, the models carry their own Intellectual Property, which might restrict their use.

Moreover, the unavailability of open-sourced code and/or model weights might further limit usage in downstream tasks. As Table 2 shows, only 4 of the 7 SotA models on the list have any open-sourced components, and no model has all three components that are required for it to be considered fully open-source (i.e. pre-training code, pre-trained weights, fine-tuning code).

RC5: Can the field incentivize enterprise stakeholders to participate in open research by investing in methodologies that address their needs?

4.3 Grounding and generation

The recent success of large generative models has prompted researchers to explore Multimodal Large Language Models (MLLMs) as a solution to the VrDU problem. This has led to the emergence of a suite of generative MLLMs with promising performance across several VrDU tasks, including VQA, KIE, and CLS. A key challenge of these generative models is the fact that they are not designed to ground their responses within the input, further complicating the grounding challenge mentioned in Section 3.3.

RC6: How can generative architectures be reconciled with grounding requirements?

4.4 Field-level vs. document-level performance

The benchmarks in the VrDU field often calculate a model's performance as an average over all fields within the dataset, whereas in real world settings, performance is often measured at the document level. In our scenario, Alice's goal is to reduce her workload by 70%. This is achievable by: 1) Using a model that processes 70% or more of documents without an error, or 2) Using a model that has an average per-document performance of 70% or more. Neither of these metrics are commonly calculated as part of standard benchmarks. As an example, Table 5 shows the performance of LayoutLMv3 when measured overall, compared to context-specific measurements such as doc-level accuracy and average F1 per document. Doc-level accuracy shows the percentage of documents that can be processed in a "straight-through" fashion, i.e. ones which the model processes without any errors. As the table shows, only 4% of documents are processed by the model without any errors, falling far behind Alice's target of 70%. On the other hand, LayoutLMv3's average F1 per document is 83.08. This means that Alice can focus her effort on reviewing the portions of each document that the model is likely to mishandle. She can strategize by analyzing the model's performance per entity type, and focus her efforts on the "Header" category for which the model performs poorly compared to other entity types. Alternatively, given a model that produces well-calibrated probabilities, Alice can focus her efforts on low-confidence outputs. This gives rise to the following research opportunity:

RC7: Studies such as Zmigrod et al. (2024a) have proposed evaluation metrics that take into account the semantic structure of documents, and combine extractive and associative performance. What additional metrics should be used to assess the performance of VrDU models on key tasks?

Reported F1 (with segments)		92.08
True F1 (no segment info)		82.86*
F1 per entity type	**Header**	57.49
	Question	86.03
	Answer	83.25
Doc-level accuracy		4%
Avg F1 per doc		83.08

Table 5: Evaluation of LayoutLMv3 performance as reported in Huang et al. (2022) (Reported F1) versus when using contextualized metrics (e.g. Doc-level F1). *The "True F1" value of 82.86 is consistent with the value reported in Lee et al. (2023), i.e. 82.53.

4.5 The problem of reading order

VrDU datasets that cover the KIE task are often evaluated using a standard IOB schema. Originally developed for the IE task in unimodal text, the IOB schema honors a sequence order that is inherent to unimodal text. In contrast, multimodal documents are 2-dimensional artifacts, and a canonical ordering of the words might not be readily available due to their complex structure.

Nevertheless, in order to support the IOB schema, many VrDU datasets provide a proposed ordering as part of their annotations, which the models in turn use during evaluation. This leads to a fundamental problem of information leakage—the datasets are providing information to the model regarding the order of the words which would not be available in real-world test settings.

Additionally, as pointed out by Lee et al. (2023), some models assume that they have access to segment-level bounding boxes at test time. They demonstrated how the performance of LayoutLMv3 on the FUNSD dataset would decrease by almost 10 points when access to segment-level information wasn't provided (see second row of Table 5).

Recent studies such as Zhang et al. (2023) have taken steps towards addressing this problem by proposing graph-based models that are sensitive to reading order. The authors have also released revised versions of the FUNSD and CORD datasets to address the problem of information leakage. Further investigation in this domain will enhance real-world outcomes for downstream users.

RC8: What other measures can be employed to ensure common benchmarks are protected against information leakage?

5 A utility-focused research agenda

Given the challenges laid out in previous sections, we will now discuss the opportunities that they offer to researchers and practitioners in the field. We will present these opportunities in the context of several high-level focus areas, each covering one or more of the research challenges discussed in previous sections. Consistent with previous sections, we will present the focus areas in the context of the main components of research: datasets, models, and evaluation methods.

5.1 Datasets: Curation

The inherent issue of copyright and ownership that limits the use of document collections in training VrDU models is compounded by the confidentiality of content in most enterprise settings. Recent work that has focused on synthetic document generation explores the possibility of creating realistic layouts (Raman et al., 2022), content (Hiebel et al., 2023), or both (Babkin et al., 2023). A challenge in using synthetic documents for VrDU tasks is that many of them are modeled after the same public-domain datasets mentioned in Section 3.1, and are thus likely to carry the same biases in their multimodal signal. To tackle this problem, a two-pronged approach is needed: 1) Enterprise researchers and practitioners can take on a leading role in releasing synthesized collections that reflect their proprietary documents with high fidelity, without violating their confidentiality. 2) Public-domain collections such as the IIT-CDIP dataset can be augmented with a larger variety of enterprise collections from a wider range of industries and time spans. As Figure 2 illustrates, the datasets that are derived from a variety of upstream sources are rare, and curators often focus on one or two specific publishers.

5.2 Datasets & Models: Grounding

As previously noted, grounding is a crucial requirement for downstream applications, and is often missing in both datasets and models. For extractive tasks such as KIE and extractive VQA, a basic level of spatio-visual grounding can be achieved by tying each token to its corresponding bounding box within the page. When annotating datasets, we encourage researchers to use tools that support visual annotations, such as PAWLS (Neumann et al., 2021). Developing models that provide bounding-boxes as part of their output further enables downstream users to verify them. Lastly, evaluation strategies that do not consider the placement of each answer within the original document risk overestimating performance. As discussed in Section 3.3, metrics such as ANLS can overestimate performance in extractive settings, and fail to capture semantic nuances in abstractive settings.

For abstractive tasks such as CLS and abstractive VQA, grounding can be a more challenging task. Nevertheless, borrowing from the rich body of research focused on verification and evidence retrieval can open new opportunities for VrDU researchers in this space.

5.3 Models: A new focus on calibration

We encourage the development of new benchmarks that probe the model's output or its internal representations from a calibration perspective. In the unimodal literature, previous studies have often done so by presenting evidence that the model generates compositional representations (Ettinger et al., 2018), that the model's confidence aligns with its performance (Jiang et al., 2021), or that the model makes "forgivable" errors (Renduchintala and Williams, 2022). Similar measures can be used by VrDU researchers to demonstrate whether the models produce well-calibrated outputs. Below are some research questions that can probe the issue of calibration in the VrDU domain:

- Is the model able to map documents to a compositional semantic space, where similar documents (in terms of content) are grouped together? How about similar documents in terms of visual style? In terms of layout? In terms of category (e.g. forms versus contracts)? Or in terms of issuer/producer?
- Does the model produce well calibrated probability distributions as part of its output? If not, does it lend itself to a post-hoc calibration approach such as Jiang et al. (2021)?
- Does the model's performance linearly scale with its confidence?
- Can the model's errors be identified at test time using contextual signals? For example, does the model consistently do well on tables but poorly on diagrams?
- How can the model be integrated into an operational pipeline where human oversight can be directed towards high-risk samples?

5.4 Evaluation: Contextualizing the performance

Lastly, we encourage researchers to report performance metrics not only at the dataset-level, but also at the document (and possibly output-type) levels. This can be used as an estimate of the amount of time that a knowledge worker can save by using the model in an operational pipeline, provided that the model is well calibrated.

Performance can also be profiled based on the category, type, visual complexity, and contents of the input. As an example, the DocVQA benchmark provides a breakdown of the different classes of questions and the type of reasoning that is required to answer them (e.g. reasoning over tables, charts, layout, or text). This can facilitate a more purposeful analysis of each model's performance compared to a singular measurement.

6 Conclusion

In this position paper, we provided a detailed breakdown of the challenges of operationalizing current VrDU models in enterprise applications. We argued that these challenges can be embedded within three key components of research: datasets, models, and evaluation strategies. Using a fictional scenario that was inspired by real-world use cases, we contextualized each challenge by demonstrating how it could hinder end-user adoption. Lastly, we proposed a set of research questions that can be investigated to address each challenge and facilitate the adoption of SotA models in enterprise settings. We hope that the agenda put forward in this paper can inspire new directions in the VrDU literature that accommodate downstream applications within operational pipelines.

7 Limitations

In Sections 1 and 2, we defined the scope of this position paper around major challenges in operationalizing VrDU models in enterprise settings from a task-centric perspective, hence excluding end-task-agnostic fields such as Optical Character Recognition and Scene Text Recognition. This task-centric perspective however might not hold for future studies due to the growing popularity of OCR-free models such as Donut Kim et al. (2022) and Dessurt Davis et al. (2022).

The authors acknowledge that certain requirements that can drive impact in enterprise settings can limit applicability in other aspects, such as performance. For example, in settings where low recall carries high risk, if grounding and explainability come at the cost of recall, end users might prefer to trade them off for better performance.

Lastly, the limitations regarding data and model availability are often motivated by guidelines around data governance and intellectual property, and might not be addressable in the context of a research agenda that is isolated from the legal context. We hope that the research challenges and recommendations made in this position paper motivate further investigations into the legal and governance implications of data- and model-sharing across academic and enterprise entities.

8 Acknowledgements

Armineh Nourbakhsh's work is supported by JPMorgan Chase & Co. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Petr Babkin, William Watson, Zhiqiang Ma, Lucas Cecchi, Natraj Raman, Armineh Nourbakhsh, and Sameena Shah. 2023. Bizgraphqa: A dataset for image-based inference over graph-structured diagrams from business domains. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2691–2700, New York, NY, USA. Association for Computing Machinery.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

TrustBank CBS. Seamless solution for customer onboarding and kyc compliance. https://www.trustbankcbs.com/solutions-Core-banking-software-customer-onboarding.html. Accessed: 2024-02-01.

Chase Media Center. 2024. Consumers are using banking apps for more than transactions, new chase study finds.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

NRZ SERVICER ADVANCE RECEIVABLES TRUST CS. 2013. Amended and restated indenture.

Brian Davis, Bryan Morse, Brian Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end document recognition and understanding with dessurt. In *European Conference on Computer Vision*, pages 280–296. Springer.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Laura Gordon-Murnane. 2010. Creative commons: Copyright tools for the 21st century. *Online Magazine*, 34(1).

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

ibml. 2024. What is document automation? how it works & benefits.

Global Growth Insights. 2024. Document automation software market size.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

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 *European Conference on Computer Vision (ECCV)*.

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.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

McKinsey. 2022. Fueling digital operations with analog data.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. Pawls: Pdf annotation with labels and structure.

OpenAI. 2023. Gpt-4v(ision) system card. Technical report, OpenAI.

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*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Paycom. 2021. The true cost of manual hr processes.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Natraj Raman, Sameena Shah, and Manuela Veloso. 2022. Synthetic document generator for annotation-free layout recognition. *Pattern Recognition*, 128:108660.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. 2023. Docile benchmark for document information localization and extraction. *arXiv preprint arXiv: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 *International Conference on Document Analysis and Recognition*, pages 564–579. Springer.

Stacey Svetlichnaya. 2020. Deepform: Understand structured documents at scale.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023a. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023b. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. 2023. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. *arXiv preprint arXiv:2312.00968*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13716–13730, Singapore. Association for Computational Linguistics.

Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. 2021. NOAHQA: Numerical reasoning with interpretable graph question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ran Zmigrod, Zhiqiang Ma, Armineh Nourbakhsh, and Sameena Shah. 2024a. Treeform: End-to-end annotation and evaluation for form document parsing. *arXiv preprint arXiv:2402.05282*.

Ran Zmigrod, Dongsheng Wang, Mathieu Sibue, Yulong Pei, Petr Babkin, Ivan Brugere, Xiaomo Liu, Nacho Navarro, Antony Papadimitriou, William Watson, Zhiqiang Ma, Armineh Nourbakhsh, and Sameena Shah. 2024b. Buddie: A business document dataset for multi-task information extraction.

A GPT-4V example

Figure 1 shows an interaction with the GPT-4V model, recorded on Jan 7, 2024. The model was prompted on a document classification task, using a document from the RVL-CDIP dataset. The document, a tax revenue report issued by a local tax council, was tagged as a “budget” report within the dataset. GPT-4V mischaracterizes the document as a scientific report based on its style and numeric content. Subsequent prompts with minor modifications to the question do not meaningfully change the output of the model, and can only change the response from “scientific publication” to “statistical report” or “memo”.

B Upstream publishers of popular VrDU datasets

Figure 2 shows the lineage of the datasets listed in Table 3, based on upstream publishers.

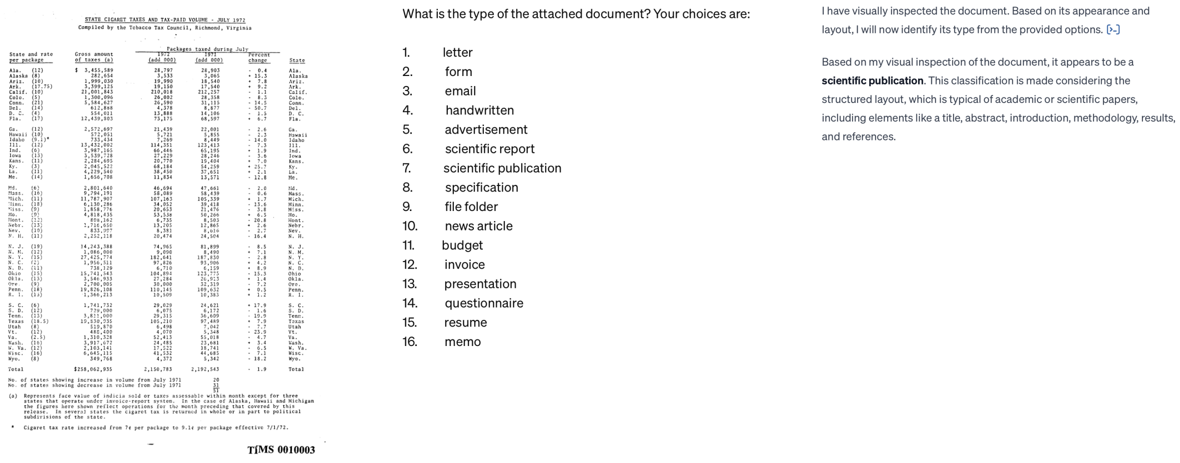


Figure 1: GPT-4V response to a document classification problem. The document is a budget report by a tax council, excerpted from the RVL-CDIP dataset (Harley et al., 2015).

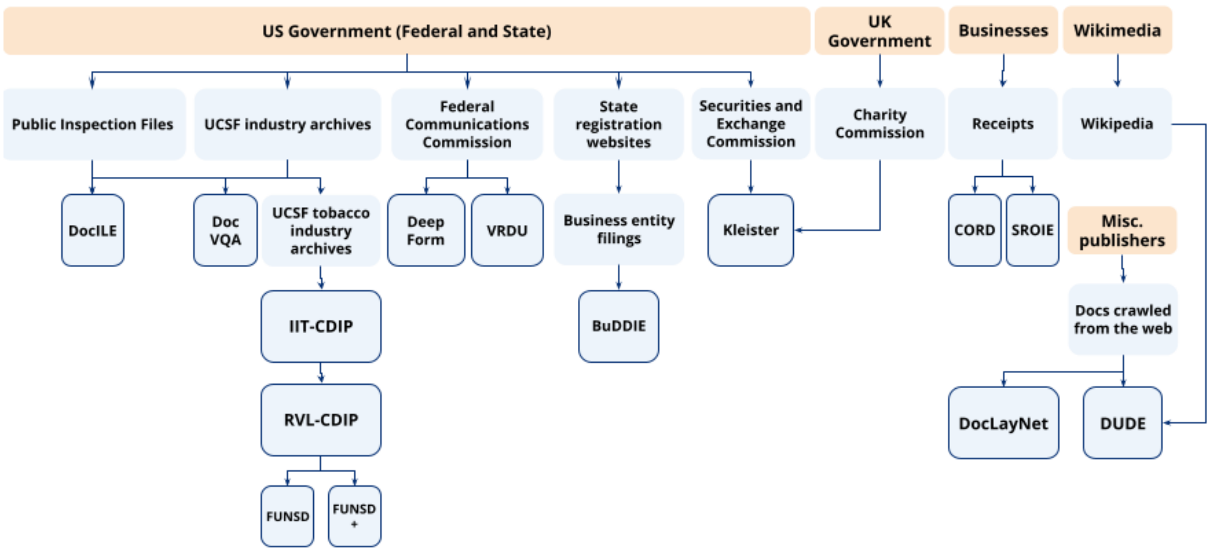


Figure 2: The lineage of the datasets listed in Table 3. Each dataset is displayed in a bordered box. The remaining boxes represent upstream sources of documents, with the most upstream publisher highlighted in orange.