DocKD: Knowledge Distillation from LLMs for Open-World Document Understanding Models

Sungnyun Kim[1*†‡], **Haofu Liao**[2*], **Srikar Appalaraju**[2], **Peng Tang**[2], **Zhuowen Tu**[2], **Ravi Kumar Satzoda**[2], **R. Manmatha**[2], **Vijay Mahadevan**[2], **Stefano Soatto**[2]

[1]KAIST AI [2]AWS AI Labs

Abstract

Visual document understanding (VDU) is a challenging task that involves understanding documents across various modalities (text and image) and layouts (forms, tables, etc.). This study aims to enhance generalizability of small VDU models by distilling knowledge from LLMs. We identify that directly prompting LLMs often fails to generate informative and useful data. In response, we present a new framework (called DocKD) that enriches the data generation process by integrating external document knowledge. Specifically, we provide an LLM with various document elements like key-value pairs, layouts, and descriptions, to elicit open-ended answers. Our experiments show that DocKD produces high-quality document annotations and surpasses the direct knowledge distillation approach that does not leverage external document knowledge. Moreover, student VDU models trained with solely DocKD-generated data is not only comparable to those trained with human-annotated data on in-domain tasks but also significantly excel them on out-of-domain tasks.

1 Introduction

Visual document understanding (VDU) requires extracting and analyzing both textual and non-textual information from a document. The textual information is usually obtained via optical character recognition (OCR), which only provides unstructured or naïvely ordered text. The non-textual information is visually-rich, demanding a solution to directly process the document image. Earlier studies of VDU (Liu et al., 2007; Hao et al., 2016; Soto and Yoo, 2019) primarily focused on identifying certain parts of a document using heuristics or simple networks. Recent approaches (Huang et al., 2022; Tang et al., 2023) have shifted towards pretraining multi-modal document understanding models to address the model's comprehension of textual, visual, and layout features. However, the existing VDU methods are limited by training on a small-scale, curated document dataset, compromising the generalizability of VDU models to diverse documents. Thus, their performance heavily relies on the annotated training document set for downstream tasks.

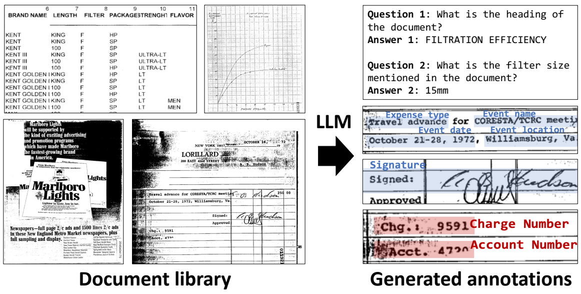

Question 1: What is the heading of the document?
Answer 1: FILTRATION EFFICIENCY
Question 2: What is the filter size mentioned in the document?
Answer 2: 15mm
LLM
Expense type
Event name
Event date
Event location
Signature
Charge Number
Account Number
Document library
Generated annotations

Figure 1: We leverage LLM to generate document annotations given the text extracted from a document image.

In this study, we aim to improve the generalizability of VDU models by distilling knowledge from large language models (LLMs). In particular, we introduce an *open-world document understanding* problem, where the model needs to address the downstream task with a broader scope of documents than covered by the available annotations. LLMs, given instructions to elicit open-ended answers, can create rich and diverse annotations, as illustrated in Fig. 1. For instance, we might instruct the LLM to "`generate question-answer pairs from this document`", along with document text extracted from OCR. However, this approach entails a critical challenge, since LLMs often struggle to comprehend unstructured OCR text (Wang et al., 2023b), leading to its generation of low-quality annotations. Moreover, there is a variety of non-textual information within documents which is not included in the LLM prompt.

To overcome these challenges, we present DocKD, *a document knowledge distillation frame-*

[*] Equal contribution
[†] Work done at AWS AI Labs
[‡] Corresponding author `ksn4397@kaist.ac.kr`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3167–3193
November 12-16, 2024 ©2024 Association for Computational Linguistics

work that leverages external document information to enhance LLM data generation. In this framework, we extract various document elements (*e.g.*, key-value pairs, layout, and descriptions) along with text and formulate a generation prompt for LLMs with this visual information. The LLM outputs then serve as annotations to train a small-scale VDU model. While large multimodal models like GPT-4V (OpenAI, 2023) are also recognized for their visual-language capabilities, they still lag behind state-of-the-art OCR systems (Fujitake, 2024), but LLMs that utilize well-structured OCR text excel in document processing and understanding. Thus, we employ LLMs aided with visual tools for data generation.

We demonstrate the efficacy of DocKD on three document understanding tasks: visual question answering, entity extraction, and classification. In each task, we introduce new tools for incorporating external document knowledge. Our experiments reveal that DocKD allows student models to attain open document understanding abilities, generalizing to unseen documents, questions, entities, or categories. Our contributions are as follows:

- We introduce DocKD, a framework designed to facilitate VDU models for open-world document understanding. It boosts the generalizability of VDU models by leveraging LLMs and external document knowledge to generate training data.
- We demonstrate that DocKD surpasses direct knowledge distillation approach that relies solely on the LLM prompt tuning to generate data without document-specific knowledge.
- In comparison to models trained with human-annotated data, student VDU models trained solely with DocKD-generated data achieve comparable performance on in-domain tasks and excel in addressing out-of-domain tasks. This showcases DocKD's potential to improve models for open-world documents understanding.

2 Related Work

Document understanding models. Research in document intelligence (Liu et al., 2007; Hao et al., 2016; Subramani et al., 2020; Wang et al., 2022b) has gained significant interest, developing machines to understand document contents and address associated tasks. Previous studies (Hong et al., 2020; Wang et al., 2022a) have proposed document understanding models to improve the comprehension of multi-modality by integrating textual and layout information. These models later have evolved to incorporate visual information as well (Appalaraju et al., 2021; Gu et al., 2021; Peng et al., 2022). These models are typically pretrained through self-supervised learning methods, such as word/line alignment (Appalaraju et al., 2023; Tang et al., 2023) or masked text/image modeling (Li et al., 2021; Huang et al., 2022). Subsequently, they undergo a fine-tuning phase for specific downstream tasks, which entails the manual annotation of documents. To facilitate the training of VDU models without the need for human labels, we propose knowledge distillation (Hinton et al., 2015; Gou et al., 2021) approach from LLMs.

Leveraging LLMs for data generation. Knowledge distillation (KD) from LLMs has been explored across various natural language processing tasks (Gu et al., 2023). LLMs like GPT-3 (Brown et al., 2020) are utilized for guided annotation of unlabeled data (Wang et al., 2021; Ding et al., 2022; Touvron et al., 2023; Chiang et al., 2023) or for distilling reasoning capabilities (Magister et al., 2022; Hsieh et al., 2023; Zhu et al., 2023) which is then used to fine-tune smaller language models. Among these, targeted distillation (Jung et al., 2023; Zhou et al., 2023) has demonstrated that identifying and amplifying the LLM's knowledge to a high-quality dataset enables student models to attain task-specific knowledge. It has the potential to make specialized language models that outperform in specific tasks, at the expense of generic performances (Fu et al., 2023).

In visual instruction tuning research (Li et al., 2023a,b,c; Liu et al., 2023b,a), LLMs are employed to generate visual-language instruction-following data. For instance, LLaVA (Liu et al., 2023b) is trained on the instruction-following dataset for conversation, description, and complex reasoning, created by prompting the LLM with bounding box coordinates of objects along with image captions. InstructBLIP (Dai et al., 2023) incorporates diverse tasks, such as image question generation and video question answering. Closest to our work is the extension of visual instruction tuning to the domain of VDU, generating data with document-specific knowledge to fine-tune downstream models. Wang et al. (2023c) use layout-aware documents to answer given questions and fine-tune LLMs, and Aubakirova et al. (2023) generate captions for patent figures to fine-tune VLMs. The community has recently focused on directly im-

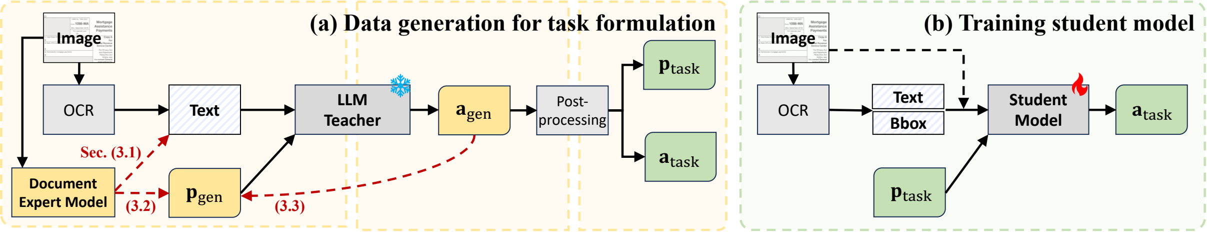

(a) Data generation for task formulation
Image
OCR
Text
LLM Teacher
$\mathbf{a}_{\text{gen}}$
Post-processing
$\mathbf{p}_{\text{task}}$
$\mathbf{a}_{\text{task}}$
Sec. (3.1)
Document Expert Model
(3.2)
$\mathbf{p}_{\text{gen}}$
(3.3)
(b) Training student model
Image
OCR
Text
Bbox
Student Model
$\mathbf{a}_{\text{task}}$
$\mathbf{p}_{\text{task}}$

Figure 2: Overview of DocKD. **(a)** To prepare training data, we provide an LLM teacher with a generation prompt $\mathbf{p}_{\text{gen}}$ given the document text. LLM generates answers $\mathbf{a}_{\text{gen}}$ which are then converted into $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$. We explore methods to inject external document knowledge (- - →) into the document text or $\mathbf{p}_{\text{gen}}$ to obtain high-quality annotations. **(b)** We train a student VDU model using the generated task prompt and answer pairs $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$.

proving the VDU performance of LLMs or LMMs by introducing new designs of encoding document images (Li et al., 2024; Luo et al., 2024; Tanaka et al., 2024; Liu et al., 2024), which are closely related and complementary to our work that focuses on distilling knowledge from strong LLMs for VDU. Our work is the first to extract knowledge from LLMs for open document understanding tasks, exploring methods to inject visual document-specific knowledge into LLM and produce high-quality data for training VDU models.

3 Document Knowledge Distillation

Problem formulation. Similar to prior work (Kim et al., 2022; Appalaraju et al., 2023; Tang et al., 2023), we formulate document understanding problem under a sequence-to-sequence (seq2seq) generation framework. That is, we design a task-specific prompt $\mathbf{p}_{\text{task}}$ which asks a VDU model to solve the task and output an answer $\mathbf{a}_{\text{task}}$. DocKD involves an LLM teacher f_{T} to generate these prompt and answer pairs. Given an image of a document page, we apply a pre-built OCR engine to extract its words and word bounding boxes. For simplicity, we represent a document input as $\mathbf{d}$.

The overall pipeline of the DocKD approach is described in Fig. 2. In Fig. 2 (a), we first construct a generation prompt $\mathbf{p}_{\text{gen}}$ for the task. Then, given $\mathbf{p}_{\text{gen}}$ and document text $\mathbf{d}_{\text{text}}$ as inputs, the LLM generates $\mathbf{a}_{\text{gen}}$, *i.e.*, $f_{\text{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen}}) \to \mathbf{a}_{\text{gen}}$. This can be readily parsed into $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$ by post-processing. Here, we can inject document-specific knowledge into the LLM inputs, so that it can better understand the document content and generate more accurate $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$ pairs. In Fig. 2 (b), we train a student model f_{S} to output an answer $\mathbf{a}_{\text{task}}$ given $\mathbf{d}$ and $\mathbf{p}_{\text{task}}$, *i.e.*, $f_{\text{S}}(\mathbf{d}, \mathbf{p}_{\text{task}}) \to \mathbf{a}_{\text{task}}$.

We exemplify the application of our training pipeline on three document understanding tasks: visual question answering (VQA), entity extraction, and document classification. To summarize each section, we leverage document knowledge by using the OCR linearization model to improve $\mathbf{d}_{\text{text}}$ (Sec. 3.1), using the key-value detection model to guide $\mathbf{p}_{\text{gen}}$ (Sec. 3.2), and introducing the document description into $\mathbf{p}_{\text{gen}}$ for better class candidates (Sec. 3.3). Refer to Appx. B for the full templates of $\mathbf{p}_{\text{gen}}$ in each task.

3.1 Document VQA

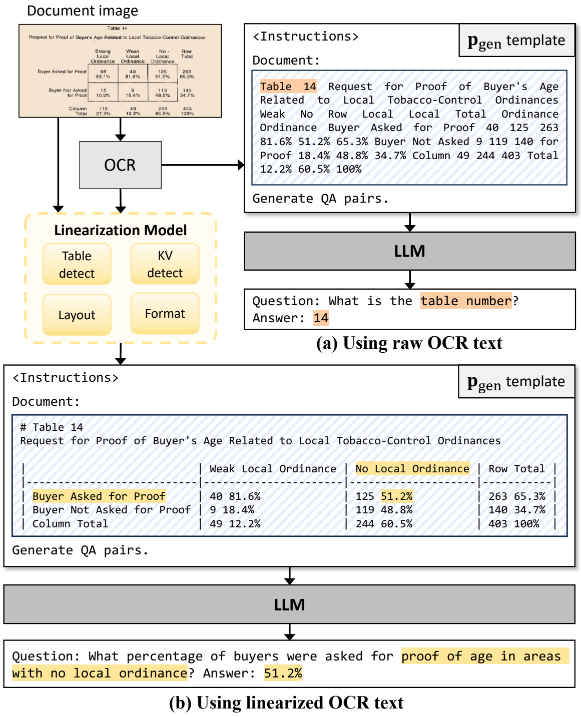
Document VQA (Borchmann et al., 2021; Mathew et al., 2021, 2022; Van Landeghem et al., 2023) is the task of answering questions about documents. Given a document $\mathbf{d}$ and a corresponding question-answer (QA) pair $(\mathbf{q}, \mathbf{a})$, we design the task prompt as $\mathbf{p}_{\text{task}}$ = "`Document:` $\mathbf{d}_{\text{text}}$. `Question:` $\mathbf{q}$", and $\mathbf{a}_{\text{task}}$ = "`Answer:` $\mathbf{a}$". To distill knowledge for a VDU model, we investigate a way to prompt LLMs to generate QA pairs from documents.

Designing QA generation task. Based on the OCR text as input context, we provide the LLM with a generation prompt $\mathbf{p}_{\text{gen}}$ to generate several QA pairs, as shown in Fig. 3 (a):

$$f_{\text{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen}}) \to \mathbf{a}_{\text{gen}} = \{(\mathbf{q}_1, \mathbf{a}_1), (\mathbf{q}_2, \mathbf{a}_2), \ldots\}$$

We randomly select one question and its corresponding answer from $\mathbf{a}_{\text{gen}}$ and create $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$ for training the student model. We find that including an instruction into $\mathbf{p}_{\text{gen}}$ helps the teacher avoid creating low-quality QAs (*e.g.*, duplicated questions or answers inconsistent with context) and enables us to control the generation output so that it can be easily parsed into $(\mathbf{p}_{\text{task}}, \mathbf{a}_{\text{task}})$.

We also note that $\mathbf{p}_{\text{gen}}$ instructs the LLM to output questions and answers *together*, which we find facilitates the generation of accurate QA pairs. Alternatively, we may ask the LLM to generate questions first and then answer them, which we observe



```
<Instructions>
Document:
Table 14 Request for Proof of Buyer's Age
Related to Local Tobacco-Control Ordinances
Weak No Row Local Local Total Ordinance
Ordinance Buyer Asked for Proof 40 125 263
81.6% 51.2% 65.3% Buyer Not Asked 9 119 140 for
Proof 18.4% 48.8% 34.7% Column 49 244 403 Total
12.2% 60.5% 100%
Generate QA pairs.
```

```
Question: What is the table number?
Answer: 14
```

(a) Using raw OCR text

```
<Instructions>
Document:
# Table 14
Request for Proof of Buyer's Age Related to Local Tobacco-Control Ordinances

|                            | Weak Local Ordinance | No Local Ordinance | Row Total |
|----------------------------|----------------------|--------------------|-----------|
| Buyer Asked for Proof      | 40 81.6%             | 125 51.2%          | 263 65.3% |
| Buyer Not Asked for Proof  | 9 18.4%              | 119 48.8%          | 140 34.7% |
| Column Total               | 49 12.2%             | 244 60.5%          | 403 100%  |
Generate QA pairs.
```

```
Question: What percentage of buyers were asked for proof of age in areas
with no local ordinance? Answer: 51.2%
```

(b) Using linearized OCR text

Figure 3: (a) When the input document text is in its raw OCR form, LLM produces simply extracted QA pairs. (b) When provided with linearized OCR text processed by a linearization model, LLM generates QA pairs that require visual layout knowledge to solve.

that the generated questions are often difficult to answer, or the answers do not match the questions.

Introducing layout knowledge to OCR text. One limitation of the LLM's QA generation lies on its text-to-text framework, where it requires the text to be organized in a semantically meaningful order. However, OCR text is a simple sequence of words typically ordered by raster scanning, which ignores the important layout and structural information of document pages. Therefore, QAs generated from such text are usually less challenging and do not cover the spatial relationship between entities.

To ensure the LLM's awareness on the text layout, we replace the raw OCR text with spatially linearized OCR text, where we organize document text into a markdown style as displayed in Fig. 3 (b). We use the linearization model inspired by (Peng et al., 2022), also extracting tables, key-value pairs, and layout information using Textract API[1] which assists the conversion to markdown. Interestingly, an LLM understands this markdown style; thus, the linearization model supplements document layout knowledge that is missing and helps the LLM to generate more diverse and higher-quality QAs. The student model trained with these QA pairs achieves notable VQA performances (Table 1). Refer to Appx. C.1 for the examples of generated QAs with raw or linearized OCR text.

3.2 Entity Extraction

Entity extraction aims to identify entities in the document that matches a given field name. Similar to the VQA task, we convert this task into a seq2seq form. For each field name $\mathbf{f}$ and the corresponding entity $\mathbf{e}$, $\mathbf{p}_{\text{task}}$ = "`Document:` $\mathbf{d}_{\text{text}}$. `Question: what are entities of` $<\mathbf{f}>$?" and $\mathbf{a}_{\text{task}}$ = "`Answer:` $\mathbf{e}$".

The challenge of this task lies in that we do not know which field will be queried for a new document. Thus, we should generate as many diverse fields as possible for different kinds of entities, and train the entity extraction model to link those fields to the entities. Indeed, LLMs are known to be proficient at the entity recognition task (Li et al., 2019; Wang et al., 2023a) and can even identify their names (Zhou et al., 2023).

Designing entity generation task. To generate data for entity extraction, we prompt LLMs to exhaustively extract any entities present in a document. We design an entity extraction prompt $\mathbf{p}_{\text{gen-}ent}$ and send it together with the document text $\mathbf{d}_{\text{text}}$ as the inputs to an LLM, which then outputs a list of entities along with their field names:

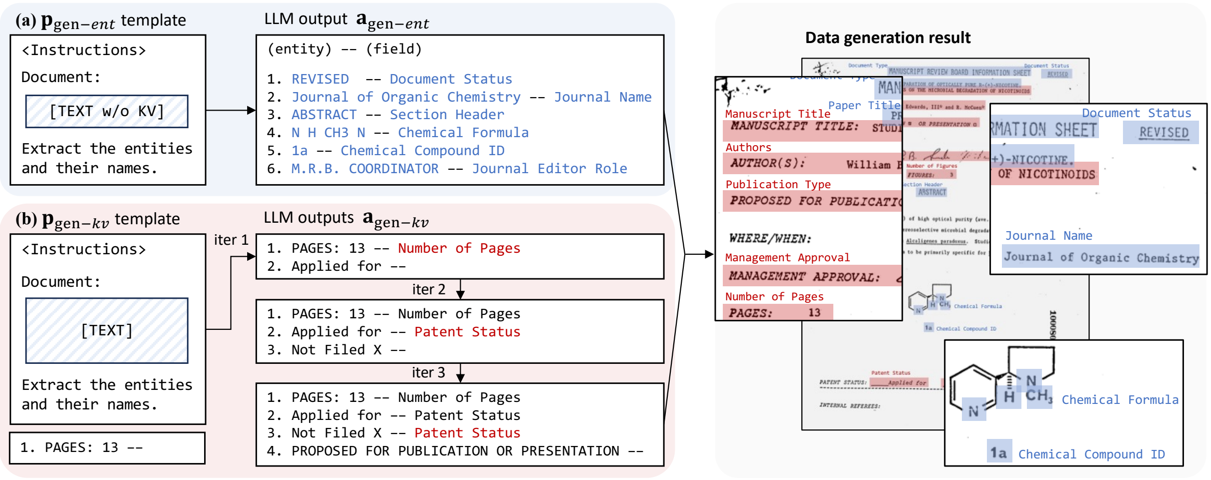
$$f_{\text{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen-}ent}) \rightarrow \mathbf{a}_{\text{gen-}ent} = \{(\mathbf{f}_1, \mathbf{e}_1), (\mathbf{f}_2, \mathbf{e}_2), \ldots\}$$

where $\mathbf{f}_i$ is a generated field name for the i-th entity $\mathbf{e}_i$. We find that LLMs are able to capture a group of words into a single entity and generate a field based on the context, as observed in Fig. 4 (a).

Introducing KV entity knowledge to $\mathbf{p}_{\text{gen}}$. Although LLMs can identify entities from documents to a certain extent, we notice that they are unable to sufficiently enumerate the entities. They tend to list mostly the major ones, especially when there are many potential entities in the document, and fail to identify diverse types. To help LLMs to enumerate them, we propose to leverage a document expert model that extracts key-value (KV) pairs from documents. KV pairs are frequently found in documents, *e.g.*, the entity "`Name: XYZ`" is composed of a key "`Name:`" and a value "`XYZ`".

We detect all KV pairs using an external KV detection model, and send the detected KV pairs to LLMs to obtain their field names. Because there

[1] https://aws.amazon.com/textract/



```
<Instructions>
Document:
[TEXT w/o KV]
Extract the entities
and their names.
```

```
(entity) -- (field)
1. REVISED  -- Document Status
2. Journal of Organic Chemistry -- Journal Name
3. ABSTRACT -- Section Header
4. N H CH3 N -- Chemical Formula
5. 1a -- Chemical Compound ID
6. M.R.B. COORDINATOR -- Journal Editor Role
```

```
<Instructions>
Document:
[TEXT]
Extract the entities
and their names.
```

```
1. PAGES: 13 --
```

```
1. PAGES: 13 -- Number of Pages
2. Applied for --
```

```
1. PAGES: 13 -- Number of Pages
2. Applied for -- Patent Status
3. Not Filed X --
```

```
1. PAGES: 13 -- Number of Pages
2. Applied for -- Patent Status
3. Not Filed X -- Patent Status
4. PROPOSED FOR PUBLICATION OR PRESENTATION --
```

Figure 4: The templates on the left serve as input prompts to the LLM, for (a) generating non-KV entities and (b) naming KV entities, respectively. For (b), in the iteration n, the n-th KV entity is provided as input as well as the output from the previous iteration. On the right, we show the result of generated entities and field names, with blue boxes representing non-KV entities and red boxes representing KV entities.

exist multiple KV pairs, we iteratively present each KV entity line by line to the LLM, with the previous line's output appended (refer to Fig. 4 (b)):

$$f_{\mathrm{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen-}kv}, (\mathbf{f}_i, \mathbf{e}_i)_{1:n}, \mathbf{e}_{n+1}) \rightarrow \mathbf{a}_{\text{gen-}kv} = \mathbf{f}_{n+1}$$

where $\mathbf{f}_{n+1}$ is a field name for the KV entity $\mathbf{e}_{n+1}$, as result of the $(n+1)$-th generation. This way, we make the LLM focus on the field generation only for the current KV entity. In addition, it has access to previous generated outputs, so if there are similar entities given, it can assign the same field.

Note that we do not eliminate the entity generation process by $\mathbf{p}_{\text{gen-}ent}$. Not all entities are detected by the KV detection model, so it is still required to extract non-KV entities. Hence, when generating non-KV entities, we provide the OCR text in which all KV entities are removed.

3.3 Document Classification

We formulate a classification task within a seq2seq framework so that a VDU model can generalize to any novel classes. Specifically, we design the input prompt as $\mathbf{p}_{\text{task}}$ = "Document: $\mathbf{d}_{\text{text}}$. Question: what is the class of this document? choose from the following: {*candidate list*}", and correspondingly, $\mathbf{a}_{\text{task}}$ = "Answer: *class label*". The candidate list contains document class labels, including the answer class. We collect the LLM-generated labels to fill out the prompt without human annotations.

Designing document class generation task. We generate candidates of class labels that can further be used to formulate a downstream classification task. For this, we need two types of generation prompts. $\mathbf{p}_{\text{gen-}pos}$ is used to generate candidates of a given document's type, and we call this output list *positive labels* that may be used as an answer. In order to build a classification task, we not only need the document types that match the given document but also the candidate types that do not match the document. LLM is instructed with $\mathbf{p}_{\text{gen-}neg}$ to suggest these types, which we call *negative labels*.

Introducing knowledge from $\mathbf{a}_{\text{gen}}$ to $\mathbf{p}_{\text{gen}}$. We notice that when an LLM is directly prompted to predict document classes, it frequently generates class labels that are overly general, resulting in low diversity. To address this, we incorporate document descriptions to $\mathbf{p}_{\text{gen}}$ which we find can facilitate LLMs to better summarize a document and generate more diverse class labels.

LLM is instructed with $\mathbf{p}_{\text{gen-}desc}$ = "Describe this document in one sentence". The output document description $\mathbf{a}_{\text{gen-}desc}$ is then appended to the generation prompt for positive labels. This strategy makes the positive labels more diverse and detailed, *e.g.*, letter → consumer letter. Subsequently, we also use the output positives in the negatives generation prompt, in order to avoid generating labels that are similar to the positives. We summarize the generation steps as follows:

(1) description: $f_{\mathrm{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen-}desc}) \rightarrow \mathbf{a}_{\text{gen-}desc}$,

(2) positives: $f_{\mathrm{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen-}pos}, \mathbf{a}_{\text{gen-}desc}) \rightarrow \mathbf{a}_{\text{gen-}pos}$,

(3) negatives: $f_{\mathrm{T}}(\mathbf{d}_{\text{text}}, \mathbf{p}_{\text{gen-}neg}, \mathbf{a}_{\text{gen-}pos}) \rightarrow \mathbf{a}_{\text{gen-}neg}$.

While this approach does not directly leverage visual information, it adopts a similar strategy to the chain-of-thought reasoning (Wei et al., 2022; Hsieh et al., 2023) that encourages better outputs by prompting the instruction steps to LLMs.

Candidate list formulation. We select one positive label the list $\mathbf{a}_{\text{gen-}pos}$, as an answer. For other non-answer candidates, we randomly sample a few from $\mathbf{a}_{\text{gen-}neg}$. We train the model to choose one among the $\{\mathtt{positive} + \mathtt{negatives}\}$ list. In addition, the generated description $\mathbf{a}_{\text{gen-}desc}$ is appended to each positive label to give a hint about the class. We also gather all unique negative classes and use the LLM to produce descriptions for these types, which are also appended to the labels. Refer to Appx. B.3 for the prompt we used based on this.

4 Experiments and Results

4.1 Implementation Details

Models. We compare the DocKD performance with the plain KD approach, naïvely using $\mathbf{d}_{\text{text}}$ and $\mathbf{p}_{\text{gen}}$ without external document knowledge, as a prompt engineering baseline. By default, we use Claude-2 [2] as a teacher LLM and DocFormerv2$_{\text{large}}$ (Appalaraju et al., 2023) as a student VDU model, while partially using DocFormerv2$_{\text{base}}$ to facilitate more efficient analysis. The training procedure of DocFormerv2 (DFv2) closely follows that of the original paper, where it jointly encodes document image, OCR text, and bounding boxes. The provided query ($\mathbf{p}_{\text{task}}$) is appended to the text ($\mathbf{d}_{\text{text}}$), and the decoder outputs the target answer ($\mathbf{a}_{\text{task}}$).

For comparison, we also employ Flan-T5$_{\text{large}}$ (Chung et al., 2022) as a student language-only model, since the DFv2 structure is based on T5 (Raffel et al., 2020). To provide a base comparison for each task, we additionally present the zero-shot performance of instruction-tuned LLMs (Chung et al., 2022; Almazrouei et al., 2023b; Chiang et al., 2023) and a vision-language multi-modal foundation model (Liu et al., 2023a).

Datasets. For the LLM's data generation, we use a randomly sampled subset of Industry Document Library (IDL, Lewis et al. (2006)) as unannotated document images. To accurately evaluate the open-world capabilities, we have removed all IDL documents that overlap with any of our downstream task datasets and excluded them from the data generation phase. For the evaluation datasets and metrics, we use DocVQA (Mathew et al., 2021) validation set in the document VQA task, measured by ANLS (average normalized Levenshtein similarity) (Biten et al., 2019) and EM (exact match). In the entity extraction, we use two datasets, CORD (Park et al., 2019) and DeepForm (Borchmann et al., 2021), evaluated by entity-level F1 score and ANLS, respectively. In the classification task, we use RVL-CDIP (Harley et al., 2015) test set, evaluated by the mean accuracy over 16 document categories. Refer to Appx. D for more details on each dataset.

4.2 Evaluation on Open-World Document Understanding Tasks

Document VQA. Claude-2 generates QAs from randomly sampled 100K IDL documents. We prompt Claude-2 to generate three QA pairs per document sample, and the trained student model is evaluated on DocVQA (Mathew et al., 2021). Table 1 (a) summarizes the DocVQA performances of the distilled students as well as the LLMs, where none of these models have been trained on human annotations for the document VQA task. We confirm that knowledge-distilled student models can effectively answer document questions, being comparable with much larger-size language models.

Compared to the plain KD with raw OCR text, DocKD significantly enhances the performance up to 81.0% ANLS. This result is comparable to using human-labeled annotations (refer to Sec. 4.3), which implies the high quality of generated data. Furthermore, the performance gain is greater with DFv2 (vision + language) than Flan-T5 (language), which shows that the linearization model supplements informative visual knowledge.

Entity extraction. For generating the entities with KV detection, we need documents with rich key and value information. Such documents are frequently found from forms or invoices. Thus, instead of using IDL, we use the invoices subset of RVL-CDIP (Harley et al., 2015) for entity generation, sampling 5K documents. Table 1 (b) demonstrates that if the data generation does not involve the KV detection model but only exploits the entity generation prompt $\mathbf{p}_{\text{gen-}ent}$, the LLM produces low-quality entities and field names, leading to the subpar performance of the student models.

Document classification. We sample 50K documents from IDL to generate class labels. For each document sample, Claude-2 generates one-sentence description, three positive labels, and ten

[2] https://www.anthropic.com/index/claude-2

		(a) VQA		(b) Entity extraction		(c) Classification	
model	size	val ANLS	val EM	test F1	test ANLS	test mAcc	test mAcc*
LLM zero-shot prediction							
Flan-T5$_{\text{large}}$ (Chung et al., 2022)	750M	59.6	48.8	0.90	2.57	46.7	54.0
Flan-T5$_{\text{XXL}}$ (Chung et al., 2022)	11B	70.4	60.0	21.2	24.1	52.0	58.1
LLaVA-1.5 (Liu et al., 2023a)	13B	49.0	37.3	9.12	5.20	36.1	43.3
Vicuna-1.3 (Chiang et al., 2023)	33B	62.4	51.9	24.3	27.6	48.4	57.7
Falcon (Almazrouei et al., 2023b)	40B	72.4	62.7	48.5	38.7	37.9	43.3
*VDU models trained with **only generated** data*							
Flan-T5$_{\text{large}}$ + KD	750M	70.4	59.4	24.4	56.3	52.3	59.8
Flan-T5$_{\text{large}}$ + **DocKD**	750M	72.9	62.7	55.9	66.1	57.0	71.7
DocFormerv2$_{\text{large}}$ + KD	750M	76.9	67.4	30.2	51.8	58.6	69.0
DocFormerv2$_{\text{large}}$ + **DocKD**	750M	**81.0**	**71.9**	**61.5**	**68.7**	**62.4**	**73.9**

Table 1: Document understanding results for LLMs and student VDU models. Note that none of these models were trained with human-labeled annotations. (a) DocVQA validation performance. KD baseline uses raw OCR text for the QA generation, while DocKD uses linearized OCR text. (b) Entity extraction performance on CORD (F1) and DeepForm (ANLS). KD baseline generates entities without KV detection. (c) RVL-CDIP test accuracy. For DocKD, both class labels and descriptions are generated. mAcc* measures the mean accuracy excluding four ambiguous categories: memo, filefolder, handwritten, and presentation.

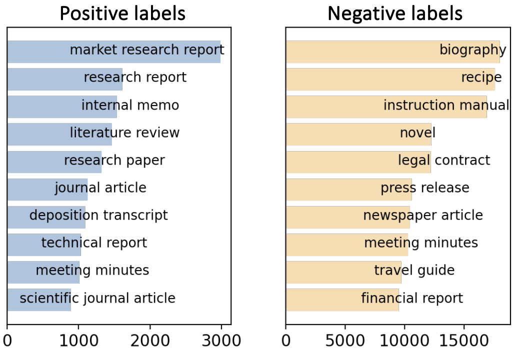


Figure 5: Top-10 frequently generated document class labels from IDL (Lewis et al., 2006).

		DocVQA	CORD	DeepForm	RVL-CDIP
teacher	student	val ANLS	test F1	test ANLS	test mAcc
Falcon-40B	DFv2$_{\text{base}}$	68.6	55.1	48.5	54.7
Falcon-180B	DFv2$_{\text{base}}$	71.3	59.8	62.0	53.8
Claude-2	DFv2$_{\text{base}}$	**77.2**	**60.2**	**64.2**	**61.9**
Falcon-40B	DFv2$_{\text{large}}$	74.9	59.8	61.2	55.6
Falcon-180B	DFv2$_{\text{large}}$	76.8	**66.6**	64.5	58.5
Claude-2	**DFv2$_{\text{large}}$**	**81.0**	61.5	**68.7**	**62.4**

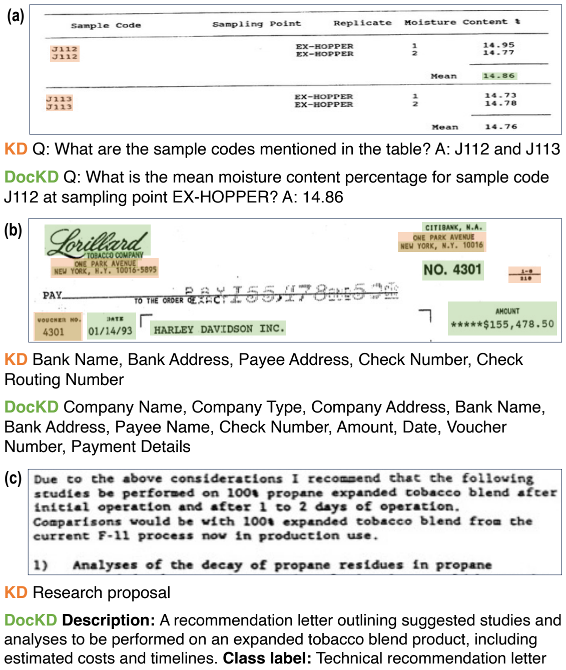
Table 2: We compare the Claude-2 teacher with Falcon-40B and Falcon-180B teacher models, and the DFv2$_{\text{large}}$ (750M) and DFv2$_{\text{base}}$ (232M) student models.

negative labels. Table 1 (c) shows that our distillation framework enables the student model to classify novel documents, removing the need to predefine categories or collect annotated documents to train a classification model. In addition, we find that DocKD's description generation induces more knowledge on documents compared to the plain KD, improving the accuracy by large margin: 58.6% → 62.4% mAcc.

Fig. 5 shows the spectrum of generated class labels from the IDL documents. After filtering out invalid labels (*e.g.*, too long or outliers), it amounts to 49.9K unique positive labels and 10.5K unique negative labels. Before introducing the description generation, we had 17.2K unique positives, implying that the provision of description contributes to increasing the label diversity.

Smaller teacher and student models. Table 2 presents the result with a smaller teacher, Falcon-40B (Almazrouei et al., 2023b), and a smaller student, DFv2$_{\text{base}}$. We find that smaller teacher and student models can degrade the data generation quality and task performances. In contrast, larger and stronger teacher models like Claude-2 or Falcon-180B (Almazrouei et al., 2023a) can generate better data, leading to the highest task performances. For instance, Claude-2 better understands the linearized OCR text than Falcon-40B does, so it generates diverse and accurate QAs from the layout-aware text. Refer to Appx. C for comparisons between different teacher models.

Visualization and statistics of generated data. Fig. 6 visualizes some sample data generated by using KD and DocKD respectively. For document VQA, DocKD generates more challenging QA pairs that requires understanding the structure of the table. In Fig. 6 (a), the question generated by DocKD requires understanding the relationship between "mean", "moisture content %", "sample code" and "sample point". For entity extraction, we show a common example in Fig. 6 (b) where we list the entity names extracted by KD and DocKD. We see that DocKD is able to capture significantly more entities than KD. For document classification, we note that DocKD generates a document description which help to give class labels that aligns better with the document content. Additional examples of DocKD-generated data are available in Appx. C.



(a)

Sample Code	Sampling Point	Replicate	Moisture Content %
J112 J112	EX-HOPPER EX-HOPPER	1 2	14.95 14.77
		Mean	14.86
J113 J113	EX-HOPPER EX-HOPPER	1 2	14.73 14.78
		Mean	14.76

KD Q: What are the sample codes mentioned in the table? A: J112 and J113

DocKD Q: What is the mean moisture content percentage for sample code J112 at sampling point EX-HOPPER? A: 14.86

(b)

Lorillard
TOBACCO COMPANY
ONE PARK AVENUE
NEW YORK, N.Y. 10016-5895

CITIBANK, N.A.
ONE PARK AVENUE
NEW YORK, N.Y. 10016

NO. 4301

1-8
210

PAY
TO THE ORDER OF

VOUCHER NO. 4301

DATE 01/14/93

HARLEY DAVIDSON INC.

AMOUNT *****$155,478.50

KD Bank Name, Bank Address, Payee Address, Check Number, Check Routing Number

DocKD Company Name, Company Type, Company Address, Bank Name, Bank Address, Payee Name, Check Number, Amount, Date, Voucher Number, Payment Details

(c)

Due to the above considerations I recommend that the following studies be performed on 100% propane expanded tobacco blend after initial operation and after 1 to 2 days of operation. Comparisons would be with 100% expanded tobacco blend from the current F-11 process now in production use.

1) Analyses of the decay of propane residues in propane

KD Research proposal

DocKD **Description:** A recommendation letter outlining suggested studies and analyses to be performed on an expanded tobacco blend product, including estimated costs and timelines. **Class label:** Technical recommendation letter

Figure 6: Comparison between data generated by KD and DocKD: (a) document VQA, (b) entity extraction, and (c) document classification.

	entity extraction		document classification	
method	# of ent. types	# ent. per doc.	# pos. labels	# neg. labels
KD	1454	11.5	4674	2476
DocKD	2316	20.1	6053	3013

Table 3: Statistics of data generated by KD and DocKD.

Table 3 shows some statistics of the data generated by KD and DocKD. For entity extraction, we calculate the number of unique entity types (# of ent. types) and average number of entities generated per document (# of ent. per doc.). We note that DocKD can generate significantly more entities and entity types than KD, by leveraging external document knowledge. Similarly, we also summarize the number unique document labels generated by KD and DocKD for document classification. For both the positive and negative class labels, DocKD generates more unique labels than KD. We attribute this to leveraging document descriptions for generation which helps LLMs generating fine-grained labels that align better with the document.

4.3 Leveraging Human-Labeled Annotations

Human annotation QAs. We demonstrate that unsupervised knowledge from an LLM remains valuable even when human annotations are available for training. As shown in Table 4 (a), augmenting DocVQA human annotations with DocKD-generated QAs, which incorporate a variety of document knowledge, results in stronger student models, achieving 83.4% ANLS on the DocVQA validation set. In a more practical scenario where human-labeled documents have different distribution, we utilize DUDE, a dataset featuring multi-domain documents with diverse VQA annotations (text, numerical, yes/no, lists, etc.). In Table 4 (b), DocKD-generated data significantly enhances student model performance, reaching 79.1% ANLS, compared to 66.0% with human annotations alone.

		DocVQA val		DUDE val	
human anno.	**DocKD**-generated	ANLS	EM	ANLS	EM
(a) human anno. = DocVQA train set					
✓		80.6	72.0	53.8	37.2
	✓	77.2	68.6	52.6	36.0
✓	✓	**83.4**	**76.2**	**55.3**	**38.8**
(b) human anno. = DUDE train set					
✓		66.0	54.9	54.4	40.0
	✓	77.2	68.6	52.6	36.0
✓	✓	**79.1**	**70.8**	**58.0**	**42.1**

Table 4: The document VQA task performance using a human-annotated training dataset. **DocKD** indicates the generated QAs from the IDL documents. The teacher model is Claude-2, and the student model is DFv2$_{\text{base}}$. For results with DFv2$_{\text{large}}$, refer to Appx. A.2.

	RVL-CDIP test		out-of-domain		
model	$\mathcal{C}_1$ (known)	$\mathcal{C}_2$ (unk.)	RVL-O	IRS-50	WikiDoc
Falcon-40B	62.3	27.4	76.3	54.0	39.8
DFv2$_{\text{base}}$ S	**86.1**	0.08	0.00	0.00	0.00
DFv2$_{\text{base}}$ U	50.5	**56.1**	42.6	74.0	44.4
DFv2$_{\text{base}}$ S+U	77.1	52.1	**52.8**	**82.0**	**45.2**

Table 5: Open-set classification performance. S: supervised training with $\mathcal{C}_1$ annotations, U: unsupervised DocKD from LLM-generated class labels.

Open-set classification. One of the main applications by distilling LLM's knowledge lies in its open-set classification ability, *i.e.*, it can classify documents of unseen categories. The diversity of generated class labels ensures robustness, while a fixed set of annotations makes it hard to adapt to unseen labels. To verify this, let $\mathcal{C}$ denote the set of all RVL-CDIP labels, and we split $\mathcal{C}$ into two sets: $\mathcal{C}_1 = \{\text{email, letter, memo, news article}\}$ and $\mathcal{C}_2 = \mathcal{C} - \mathcal{C}_1$. We train the model with documents from the web, crawled by $\mathcal{C}_1$ labels (Larson et al., 2022). Table 5 shows that this supervised model (S) makes highly biased predictions—while it predicts known classes accurately (86.1%), it struggles to identify unknown categories in $\mathcal{C}_2$. In contrast, DocKD without any supervised data (U) enables generalization to unseen types of documents. Further, merging the $\mathcal{C}_1$ annotations with the generated data (S+U) leverages the advantages of both supervised and unsupervised learning.

We also evaluate our model in a more realistic distribution of data and labels, using the documents out of the domain of IDL or RVL-CDIP. To this end, we use three evaluation sets, RVL-O (Larson et al., 2022), IRS-50, and WikiDoc (Fujinuma et al., 2023), all of which contain out-of-domain documents (refer to Appx. D for the details of datasets). While the supervised model cannot handle these novel categories, unsupervised DocKD makes the student model even adaptable to out-of-domain classification and outlier detection, following the LLM teacher's robust predictions.

5 Conclusion

We address the open-world document understanding problem by instructing the LLMs to generate document annotations, given the generation prompt and OCR text. To successfully achieve this, we suggest DocKD framework, designing task prompts and answers that LLMs can easily generate, and incorporate external document knowledge from various sources. Consequently, the student models distilled by DocKD annotations demonstrate remarkable performance improvements compared to the plain KD approach in various document tasks. The integration with human-labeled annotations further enhances model performance.

Limitations

This study represents the pioneering work to utilize LLMs for open-world document understanding, specifically focusing on relatively simpler documents and tasks. We have applied LLMs to generate document annotations, and subsequently, trained student VDU models using these annotations. Our primary focus has been on common document understanding tasks such as visual question answering, entity extraction, and classification, which primarily involve documents containing tables, layouts, and forms.

However, extending our approach to handle documents with more complex visual elements, such as intricate figures, diagrams, or dense equations, remains an area for future exploration. While addressing more sophisticated problems could significantly enhance the model's applicability, such advancements would require efforts in developing new generative prompts. Furthermore, integrating LLMs with document expert models and large multimodal models, such as GPT-4V, holds potential to synthesize visually-rich, informative annotations. This integration has not yet been explored and represents a promising avenue for future research. Despite these limitations, our study lays foundational work for more complex applications in the field of document understanding using LLMs.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023a. The falcon series of language models: Towards open frontier models.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023b. Falcon-40B: an open large language model with state-of-the-art performance.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2023. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*.

Dana Aubakirova, Kim Gerdes, and Lufei Liu. 2023. Patfig: Generating short and long captions for patent figures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2843–2849.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE.

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion

Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Yoshinari Fujinuma, Siddharth Varia, Nishant Sankaran, Srikar Appalaraju, Bonan Min, and Yogarshi Vyas. 2023. A multi-modal multilingual benchmark for document image classification. *arXiv preprint arXiv:2310.16356*.

Masato Fujitake. 2024. Dtrocr: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8025–8035.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang. 2016. A table detection method for pdf documents based on convolutional neural networks. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292. IEEE.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2020. Bros: A pre-trained language model for understanding texts in document.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

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.

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.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *arXiv preprint arXiv:2305.16635*.

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*, pages 498–517. Springer.

Stefan Larson, Yi Yang Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating out-of-distribution performance on document image classifiers. *Advances in Neural Information Processing Systems*, 35:11673–11685.

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.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023c. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2024. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15546–15555.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Ying Liu, Kun Bai, Prasenjit Mitra, and C Lee Giles. 2007. Tableseer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 91–100.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

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.

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

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, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Carlos Soto and Shinjae Yoo. 2019. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3464–3470.

Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19071–19079.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jordy Van Landeghem, Rubén 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.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023b. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023c. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2022b. A benchmark for structured extractions from complex documents. *arXiv preprint arXiv:2211.15421*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xingwei Long, and Bowen Zhou. 2023. Pad: Program-aided distillation specializes large models in reasoning. *arXiv preprint arXiv:2305.13888*.

Appendix

A Additional Experiments

A.1 Statistical Significance of Document Understanding Results

We have conducted further experiments to substantiate our findings about statistical significance. Specifically, we reproduced the main results across all three tasks (Table 1) by rerunning the experiments for the configurations DocFormerv2$_{\text{large}}$ + KD and DocFormerv2$_{\text{large}}$ + DocKD using three different random seeds. The results of these additional runs are summarized in Table 6. These results underscore the statistical significance and reliability of our approach.

	(a) VQA		(b) Entity extraction		(c) Classification	
Model	val ANLS	val EM	test F1	test ANLS	test mAcc	test mAcc*
KD run #1	76.88	67.38	30.20	51.81	58.57	68.99
KD run #2	76.28	66.97	32.70	48.72	60.07	66.81
KD run #3	75.71	66.24	28.90	49.77	61.30	70.90
KD	$76.29_{\pm 0.59}$	$66.86_{\pm 0.58}$	$30.60_{\pm 1.93}$	$50.10_{\pm 1.57}$	$59.98_{\pm 1.37}$	$68.90_{\pm 2.05}$
DocKD run #1	81.00	71.85	61.46	68.66	62.40	73.93
DocKD run #2	80.59	72.16	62.95	70.29	63.17	74.76
DocKD run #3	80.10	71.60	62.95	69.58	63.88	73.93
DocKD	$80.56_{\pm 0.45}$	$71.87_{\pm 0.28}$	$62.45_{\pm 0.86}$	$69.51_{\pm 0.82}$	$63.15_{\pm 0.74}$	$74.21_{\pm 0.48}$

Table 6: Statistical significance of our experiments on document understanding tasks. Run #1 are the results reported in the main paper. KD and DocKD are the results with mean ± standard deviation of the three runs.

human anno.	**DocKD**-generated	val ANLS	val EM
(a) human anno. = DocVQA train set			
✓		85.4	77.7
	✓	81.0	71.9
✓	✓	**86.1**	**79.1**
(b) human anno. = DUDE train set			
✓		74.8	64.0
	✓	**81.0**	**71.9**
✓	✓	80.3	71.6

Table 7: DocVQA validation performance using a human-annotated training dataset, (a) DocVQA train set and (b) DUDE train set. **DocKD** indicates the generated QAs from the IDL documents. The teacher model is Claude-2, and the student model is DFv2$_{\text{large}}$.

A.2 Additional Results on DocVQA

DFv2$_{\text{large}}$ model performance. Table 7 presents the DocVQA validation performance with DFv2$_{\text{large}}$ trained on the human-annotated dataset, as in Table 4 with DFv2$_{\text{base}}$. Generated QAs by DocKD are comparable to the human-labeled train set, whereas human annotations with a significantly different distribution (*e.g.*, DUDE (Van Landeghem et al., 2023)) may even degrade performance.

DocVQA test set performance. In Table 8, we provide the test set performance on DocVQA (Mathew et al., 2021), in order to compare with the previous VDU models, which were all trained on the DocVQA training set.

A.3 Data Volume and Quality

In Fig. 7, we emphasize the significance of the distilled data volume in capturing diverse knowledge. Additionally, the introduction of a small set of human annotations (*e.g.*, DUDE (Van Landeghem et al., 2023)) from a different domain proves beneficial, especially when the teacher model size is small and thus generates data of lower quality.

However, it is crucial to note that a larger vol-

model	size	ANLS
DocVQA supervised learning		
Donut$_{\text{base}}$ (Kim et al., 2022)	143M	67.5
T5$_{\text{large}}$ (Raffel et al., 2020)	750M	70.4
LayoutLMv2$_{\text{large}}$ (Xu et al., 2020)	426M	86.7
LayoutLMv3$_{\text{large}}$ (Huang et al., 2022)	368M	83.4
UDOP (Tang et al., 2023)	794M	84.7
DocFormerv2$_{\text{large}}$ (Appalaraju et al., 2023)	750M	86.3†
Training with Claude-2-generated data		
DocFormerv2$_{\text{large}}$ + KD QA	750M	75.8
DocFormerv2$_{\text{large}}$ + **DocKD QA**	750M	80.6
DocFormerv2$_{\text{large}}$ + **DocKD QA** (+ DocVQA anno.)	750M	**86.9**

Table 8: DocVQA test set performance. The KD baseline uses raw OCR text for the QA generation, while DocKD uses the linearized OCR text. †: reproduced without searching hyperparameters. The same hyperparameters were used for training with DocKD QAs.

ume of generated data does not always guarantee superior performance, *i.e.*, quality of the dataset is also important. For the classification task, we established evaluation criteria for generated labels, accounting for both word length and frequency within the dataset. Labels exceeding a word length of 5 (considered overly specific) or occurring less than 3 times throughout the dataset (outliers) were excluded. Documents without remaining positive labels were removed, consequently reducing our IDL training set size from 50K to 43K. This refinement enhanced overall data quality, resulting in an improved test accuracy (+3.5%). Similarly, in VQA and entity extraction tasks, we filtered out excessively long or short questions/answers and field names identified as outliers.

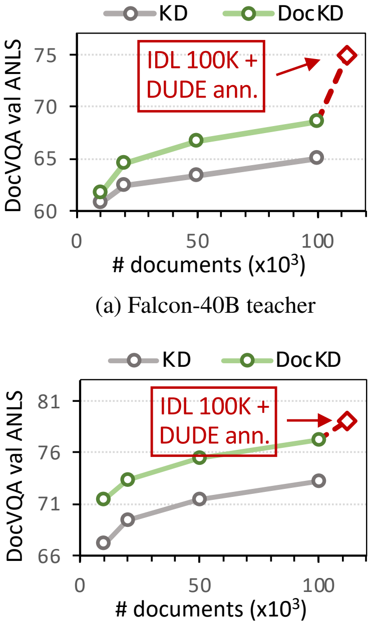
A.4 Using Human-Labeled FUNSD Entities

For the entity extraction task, we utilized RVL-CDIP invoices (Harley et al., 2015), extracting keys and values, and applying the entity generation prompts. Here, we use FUNSD (Jaume et al., 2019) dataset, which is a small subset of RVL-CDIP forms, and all the KV entities are manually annotated. In this case, we use their annotations for the KV entity inputs. Table 9 shows that, although FUNSD contains only a small number of document samples, an LLM can generate reliable KV entity fields based on the manual annotations. Combining with invoices documents that have abundant entities, the student model is effectively distilled with diverse knowledge and can exhibit the highest entity extraction performances.

A.5 Ablation Study on Entity Generation Strategies

In the entity extraction task, we have utilized the LLM's entity recognition ability and the KV detection model's key-value extraction ability. To unveil the individual contributions of each component, Table 10 presents an ablation study on different entity generation methods. Using only $\mathbf{p}_{\text{gen-}ent}$ represents the plain KD baseline without external document knowledge. On the other hand, using only $\mathbf{p}_{\text{gen-}kv}$ eliminates the LLM's automatic extraction of entities that are not detected as keys or values. In addition to these approaches, we conduct key normalization method, where the LLM generates variants for each key name, and these normalized variants serve as the field for the KV entities. This method does not utilize KV entity constraints, which have been used in DocKD as an iterative presentation of KV entities for consistency with previous entities and fields.

The ablation study results confirm the significace of both $\mathbf{p}_{\text{gen-}ent}$ and $\mathbf{p}_{\text{gen-}kv}$, coupled with KV



(a) Falcon-40B teacher

(b) Claude-2 teacher

Figure 7: DocVQA (Mathew et al., 2021) results according to the number of generated data. x-axis is the number of IDL (Lewis et al., 2006) documents used by the LLM to generate the QA pairs.

teacher	gen. data (# doc.)	# entities	CORD	DeepForm
Falcon-40B	FUNSD (149)	2,308	33.2	44.6
Falcon-40B	Invoices (5,000)	38,121	**55.1**	48.5
Falcon-40B	FUNSD + Invoices	40,429	54.9	**52.2**
Claude-2	FUNSD (149)	2,608	42.8	49.1
Claude-2	Invoices (5,000)	74,289	60.2	64.2
Claude-2	FUNSD + Invoices	76,897	**60.4**	**67.5**

Table 9: Entity extraction from FUNSD (Jaume et al., 2019) and RVL-CDIP invoices (Harley et al., 2015) documents. The student model is DFv2$_{\text{base}}$.

method	Entity recognition	KV detection	KV constraints	F1
$\mathbf{p}_{\text{gen-}ent}$ (KD)	✓	✗	✗	20.9
key normalization	✗	✓	✗	39.2
$\mathbf{p}_{\text{gen-}kv}$	✗	✓	✓	45.6
$\mathbf{p}_{\text{gen-}ent} + \mathbf{p}_{\text{gen-}kv}$ (**DocKD**)	✓	✓	✓	**55.1**

Table 10: Ablation study on CORD (Park et al., 2019) entity extraction. Entities and field names are generated from 5K RVL-CDIP invoices (Harley et al., 2015) by the Falcon-40B (Almazrouei et al., 2023b) teacher. The student model is DFv2$_{\text{base}}$. Note that $\mathbf{p}_{\text{gen-}kv}$ always requires the KV detection in prior.

method	mAcc	mAcc*
KD L	56.3	63.4
KD L+D	57.9	68.4
DocKD L+D	**61.9**	**74.0**

Table 11: Ablation study on RVL-CDIP (Harley et al., 2015) classification. The student model is DFv2$_{\text{base}}$, and the teacher model is Claude-2.

detection. Notably, providing the LLM with detected KV pairs yields substantial improvement ($\mathbf{p}_{\text{gen-}ent}$ vs. DocKD), while the extraction of non-KV entities also proves to be crucial ($\mathbf{p}_{\text{gen-}kv}$ vs. DocKD). Injecting context on previous KV entities and the generated fields further enhances the reliability of subsequent generation (key normalization vs. DocKD).

A.6 Ablation Study on the Effect of Descriptions

In the document classification task, descriptions play a crucial role in two key aspects: generating positive labels and appending descriptions when constructing the candidate list. To assess the effect of each aspect, we establish an ablation baseline, KD L+D, and compare three distillation methods:

- **KD L**: LLM generates only class labels without any description.
- **KD L+D**: LLM generates description and, in sequence, class labels based on the description. However, it does not append the desciptions to the class labels during the formulation of the candidate list.
- **DocKD L+D**: LLM generates description and, in sequence, class labels based on the description. These descriptions are appended to the candidate list to give a hint about the class.

Table 11 substantiates the efficacy of utilizing descriptions in both aspects. However, the superior performance gain is observed when appending descriptions to the candidate list. This suggests that designing the task prompt to incorporate rich information about the labels is an effective strategy in training the student model.

A.7 Full Results of RVL-CDIP Classification

Table 12 shows the full category results for document classification, which were sumarized into mean accuracy in Table 1 (c).

B Generation Prompts for LLMs

We provide full templates for the generation prompts $\mathbf{p}_{\text{gen}}$, which are input to the LLM in conjunction with the document text. The generation prompts enable the LLM to proficiently generate document annotations, which are further used to train student models.

B.1 Generation Prompt for Document VQA

In the document VQA task, the generation prompt serves as a guidance for the LLM to generate a fixed number of question-answer (QA) pairs, which can be answered by referencing the document's OCR text. To facilitate this process, we provide two instructive examples and articulate several rules. Then, for the specific target document, which is an IDL (Lewis et al., 2006) document in our study, we extract OCR text from the image, convert it to linearized text (refer to Sec. 3.1), and embed this text into the placeholder {LINEARIZED_TEXT_PLACE_HOLDER} in $\mathbf{p}_{\text{gen}}$. We set {COUNT_PLACE_HOLDER} to three.

$\mathbf{p}_{\text{gen}}$ for QA pair generation

```
[Example 1]

Document: Confidential RJRT PR APPROVAL DATE:
1/8/93 SUBJECT: Ru IVAs PROPOSED RELEASE DATE: for
response FOR RELEASE TO: CONTACT: P. CARTER ROUTE
TO: Name Initials Date Peggy Carter Ace 1/1/15
Kaura Payne nt. T/R Return to Peggy Carter, PR, 16
Reynolds Building Not

Generate three question-answer pairs from this
document.

Question: what is the date mentioned in this
letter?
Answer: 1/8/93

Question: what is the contact person name
mentioned in this letter?
Answer: P. Carter

Question: What is the address of Peggy Carter?
Answer: 16 Reynolds Building


[Example 2]

Document: Link between IR and CVD THE ROUTE TO
```

model	letter	form	email	handwritten	advertisement	scientific report	scientific publication	specification	file folder	news article	budget	invoice	presentation	questionnaire	resume	memo	mAcc
LLM zero-shot prediction																	
Flan-T5$_{\text{large}}$ (Chung et al., 2022)	15.0	8.2	66.5	0.3	68.3	50.2	91.0	62.5	4.2	59.9	29.6	83.7	19.9	62.5	50.1	73.0	46.6
Flan-T5$_{\text{XXL}}$ (Chung et al., 2022)	36.5	31.7	88.8	5.0	65.0	50.8	44.2	58.7	11.3	80.4	26.7	75.4	32.5	77.5	61.6	86.4	52.0
LLaVA-1.5 (Liu et al., 2023a)	88.2	53.8	7.5	21.3	72.5	45.3	22.3	35.4	6.7	60.0	40.8	69.6	3.8	6.4	17.9	26.9	36.1
Vicuna-1.3 (Chiang et al., 2023)	62.3	30.4	87.8	1.7	68.5	84.6	67.4	76.7	0.2	73.1	28.3	60.5	21.9	52.0	0.9	57.9	48.4
Falcon (Almazrouei et al., 2023b)	67.3	14.8	65.7	10.2	50.3	59.0	18.4	49.5	4.9	66.9	10.5	55.7	11.5	39.2	21.9	60.7	37.9
VDU models trained with ***only generated*** *data*																	
Flan-T5$_{\text{large}}$ + KD	36.6	23.0	21.7	2.3	89.5	64.5	90.6	76.1	20.7	61.4	31.4	68.7	34.8	74.4	79.2	61.5	52.3
Flan-T5$_{\text{large}}$ + **DocKD**	72.6	9.1	89.7	3.2	86.4	68.9	77.2	73.9	5.1	76.1	40.4	84.4	29.8	85.3	96.7	12.4	57.0
DocFormerv2$_{\text{large}}$ + KD	59.3	17.5	75.2	0.9	91.5	69.9	87.4	76.2	22.2	67.9	29.3	73.5	38.5	85.7	94.6	47.7	58.6
DocFormerv2$_{\text{large}}$ + **DocKD**	55.8	21.4	89.6	6.7	78.2	55.5	89.8	87.4	6.6	85.4	56.1	79.4	26.3	92.2	96.3	71.8	**62.4**

Table 12: RVL-CDIP classification results of all 16 categories.

```
CARDIOVASCULAR DISEASE 2.11.15-19 Hyperglycemia
Insulin Hyper a path that leads to increased risk
for MI Resistance Dys TYPE 2 DIABETES EQUALS PRIOR
MI AS A CHD RISK FACTOR Pr S 7-year incidence
of myocardial infarction (MI) (%) 25% 20% 15%
18.8% 20.2% 10% 5% 0% Nondiabetic patients Type
2 diabetics with prior MI without prior MI

Generate two question-answer pairs from this
document.

Question: Heading of the document?
Answer: Link between IR and CVD

Question: what does MI stand for?
Answer: myocardial infarction


Rules:
- Use the following test document as the only
source of information.
- Make questions diverse as possible.
- Answers should be simple and specified in the
document.
- Generate ONLY questions and answers, do not give
any explanations.

[Test]

Document: {LINEARIZED_TEXT_PLACE_HOLDER}

Generate {COUNT_PLACE_HOLDER} question-answer pairs
from this document.
```

B.2 Generation Prompt for Entity Extraction

We separate the generation of entities and field names into two parts: for non-KV entities and for KV entities. For the former, the generation prompt $\mathbf{p}_{\text{gen-}ent}$ is employed to extract entities from the document text as well as assigning their names. This process is exemplified through two instructive examples. Provided with the document text, the LLM is instructed to extract entities enclosed with <regular> and </regular> tags. Also, each line of entity is delimited by a separator " -- ", followed by the corresponding generated field name. Note that, to avoid duplicated generations for KV entities, we remove all the detected KV entities from the document text: {TEXT_WITHOUT_KV_PLACE_HOLDER} (refer to Sec. 3.2).

For the KV entities identified by a KV detection model, $\mathbf{p}_{\text{gen-}kv}$ instructs the LLM to generate only the field names for these entities. In the OCR text, the KV entities are enclosed by the tags <kv> and </kv> to provide explicit guidance to the model regarding which part it should refer to. The iterative presentation of each KV entity, line by line, involves inputting each line into {CONSTRAINTS_PLACE_HOLDER} in the format of "<kv>key value</kv> -- ". The generated field name is then appended to the constraint for the next iteration.

$\mathbf{p}_{\text{gen-}ent}$ for entity generation

```
Task: I want to get entities and their entity
types from OCR text of documents.

OCR text1: Invoice us EK Packaging Goras Ice Cream
$ Kathwada GIDC EK Packaging Ahmedabad, Gujarat.

<regular entities for OCR text1>
1. <regular>EK Packaging</regular> -- Company
Name
2. <regular>Goras Ice Cream</regular> --
Customer Name
3. <regular>Kathwada GIDC</regular> -- Customer
Address
4. <regular>EK Packaging Ahmedabad,
Gujarat.</regular> -- Company Address


OCR text2: 1 REAL GANACHE 16,500 1 egg tart 13,000
1 pizza toast 16,000

<regular entities for OCR text2>
1. <regular>REAL GANACHE</regular> -- Item Name
2. <regular>16,500</regular> -- Item Price
3. <regular>egg tart</regular> -- Item Name
4. <regular>13,000</regular> -- Item Price
5. <regular>pizza toast</regular> -- Item Name
6. <regular>16,000</regular> -- Item Price
```

```
7.  <regular>1</regular> -- Item Quantity


OCR text3:  {TEXT_WITHOUT_KV_PLACE_HOLDER}

<regular entities for OCR text3>
1.  <regular>
```

$\mathbf{p}_{\text{gen-}\mathit{kv}}$ for KV entity generation

```
Task:  I want to get entities and their entity
types from OCR text of documents.

OCR text1:  Invoice us EK Packaging Goras Ice Cream
$ Kathwada GIDC <kv>Inv.  date 14-03-20</kv>
EK Packaging Ahmedabad, Gujarat.  <kv>Due
29-03-20</kv> <kv>Inv.  # 1248</kv>

<kv entities for OCR text1>
1.  <kv>Inv.  date 14-03-20</kv> -- Invoice Date
2.  <kv>Due 29-03-20</kv> -- Due Date
3.  <kv>Inv.  # 1248</kv> -- Invoice Number


OCR text2:  1 REAL GANACHE 16,500 1 egg tart 13,000
1 pizza toast 16,000 <kv>TOTAL 45,500</kv>
<kv>CASH 50,000</kv> <kv>CHANGE 4,500</kv>

<kv entities for OCR text2>
1.  <kv>TOTAL 45,500</kv> -- Total Amount
2.  <kv>CASH 50,000</kv> -- Payment Amount
3.  <kv>CHANGE 4,500</kv> -- Change


OCR text3:  {TEXT_WITH_KV_TAGS_PLACE_HOLDER}

<kv entities for OCR text3>
{CONSTRAINTS_PLACE_HOLDER}
```

B.3 Generation and Inference Prompts for Document Classification

In the document classification task, we need three distinct generation prompts designed for generating descriptions, positive labels list, and negative labels list, respectively. Initially, $\mathbf{p}_{\text{gen-}\mathit{desc}}$ prompts the LLM to generate a description by characterizing the document type based on the document text. Subsequently, the generated output $\mathbf{a}_{\text{gen-}\mathit{desc}}$ is incorporated into the following prompt, $\mathbf{p}_{\text{gen-}\mathit{pos}}$, specifically within the placeholder {DESCRIPTION_PLACE_HOLDER}. This serves the purpose of providing contextual information about the document, thereby facilitating the accurate generation of positive labels. Finally, the output $\mathbf{a}_{\text{gen-}\mathit{pos}}$ is introduced to {POSITIVES_PLACE_HOLDER} in the negative generation prompt $\mathbf{p}_{\text{gen-}\mathit{neg}}$. This instructs the LLM to avoid suggesting types similar to those in the positives list.

$\mathbf{p}_{\text{gen-}\mathit{desc}}$ for document description generation

```
Document:  {TEXT_PLACE_HOLDER}

Question:  Can you describe the document type of
the above document in one sentence?

Answer:
```

$\mathbf{p}_{\text{gen-}\mathit{pos}}$ for positive label generation

```
Text of the document:  {TEXT_PLACE_HOLDER}

Short description of the document:
{DESCRIPTION_PLACE_HOLDER}

Question:  Given the above text of a document and
its short description, can you suggest a list of
{COUNT_PLACE_HOLDER} possible types (or names) of
the document?  Please list only types, without any
explanation or description.

Answer:
```

$\mathbf{p}_{\text{gen-}\mathit{neg}}$ for negative label generation

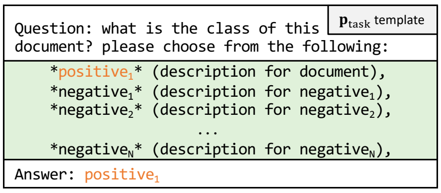
```
Document:  {TEXT_PLACE_HOLDER}

Matching types list:  {POSITIVES_PLACE_HOLDER}

Question:  Given the above text extracted from
a document using OCR, can you suggest a list of
{COUNT_PLACE_HOLDER} possible document types (or
names) that do NOT match the document?  Do not
include types similar to the matching list.

Answer:
```

For inference, we support open-world classification by dynamically constructing a candidate list in the prompt. We ask the model to select the class label that matches best with given document. Fig. 8 shows the prompt $\mathbf{p}_{\text{task}}$ we used in our experiment.



$\mathbf{p}_{\text{task}}$ template

Question: what is the class of this document? please choose from the following:

positive$_1$ (description for document),

negative$_1$ (description for negative$_1$),

negative$_2$ (description for negative$_2$),

...

negative$_N$ (description for negative$_N$),

Answer: positive$_1$

Figure 8: Classification task prompt template. The candidate list is composed of one positive label and a few negative labels, appended with descriptions.

B.4 Connectivity Between the Proposed Methods

In this study, tailoring generation prompts and document text formats for specific tasks has been proposed, and there is a potential for synergy when combining these approaches. However, the effectiveness of such combination depends on the chosen document knowledge injection method and the nature of the task. For instance, we observed that text linearization did not enhance classification accuracy and could not be transferred to entity extraction, as the field name generation also involves distinct modifications to $\mathbf{d}_{\text{text}}$ (refer to Appx. B.2). On the other hand, leveraging document descriptions or reasoning steps may hold promise for improving the QA generation. Yet, this would require nontrivial efforts in designing new generative prompts, and it is identified as a prospective direction for future research.

B.5 Improving the Instructions for LLM Zero-Shot Prediction

While numerous strategies exist for enhancing LLM zero-shot predictions through instruction modulation, the optimal approach varies depending on the model type. Although we have not explored optimal instruction strategies for every language model, our work involves minimal engineering efforts to identify the LLM's performance in document understanding tasks and show that small student models trained by DocKD are as effective as the LLMs. In this section, we describe our enhancements to the prompt for improving zero-shot predictions of Claude-2 and Falcon-40B models, in document VQA and classification tasks. Essentially, we provide the LLM with $\mathbf{p}_{\text{task}}$ and $\mathbf{d}_{\text{text}}$ as inputs, employing the same design as utilized for the student models. Within $\mathbf{p}_{\text{task}}$, we input instructions to regulate the output format for each LLM, facilitating the parsing of the answer into the desired format.

Instructions for DocVQA. We leverage linearized OCR text, a method previously employed in generating QA pairs from the LLM. Given the LLM's ability in comprehending linearized text, we convert the OCR text into the linearized form and ask the document question. In addition, since DocVQA is an extractive QA dataset, *i.e.*, answers are directly extracted from the provided context, we use the dataset-specific prompt to control the outputs. To achieve this, we implement instructing rules as suggested in (Wang et al., 2023b). This strategy has significantly increased DocVQA val ANLS to 58.3 → 79.6 for Claude-2, and 52.6 → 72.4 for Falcon-40B. In summary, the task prompt for DocVQA is provided as follows.

$\mathbf{p}_{\text{task}}$ for DocVQA zero-shot prediction

```
You are asked to answer the question based on the
given document OCR text.

For example,
Context: Confidential RJRT PR APPROVAL DATE:
1/8/93 SUBJECT: Ru IVAs PROPOSED RELEASE DATE: for
response FOR RELEASE TO: CONTACT: P. CARTER ROUTE
TO: Name Initials Date Peggy Carter Ace 1/1/15
Kaura Payne nt. T/R Return to Peggy Carter, PR, 16
Reynolds Building Not
Answer the question: What is the contact person
name mentioned in this letter?
Answer: P. Carter

Rules:
- The answers to questions are short text spans
taken verbatim from the document. This means
that the answers comprise a set of contiguous text
tokens present in the document.
- Directly extract the answer of the question from
the document with as few words as possible.


Context: {LINEARIZED_TEXT_PLACE_HOLDER}
Answer the question: {QUESTION_PLACE_HOLDER}
Answer:
```

Instructions for RVL-CDIP. Recognizing the significance of document descriptions in enhancing knowledge utilization and improving class label generation, we adopt a 2-step classification approach. In the initial step, the LLM does not classify directly but instead generates the possible document type according to its own interpretation. Subsequently, in the second step, we provide the output from the first step into {TYPE_PLACE_HOLDER} as a suggested document name, and instruct the model to select the document type from the candidate list. In addition, we recognize that Falcon-40B struggles in accurately naming the exact category, even when provided with a list. To address this, we emphasize all 16 evaluation categories. This strategic modulation has improved RVL-CDIP test mAcc to 31.8 → 37.9 for Falcon-40B, compared to direct classification. However, Claude-2 does not achieve further performance gain through this instruction. Additionally, attempts to replace the document text with linearized text, as done in DocVQA, do not yield improvements in this task.

$\mathbf{p}_{\text{task}}$ for RVL-CDIP zero-shot prediction

```
Choose the document type based on the given context.
We have 16 categories.

- letter
- form
- email
- handwritten
- advertisement
- scientific report
- scientific publication
- specification
- file folder
- news article
- budget
- invoice
- presentation
- questionnaire
- resume
- memo
```

```
Context: {TEXT_PLACE_HOLDER}
Suggested document name: {TYPE_PLACE_HOLDER}
Question: What is the document type of this
document? Please choose from the following:
{letter; form; email; handwritten; advertisement;
scientific report; scientific publication;
specification; file folder; news article; budget;
invoice; presentation; questionnaire; resume; memo}
Answer:
```

C Examples of Generated Annotations

We present the examples of LLM-generated annotations, for document VQA in Appx. C.1, for entity extraction in Appx. C.2, and for document classification in Appx. C.3.

C.1 Generated QAs for Document VQA

Using raw OCR text vs. linearized OCR text. Table 13 and Table 14 describe the generated QAs from Claude-2, comparing the results from the plain KD (using raw OCR text) and DocKD (using linearized OCR text). In Table 13, the document includes line numbers for each line of text, but raw OCR text lacks this structural detail, resulting in misplaced numbers in the middle of text. Consequently, Claude-2 generates inaccurate questions, such as `Question 1` erroneously referencing a non-existent question number 2, or `Question 2` inquiring about the percentage of children, which cannot be directly answered from the document. In contrast, when linearized OCR text is utilized, questions align with the document context, ensuring correct answers. Notably, questions explicitly refer to line numbers, *e.g.*, inquiring about the contents in `line 1` or in `lines 5-8`, which requires visual knowledge to answer.

In Table 14, the document contains words and numbers in a structured form, posing a challenge for the LLM in generating informative QAs from the OCR text. In KD QAs, `Question 1` and `Question 3` are easily extracted and straightforward to answer without visual knowledge. `Question 2`, which pertains to tabular information, is paired with `Answer 2`, which is incorrect. In contrast, `Question 2` of DocKD requires reference to the table format, specifically in the third row and the second column, for a correct response. Also, the paired `Answer 2` is correct. Similarly, `Question 3` and `Answer 3` are about the contents in the second row and the last column of the table.

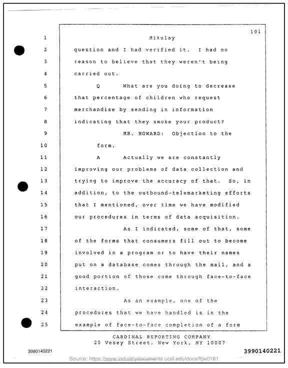
LLM teachers: Falcon-40B vs. Falcon-180B vs. Claude-2. Table 15 and Table 16 describe the generated QAs from different teacher models, using Falcon-40B, Falcon-180B, and Claude-2. Every teacher utilizes the linaerized OCR text. The target document in Table 15 corresponds to the one used in Table 13, and the document for Table 16 corresponds to the one used in Table 14. While Claude-2 adeptly incorporates layout knowledge into QA generation, Falcon-40B tends to produce simple questions and answers, occasionally resulting in duplicates or only slight variations. In contrast, the Falcon-180B model better generates diverse QA pairs, and they are mostly accurate. The primary distinction from Claude-2 lies in the observation that Claude-2 is more inclined to explicitly mention layout information in the document.

2-step generation of Q → A. In QA generation for the document VQA task, we have directed the LLM to simultaneously produce both questions and answers. This approach aims to ensure consistency with the document contents and establish more accurate relationship between the generated question and its corresponding answer. Alternatively, we explore a 2-step generation process where the LLM initially generates a list of questions and subsequently provides answers for them.

Table 17 and Table 18 delineate questions and answers generated by Claude-2, comparing the two distinct generation schemes: 2-step generation and QA simultaneous generation. In Table 17, the target document features a table with limited extractable information. During the first step of question generation, Claude-2 manages to produce questions related to the table headers or the index, yet these remain challenging to answer based on the text. As result, the second step generates random number answers. Conversely, QA pair simultaneous generation yields better questions and answers, effectively leveraging structural information, *e.g.*, column headers or numbers and ratios listed in the table, and creating easy-to-answer questions from them.

Similar observations are found in Table 18, where the document contains a plot and there is not much information other than the header, axes, and axis labels. In the 2-step generation, questions are formulated regarding the efficiency and percentage of the filtraion, which cannot be addressed using the available document content. The resulting answers include phrases like "`not mentioned`" or "`not provided`". Conversely, QA pair generation produces questions that are easily answerable.

Document image (ID: ftjw0181)



101

Mikulay

question and I had verified it. I had no reason to believe that they weren't being carried out.

Q What are you doing to decrease that percentage of children who request merchandise by sending in information indicating that they smoke your product?

MR. HOWARD: Objection to the form.

A Actually we are constantly improving our problems of data collection and trying to improve the accuracy of that. So, in addition, to the outbound-telemarketing efforts that I mentioned, over time we have modified our procedures in terms of data acquisition.

As I indicated, some of that, some of the forms that consumers fill out to become involved in a program or to have their names put on a database comes through the mail, and a good portion of those come through face-to-face interaction.

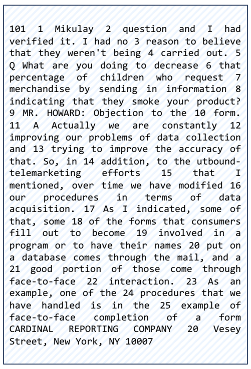
As an example, one of the procedures that we have handled is in the example of face-to-face completion of a form

CARDINAL REPORTING COMPANY
20 Vesey Street, New York, NY 10007

3990140221 3990140221

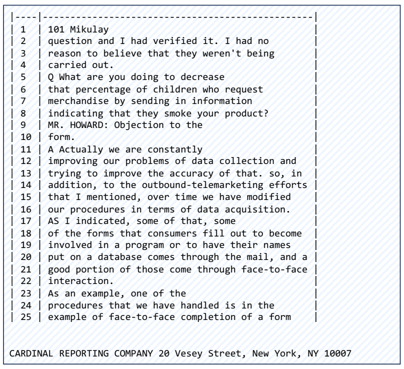
Source: https://www.industrydocuments.ucsf.edu/docs/ftjw0181

OCR text



101 1 Mikulay 2 question and I had verified it. I had no 3 reason to believe that they weren't being 4 carried out. 5 Q What are you doing to decrease 6 that percentage of children who request 7 merchandise by sending in information 8 indicating that they smoke your product? 9 MR. HOWARD: Objection to the 10 form. 11 A Actually we are constantly 12 improving our problems of data collection and 13 trying to improve the accuracy of that. So, in 14 addition, to the utbound-telemarketing efforts 15 that I mentioned, over time we have modified 16 our procedures in terms of data acquisition. 17 As I indicated, some of that, some 18 of the forms that consumers fill out to become 19 involved in a program or to have their names 20 put on a database comes through the mail, and a 21 good portion of those come through face-to-face 22 interaction. 23 As an example, one of the 24 procedures that we have handled is in the 25 example of face-to-face completion of a form CARDINAL REPORTING COMPANY 20 Vesey Street, New York, NY 10007

Linearized OCR text



```
|----|-------------------------------------------------|
| 1  | 101 Mikulay                                     |
| 2  | question and I had verified it. I had no        |
| 3  | reason to believe that they weren't being       |
| 4  | carried out.                                    |
| 5  | Q What are you doing to decrease                |
| 6  | that percentage of children who request         |
| 7  | merchandise by sending in information           |
| 8  | indicating that they smoke your product?        |
| 9  | MR. HOWARD: Objection to the                    |
| 10 | form.                                           |
| 11 | A Actually we are constantly                    |
| 12 | improving our problems of data collection and   |
| 13 | trying to improve the accuracy of that. so, in  |
| 14 | addition, to the outbound-telemarketing efforts |
| 15 | that I mentioned, over time we have modified    |
| 16 | our procedures in terms of data acquisition.    |
| 17 | AS I indicated, some of that, some              |
| 18 | of the forms that consumers fill out to become  |
| 19 | involved in a program or to have their names    |
| 20 | put on a database comes through the mail, and a |
| 21 | good portion of those come through face-to-face |
| 22 | interaction.                                    |
| 23 | As an example, one of the                       |
| 24 | procedures that we have handled is in the       |
| 25 | example of face-to-face completion of a form    |


CARDINAL REPORTING COMPANY 20 Vesey Street, New York, NY 10007
```

KD QAs (using OCR text)

Question 1: What is the question number 2 referring to in the document?
Answer 1: Mikulay

Question 2: What percentage of children request merchandise by sending in information indicating that they smoke the company's product?
Answer 2: 101

Question 3: What are some of the ways the company collects data mentioned in the document?
Answer 3: Through outbound telemarketing efforts, mail, and face-to-face interaction.

DocKD QAs (using linearized OCR text)

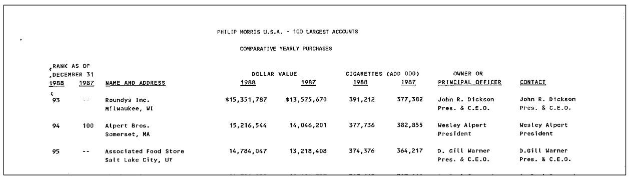
Question 1: What is the name mentioned in line 1?
Answer 1: Mikulay

Question 2: What does the questioner ask about in lines 5-8?
Answer 2: What the person is doing to decrease the percentage of children who request merchandise by sending in information indicating that they smoke the product.

Question 3: What does the respondent say they are constantly improving in lines 11-12?
Answer 3: Their problems of data collection.

Table 13: Generated QA pairs for the IDL document (ID: ftjw0181).

Document image (ID: gtbw0041)

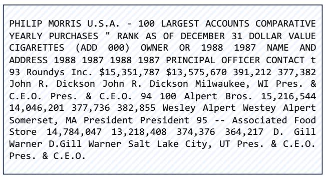


PHILIP MORRIS U.S.A. - 100 LARGEST ACCOUNTS

COMPARATIVE YEARLY PURCHASES

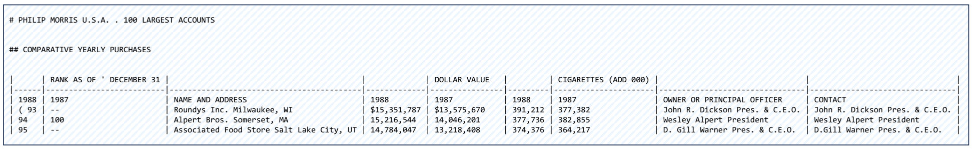
RANK AS OF DECEMBER 31			DOLLAR VALUE		CIGARETTES (ADD 000)		OWNER OR	
1988	1987	NAME AND ADDRESS	1988	1987	1988	1987	PRINCIPAL OFFICER	CONTACT
93	--	Roundys Inc. Milwaukee, WI	$15,351,787	$13,575,670	391,212	377,382	John R. Dickson Pres. & C.E.O.	John R. Dickson Pres. & C.E.O.
94	100	Alpert Bros. Somerset, MA	15,216,544	14,046,201	377,736	382,855	Wesley Alpert President	Wesley Alpert President
95	--	Associated Food Store Salt Lake City, UT	14,784,047	13,218,408	374,376	364,217	D. Gill Warner Pres. & C.E.O.	D.Gill Warner Pres. & C.E.O.

OCR text



PHILIP MORRIS U.S.A. - 100 LARGEST ACCOUNTS COMPARATIVE YEARLY PURCHASES " RANK AS OF DECEMBER 31 DOLLAR VALUE CIGARETTES (ADD 000) OWNER OR 1988 1987 NAME AND ADDRESS 1988 1987 1988 1987 PRINCIPAL OFFICER CONTACT t 93 Roundys Inc. $15,351,787 $13,575,670 391,212 377,382 John R. Dickson John R. Dickson Milwaukee, WI Pres. & C.E.O. Pres. & C.E.O. 94 100 Alpert Bros. 15,216,544 14,046,201 377,736 382,855 Wesley Alpert Westey Alpert Somerset, MA President President 95 -- Associated Food Store 14,784,047 13,218,408 374,376 364,217 D. Gill Warner D.Gill Warner Salt Lake City, UT Pres. & C.E.O. Pres. & C.E.O.

Linearized OCR text



```
# PHILIP MORRIS U.S.A. . 100 LARGEST ACCOUNTS


## COMPARATIVE YEARLY PURCHASES


|      | RANK AS OF ' DECEMBER 31 |                                        |             | DOLLAR VALUE |         | CIGARETTES (ADD 000) |                                |                                |
|------|--------------------------|----------------------------------------|-------------|--------------|---------|----------------------|--------------------------------|--------------------------------|
| 1988 | 1987                     | NAME AND ADDRESS                       | 1988        | 1987         | 1988    | 1987                 | OWNER OR PRINCIPAL OFFICER     | CONTACT                        |
| ( 93 | --                       | Roundys Inc. Milwaukee, WI             | $15,351,787 | $13,575,670  | 391,212 | 377,382              | John R. Dickson Pres. & C.E.O. | John R. Dickson Pres. & C.E.O. |
| 94   | 100                      | Alpert Bros. Somerset, MA              | 15,216,544  | 14,046,201   | 377,736 | 382,855              | Wesley Alpert President        | Wesley Alpert President        |
| 95   | --                       | Associated Food Store Salt Lake City, UT | 14,784,047 | 13,218,408  | 374,376 | 364,217              | D. Gill Warner Pres. & C.E.O.  | D.Gill Warner Pres. & C.E.O.   |
```

KD QAs (using OCR text)

Question 1: What is the name of the document?
Answer 1: PHILIP MORRIS U.S.A. - 100 LARGEST ACCOUNTS COMPARATIVE YEARLY PURCHASES

Question 2: What are the two columns after the name and address?
Answer 2: OWNER OR PRINCIPAL OFFICER and CONTACT

Question 3: What is the address of Roundys Inc?
Answer 3: Milwaukee, WI

DocKD QAs (using linearized OCR text)

Question 1: What is the title of this document?
Answer 1: # PHILIP MORRIS U.S.A.. 100 LARGEST ACCOUNTS

Question 2: What is the name and address of the 94th ranked company in 1988?
Answer 2: Alpert Bros. Somerset, MA

Question 3: Who is the contact person for Roundys Inc.?
Answer 3: John R. Dickson Pres. & C.E.O.

Table 14: Generated QA pairs for the IDL document (ID: gtbw0041).

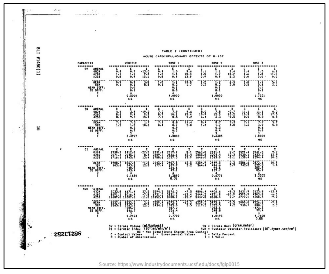
Falcon-40B-generated QAs
Question 1: What is the name of the person being questioned? Answer 1: Mr. Howard Question 2: What is the name of the person being questioned? Answer 2: Mr. Howard Question 3: What is the name of the person being questioned? Answer 3: Mr. Howard
Falcon-180B-generated QAs
Question 1: What is the name of the person mentioned in the document? Answer 1: Mikulay Question 2: What is the objection mentioned in the document? Answer 2: Objection to the form. Question 3: What is the purpose of outbound telemarketing efforts mentioned in the document? Answer 3: To improve the accuracy of data collection.
Claude-2-generated QAs
Question 1: What is the name mentioned in line 1? Answer 1: Mikulay Question 2: What does the questioner ask about in lines 5-8? Answer 2: What the person is doing to decrease the percentage of children who request merchandise by sending in information indicating that they smoke the product. Question 3: What does the respondent say they are constantly improving in lines 11-12? Answer 3: Their problems of data collection.

Table 15: Generated QAs from different teacher models. The IDL document ftjw0181 in Table 13 is provided as the linearized text form.

Falcon-40B-generated QAs
Question 1: What is the name of the company that is ranked 94th in the list? Answer 1: Alpert Bros. Question 2: What is the name of the company that is ranked 95th in the list? Answer 2: Associated Food Store
Falcon-180B-generated QAs
Question 1: What is the name of the company ranked 93rd in 1988? Answer 1: Roundys Inc. Question 2: What is the name of the contact person for Roundys Inc. in 1988? Answer 2: John R. Dickson
Claude-2-generated QAs
Question 1: What is the title of this document? Answer 1: # PHILIP MORRIS U.S.A.. 100 LARGEST ACCOUNTS Question 2: What is the name and address of the 94th ranked company in 1988? Answer 2: Alpert Bros. Somerset, MA Question 3: Who is the contact person for Roundys Inc.? Answer 3: John R. Dickson Pres. & C.E.O.

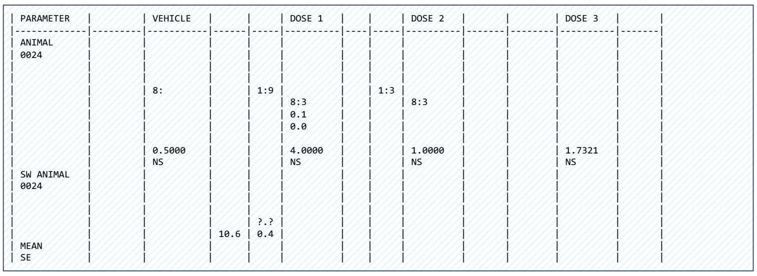
Table 16: Generated QAs from different teacher models. The IDL document gtbw0041 in Table 14 is provided as the linearized text form. If the last answer surpasses the maximum generation sequence length, the resulting QA pairs consist of only the first two elements.

Document image (ID: fglp0015)



Source: https://www.industrydocuments.ucsf.edu/docs/fglp0015

Linearized OCR text



```
| PARAMETER    |          | VEHICLE    |       |      | DOSE 1    |     |      | DOSE 2    |        |         | DOSE 3    |        |
|--------------|----------|------------|-------|------|-----------|-----|------|-----------|--------|---------|-----------|--------|
| ANIMAL       |          |            |       |      |           |     |      |           |        |         |           |        |
| 0024         |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          | 8:         |       | 1:9  |           |     | 1:3  |           |        |         |           |        |
|              |          |            |       |      | 8:3       |     |      | 8:3       |        |         |           |        |
|              |          |            |       |      | 0.1       |     |      |           |        |         |           |        |
|              |          |            |       |      | 0.0       |     |      |           |        |         |           |        |
|              |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          | 0.5000     |       |      | 4.0000    |     |      | 1.0000    |        |         | 1.7321    |        |
|              |          | NS         |       |      | NS        |     |      | NS        |        |         | NS        |        |
| SW ANIMAL    |          |            |       |      |           |     |      |           |        |         |           |        |
| 0024         |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          |            |       |      |           |     |      |           |        |         |           |        |
|              |          |            |       | ?.?  |           |     |      |           |        |         |           |        |
|              |          |            |  10.6 | 0.4  |           |     |      |           |        |         |           |        |
| MEAN         |          |            |       |      |           |     |      |           |        |         |           |        |
| SE           |          |            |       |      |           |     |      |           |        |         |           |        |
```

Q → A generation

Question 1: What is the parameter mentioned in this document?
Question 2: What is the animal mentioned in this document?
Question 3: What is the dose mentioned in this document?

Answer 1: 8:
Answer 2: 0024
Answer 3: 8:

QA pair generation

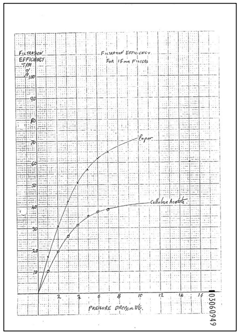
Question 1: What are the column headers in the table?
Answer 1: PARAMETETR, VEHICLE, DOSE 1, DOSE 2, DOSE 3

Question 2: What is the animal number listed in the table?
Answer 2: 0024

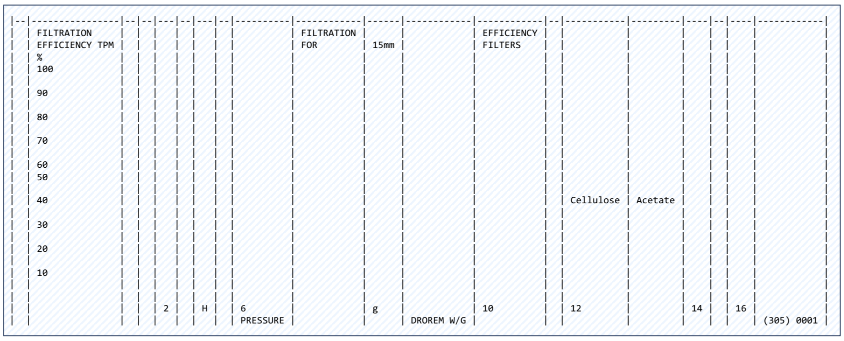
Question 3: What are the dose ratios listed for Dose 1 and Dose 2?
Answer 3: 1:9 and 1:3

Table 17: Generated questions and answers for the IDL document (ID: fglp0015).

Document image (ID: gyfm0207)


Filtration Efficiency TPM %
Filtration efficiency for 15mm Filters
Paper
Cellulose Acetate
Pressure Drop mm WG

Linearized OCR text



```
|--|----------------|--|--|---|--|---|--|----------|------------|------|------------|------------|--|-----------|---------|----|--|----|------------|
|  | FILTRATION     |  |  |   |  |   |  |          | FILTRATION |      |            | EFFICIENCY |  |           |         |    |  |    |            |
|  | EFFICIENCY TPM |  |  |   |  |   |  |          | FOR        | 15mm |            | FILTERS    |  |           |         |    |  |    |            |
|  | %              |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 100            |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 90             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 80             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 70             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 60             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 50             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 40             |  |  |   |  |   |  |          |            |      |            |            |  | Cellulose | Acetate |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 30             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 20             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  | 10             |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  |   |  |   |  |          |            |      |            |            |  |           |         |    |  |    |            |
|  |                |  |  | 2 |  | H |  | 6        |            | g    |            | 10         |  | 12        |         | 14 |  | 16 |            |
|  |                |  |  |   |  |   |  | PRESSURE |            |      | DROREM W/G |            |  |           |         |    |  |    | (305) 0001 |
```

Q → A generation

Question 1: What is the efficiency of the filtration?
Question 2: What is the efficiency of the TPM?
Question 3: What is the percentage of the filtration?

Answer 1: The efficiency of the filtration is not explicitly mentioned in the given context.
Answer 2: The efficiency of the TPM is not mentioned in the given context.
Answer 3: The answer to the question is not provided in the given context.

QA pair generation

Question 1: What is the heading of the document?
Answer 1: FILTRATION EFFICIENCY

Question 2: What is the abbreviation for the filtration efficiency?
Answer 2: TPM

Question 3: What is the filter size mentioned in the document?
Answer 3: 15mm

Table 18: Generated questions and answers for the IDL document (ID: gyfm0207).

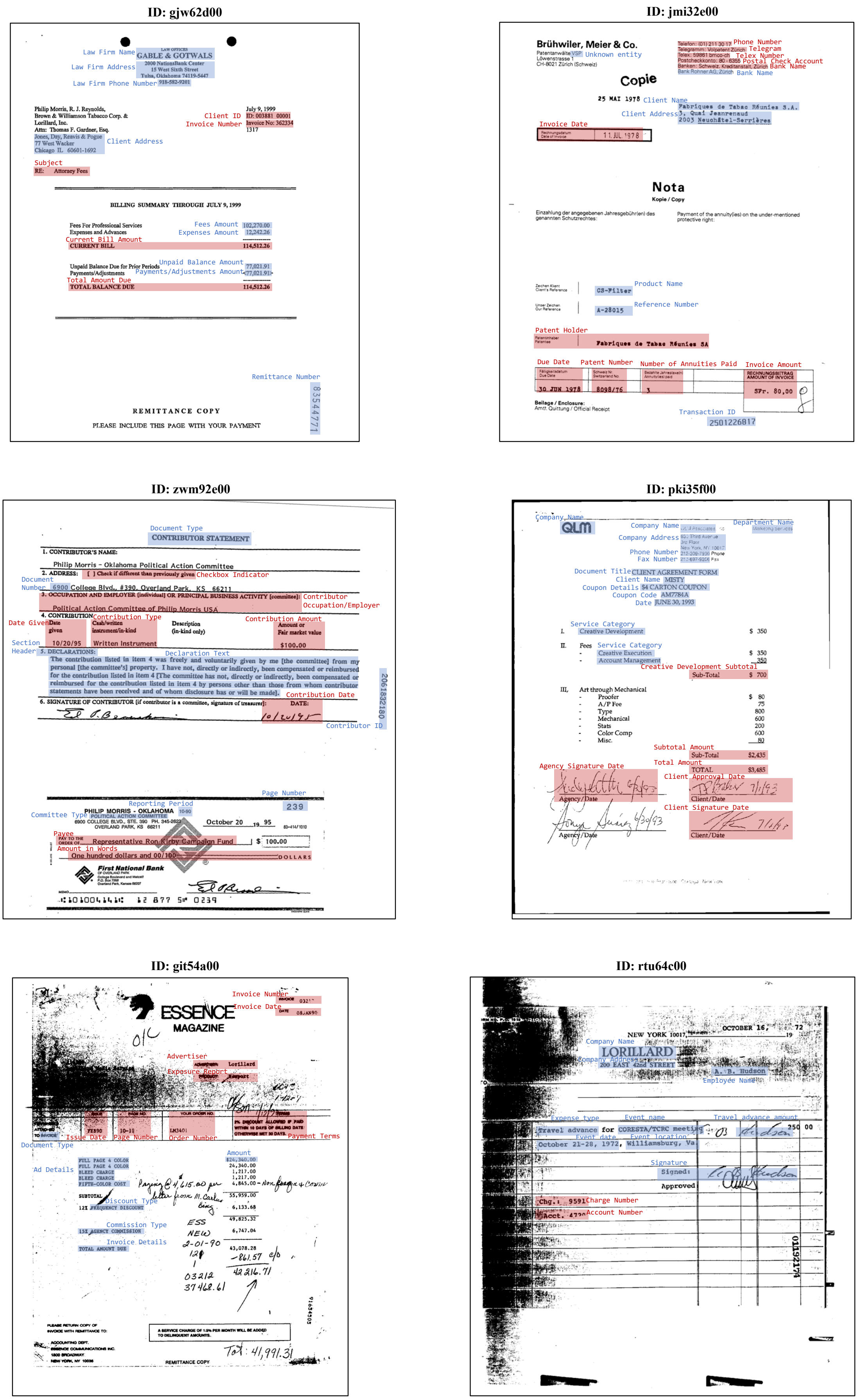
C.2 Generated Entities and Fields for Entity Extraction

Fig. 9 displays the generated entities and fields for the RVL-CDIP (Harley et al., 2015) invoice documents. Similar to Fig. 4 in the main paper, non-KV entities and their respective field names are represented by blue boxes and text, while detected KV entities and their corresponding field names are denoted by red boxes and text. It includes an example where the document is non-English (id: jmi32e00); surprisingly, leveraging the multilingual capability of the LLM, informative entities are extracted and field names are generated in English. Throughout the examples in Fig. 9, a diverse range of field names is observed.

Upon generating entities and fields, an aggregation process is employed prior to training the student model. There exist multiple entities within a single document sharing the same field name. We group these entities under the shared field, so that the student model can be trained to match the field to every entity in the group. Specifically, we gather all generated field-entity pairs $\{(\mathbf{f}_1, \mathbf{e}_1), (\mathbf{f}_2, \mathbf{e}_2), \dots\}$ and identify the entity group for each field $\mathbf{f}$, $\{\mathbf{e}_j\}$ for all j such that $\mathbf{f}_j = \mathbf{f}$. Consequently, $\mathbf{f}$ is incorporated into $\mathbf{p}_{\text{task}}$, and $\{\mathbf{e}_j\}$ is included in $\mathbf{a}_{\text{task}}$.

C.3 Generated Class Labels for Document Classification

Fig. 10 illustrates the generated description, positive class labels, and negative class labels for each IDL (Lewis et al., 2006) document. The results demonstrate that the LLM generates broad spectrum of class candidates, including report, email, business plan, to-do list, brochure, recipe, poetry, etc. This diversity enables the open document classification capabilities of student models.


ID: gjw62d00

ID: jmi32e00

ID: zwm92e00

ID: pki35f00

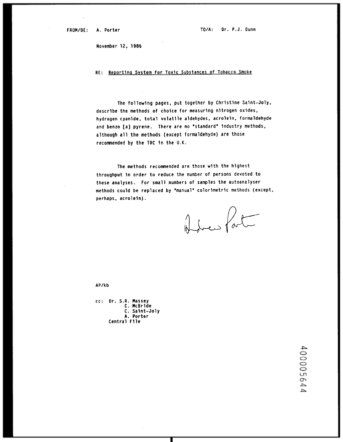
ID: git54a00

ID: rtu64c00

Law Firm Name
Law Firm Address
Law Firm Phone Number
Client ID
Invoice Number
Client Address
Subject
Fees Amount
Expenses Amount
Current Bill Amount
Unpaid Balance Amount
Payments/Adjustments Amount
Total Amount Due
Remittance Number
REMITTANCE COPY
PLEASE INCLUDE THIS PAGE WITH YOUR PAYMENT
Phone Number
Telegram
Telex Number
Postal Check Account
Bank Name
Unknown entity
Copie
Client Name
Client Address
Invoice Date
Nota
Product Name
Reference Number
Patent Holder
Due Date
Patent Number
Number of Annuities Paid
Invoice Amount
Transaction ID
Document Type
CONTRIBUTOR STATEMENT
Checkbox Indicator
Document Number
Contributor Occupation/Employer
Contribution Type
Contribution Amount
Date Given
Section Header
Declaration Text
Contribution Date
Contributor ID
Page Number
Reporting Period
Committee Type
Payee
Amount in Words
Company Name
Company Address
Department Name
Phone Number
Fax Number
Document Title
Client Name
Coupon Details
Coupon Code
Date
Service Category
Creative Development Subtotal
Subtotal Amount
Total Amount
Agency Signature Date
Client Approval Date
Client Signature Date
Invoice Number
Invoice Date
Advertiser
Exposure Report
Document Type
Issue Date
Page Number
Order Number
Payment Terms
Amount
Ad Details
Discount Type
Commission Type
Invoice Details
Company Name
Company Address
Employee Name
Expense type
Event name
Travel advance amount
Event date
Event location
Signature
Charge Number
Account Number

Figure 9: Generated entities and fields for RVL-CDIP invoice documents.

ID: fjmw0199

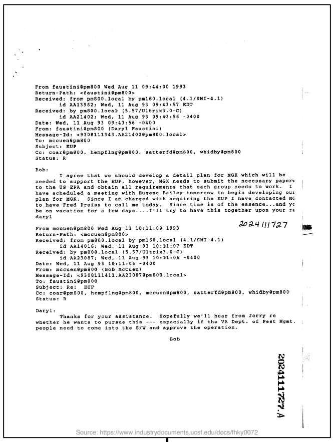


description:
An internal memo from A. Porter to Dr. P.J. Dunn on reporting methods for measuring toxic substances in tobacco smoke.

positives:
- Technical report
- Laboratory methods memo
- Research methods proposal

negatives:
- Textbook chapter
- Instruction manual
- Magazine feature
- Poetry
- Encyclopedia entry
- Novel excerpt
- Drama script
- Financial statement
- Newspaper article
- Short story

ID: fhky0072

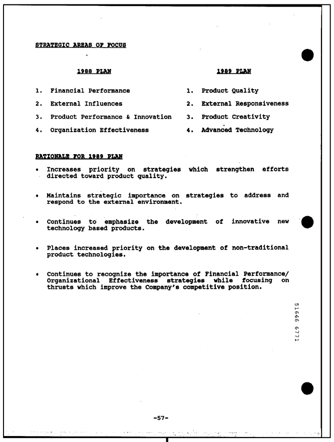


description:
An email chain discussing plans and coordination around obtaining an Experimental Use Permit (EUP) for a pesticide product.

positives:
- Email thread
- Internal correspondence
- Business communication

negatives:
- News article
- Fiction story
- Financial report
- Instruction manual
- Poetry
- Legal contract
- Technical specifications
- Personal diary
- Academic research paper
- Biography

ID: fhjf0077

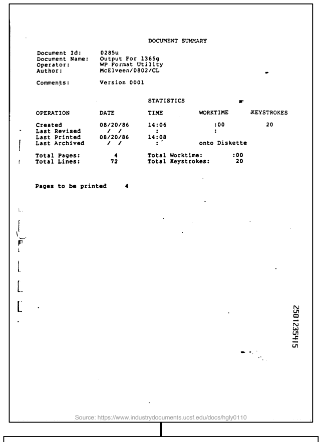


description:
An overview of a company's strategic areas of focus for two consecutive years, showing changes in priorities from one year to the next.

positives:
- Strategic plan
- Annual business plan
- Corporate strategy memo

negatives:
- Product specifications
- Budget proposal
- Meeting minutes
- Policy manual
- Employee handbook
- Marketing plan
- Financial statements
- Sales report
- Invoice
- Contract

ID: hgly0110

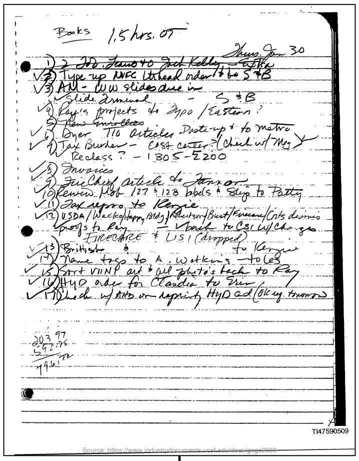


description:
A summary document that provides statistics and metadata about another document with ID 0285u, including when it was created, revised, printed, and archived, as well as the number of pages, lines, keystrokes, and total work time.

positives:
- Document statistics report
- Output summary
- Metadata record

negatives:
- Recipe
- Budget spreadsheet
- Meeting agenda
- Lab report
- Press release
- Resume
- Research paper
- Product brochure
- Email
- Invoice

ID: gngc0030

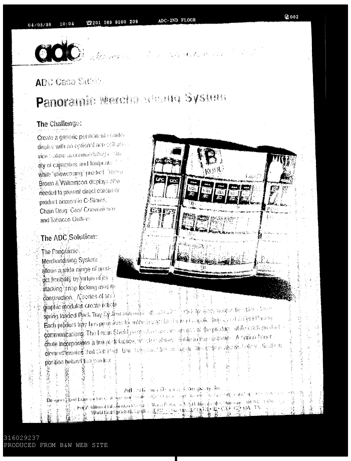


description:
A to-do list or set of notes outlining various tasks and deadlines.

positives:
- Meeting notes
- Task list
- To-do list

negatives:
- Journal article
- Short story
- Letter to teacher
- Lab report
- Shopping list
- News article
- Email
- Diary entry
- Financial report
- Recipe

ID: glxy0078



description:
Descriptions and specifications for a retail product display system.

positives:
- Product brochure
- Product specifications
- Retail display proposal

negatives:
- Medical records
- School transcript
- Wedding invitation
- Novel excerpt
- Recipe
- Financial statement
- Newspaper article
- Tax returns
- Meeting agenda
- Employee handbook

Figure 10: Generated description and class labels for the IDL documents.

D Dataset Specifications

We provide additional information on the datasets that were not fully described in the main paper.

Evaluation datasets. In the document VQA task, we use DocVQA (Mathew et al., 2021) as an evaluation dataset. The DocVQA validation set contains manually annotated 5.3K questions related to the real-world industrial documents. For metrics, we use ANLS (average normalized Levenshtein similarity) (Biten et al., 2019) and EM (exact match) which checks if the predicted answer's characters exactly match those of the ground truth.

For the entity extraction, we use two evaluation datasets, CORD (Park et al., 2019) and DeepForm (Borchmann et al., 2021), a collection of restaurant receipts and invoices for political TV ads, respectively. The model should extract entities for the field such as <*menu name*> or <*total cashprice*> for CORD, and <*advertiser*> or <*flight to*> for DeepForm. The CORD test set is evaluated by entity-level F1 score, while the DeepForm test set is evaluated by ANLS since DeepForm's ground-truth entities are re-formatted from the original document text.

In the classification task, we use RVL-CDIP (Harley et al., 2015) test set, where 40K documents are labeled into 16 categories, including letter, memo, invoice, form, etc. The performance is measured by the mean accuracy of these 16 categories, while mAcc* measures the mean accuracy excluding four ambiguous categories: memo, file-folder, handwritten, and presentation.

Open-set classification. In Sec. 4.3, we have used three out-of-domain datasets for the open-set classification. Here, we outline their setups. (*i*) RVL-O (Larson et al., 2022) has documents that do not belong to any of 16 categories of RVL-CDIP. These outliers should be classified (or detected) as *other*, with the RVL-CDIP labels also given as candidates. (*ii*) For IRS-50, we collect 50 types of forms, instructions, and publications from the US Internal Revenue Service.[3] (*iii*) WikiDoc (Fujinuma et al., 2023) consists of 33K Wikipedia screenshots on 111 different subjects.

Table 19 presents a summary of the 50 IRS class labels which were used in Table 5. Each class label corresponds to one document sample sourced from the US Internal Revenue Service. We also present the precdiction results from Falcon-40B (zero-shot) and DocFormerv2$_{\text{base}}$ (DocKD).

WikiDoc categories. The WikiDoc dataset, as described in Fujinuma et al. (2023), comprises 111 diverse categories. For each category, the dataset includes screenshots of Wikipedia articles, encompassing a wide range of subjects. Examples of categories in the dataset inlcude Album, BasketballTeam, Cardinal, Dam, Economist, Fish, Glacier, Historian, IceHockeyLeague, Journalist, Lighthouse, Magazine, Noble, OfficeHolder, Poem, Racecourse, School, TradeUnion, University, Volcano, and WrestlingEvent.

DUDE single-page QAs. Throughout this paper, our primary focus was on training the student model using single-page document annotations, *i.e.*, document annotation is derived from the contents in a single page. There are document datasets annotated with multi-page information, such as DUDE (Borchmann et al., 2021) that is employed for the document VQA task in Table 4. In this case, we only used the QA annotations that can be addressed within a single page.

[3] https://www.irs.gov/forms-instructions

GT label	Falcon-40B prediction	DFv2$_{\text{base}}$ S+U prediction
Form 1000	Form 1000	Form 1000
Form 1040 (Schedule A)	Form 1040 (Schedule A)	Form W-2
Form 1040 (Schedule B)	Form 1040 (Schedule B)	Form W-2
Form 1040 (Schedule 1)	Form 1040 (Schedule 1)	Form W-2
Form 1040 (Schedule 2)	Tax form	Form W-2
Form 1040-NR (Schedule NEC)	Form 1040-NR (Schedule NEC)	Form 1040-NR (Schedule NEC)
Form 1040-NR (Schedule OI)	NULL	Form 1040-NR
Form 1040-X	Tax form	Form 1040-X
Form 1098-C	Form 1098-C	Form 1098-C
Form 1098-E	Form 1098-E	Form 1098-E
Form 1098-MA	Form 1098-MA	Form 1098-MA
Form 1098-Q	Form 1098-Q	Form 1098-Q
Form 4506	Form 4506	Form 4506
Form 4506-T	Tax form	Form 4506-T
Form 4852	Form 4852	Form 4852
Form 8994	Form	Form 8994
Form 9779	Form	Form 9779
Form 9783	Form 1000	Form 9783
Form 15103	Form 15103	Form 15103
Form W-2	Form W-2	Form W-2
Form W-2AS	Form W-2AS	Form W-2AS
Form W-2C	Form W-2C	Form W-2C
Form W-2G	Form W-2G	Form W-2G
Form W-3	Form W-3	Form W-2
Form W-3C	Form W-2C	Form W-2C
Form W-3SS	Form W-3SS	Form W-2AS
Form W-4	Form 1040 (Schedule 1)	Form W-4
Form W-4P	Form W-4P	Form W-4P
Form W-4R	Form 1040 (Schedule 1)	Form W-4R
Form W-4S	Form W-4S	Form W-4S
Form W-7	Form W-7	Form W-7
Form W-7A	Form W-7A	Form W-7A
Instruction 1040 (Schedule A)	Form 1040 (Schedule A)	Instruction 1040 (Schedule A)
Instruction 1040 (Schedule B)	Form 1040 (Schedule B)	Notice 1016
Instruction 1040-NR	Form	Instruction 1040-NR
Instruction 1098-Q	Instruction 1098-Q	Instruction 1098-Q
Instruction 8994	Form 8994	Instruction 8994
Notice 1015	Form 1000	Notice 1015
Notice 1016	Notice	Notice 1016
Notice 1027	Notice	Notice 1027
Notice 1392	Publication	Notice 1392
Publication 15	Publication 15	Publication 15
Publication 16	Publication 16	Publication 16
Publication 17	Publication 17	Publication 17
Publication 216	Publication	Publication 216
Publication 1141	Publication	Publication 1141
Publication 1223	Publication	Publication 1223
Publication 1516	Publication 1516	Publication 1516
Publication 1518-A	Publication	Publication 1518-A
Publication 1546	Publication	Publication 1546
Total count: 50		

Table 19: IRS-50 labels and predictions of Falcon-40B and DFv2$_{\text{base}}$ S+U, which was trained with supervised annotations and unsupervised distillation in Table 5. Red-colored text indicates false predictions.