A Minor dataset modifications

Deduplication. Through the systematic analysis and validation of the chosen datasets, we noticed one of the commonly appearing defects is the presence of duplicated annotations. We decided to remove these duplicates from InfographicsVQA (14 annotations from train, two from the dev set), DocVQA (four from train and test sets each), TabFact (309 from train, 53 from dev, and 52 the test set), and WikiTableQuestions (one annotation from each train and test sets).

B Tasks processing and reformulation

Since part of the datasets were reformulated or modified to improve the benchmark quality or align the task with the Document Understanding paradigm, we describe the introduced changes in detail below.

WikiTableQuestions★. We prepare input documents by rendering table-related HTML distributed by authors in *wkhtmltopdf* and crop the resulting files with *pdfcrop*. As these code excerpts do not contain *head* tag with JavaScript and stylesheet references, we use the header from the present version of the Wikipedia website.

Approximately 10% of tables contained at least one *img* tag with a source that is no longer reachable. It results in a question mark icon displayed instead of the image and does not impact the evaluation procedure since the questions here do not require image comprehension.

Year	Venue	Winners	Runner-up	3rd place
2005	Pardubice	Poland (41 pts)	Sweden (35 pts)	Denmark (24 pts)
2006	Rybnik	Poland (41 pts)	Sweden (27 pts)	Denmark (26 pts)
2007	Abensberg	Poland (40 pts)	Great Britain (36 pts)	Czech Republic (30 pts)
2008	Holsted	Poland (40 pts)	Denmark (39 pts)	Sweden (38 pts)
2009	Gorzów Wlkp.	Poland (57 pts)	Denmark (45 pts)	Sweden (32 pts)
2010	Rye House	Denmark (51 pts)	Sweden (37 pts)	Poland (35 pts)
2011	Balakovo	Russia (61 pts)	Denmark (31 pts)	Ukraine (29+3 pts)
2012	Gniezno	Poland (61 pts)	Australia (44 pts)	Sweden (26 pts)
Year	Venue	Winners	Runner-up	3rd place

Figure 4: Document in WikiTableQuestions reformulated as Document Understanding.
(Question) After their first place win in 2009, how did Poland place the next year at the speedway junior world championship? (Answer) 3rd place

The original WTQ dataset consists of *training*, *pristine-seen-tables*, and *pristine-unseen-tables* subsets. We treat *pristine-unseen-tables* as a test set and create new training and development sets by rearranging data from *training* and *pristine-seen-tables*. The latter operation is dictated by the leakage of documents in the original formulation, i.e., we consider it undesirable for a document to appear in different splits, even if the question differs. The resulting dataset consists of approximately 2100 documents divided in the proportion of 65%, 15%, 20% into training, development, and test sets.

TabFact★. As the authors of TabFact distribute only CSV files, we resorted to HTML from the WikiTables dump their CSV were presumably generated from.[4] As Chen et al. [6] dropped some of the columns present in used WikiTable tables, we remove them too, to ensure compatibility with the original TabFact. Rendered files are used analogously to the case of WTQ.

Results differ from TabFact in several aspects, i.e., text in our variant is not normalized, it includes the original formatting, and the tables are more complex due to restoring the original cell merges. All mentioned differences are desired, as we intended to consider raw, unprocessed files without any heuristics or normalization applied.

Another difference we noticed is that tables in the original TabFact are sometimes one row shorter, i.e., they do not contain the last row present in the WikiTable dump. As it should not impact expected answers, we decided to maintain the fidelity to Wikipedia and use the complete table.

We use the original splits into training, development, and test sets.

[4] `http://websail-fe.cs.northwestern.edu/TabEL/tables.json.gz`

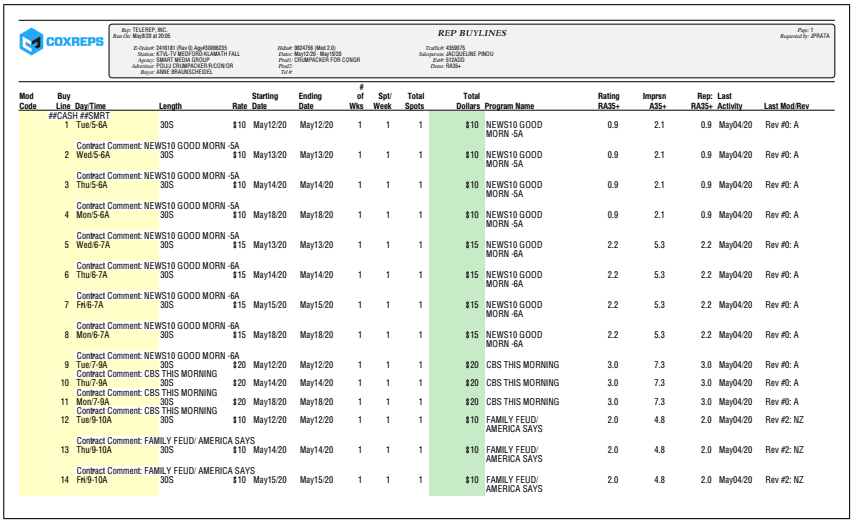
Superleague (Final League) Table (Places 1-6)

	Nation	v t e Games				Points			Table points
		Played	Won	Drawn	Lost	For	Against	Difference	
1	VVA-Podmoskovye Monino	10	9	0	1	374	119	+255	37
2	Krasny Yar Krasnoyarsk	10	6	0	4	198	255	-57	28
3	Slava Moscow	10	5	1	4	211	226	-15	26
4	Yenisey-STM Krasnoyarsk	10	5	0	5	257	158	+99	25
5	RC Novokuznetsk	10	4	1	5	168	194	-26	23
6	Imperia-Dynamo Penza	10	0	0	10	138	395	-257	10

Figure 5: Document in TabFact reformulated as Document Understanding.
(Claim) To calculate table point, a win be worth 3, a tie be worth 1 and a loss be worth 0

DeepForm★. The original DeepForm dataset consists of 2012, 2014, and 2020 subsets differing in terms of annotation quality and documents' diversity. We decided to use only the 2020 subset as for 2014, and 2020 annotations were prepared either automatically or by volunteers, leading to questionable quality. The selected subset was randomly divided into training, development and test set.

We noticed several inconsistencies during the initial analysis that lead us to the manual correction of autodetected: (1) invalid date format; (2) flight start dates earlier than flight end; (3) documents lacking one or more data points.


COXREPS

REP BUYLINES

Mod Code	Buy Line	Day/Time	Length		Starting Date	Ending Date	# of Wks	Spt/ Week	Total Spots	Total Dollars	Program Name	Rating RA35+	Imprsn A35+	Rep: RA35+	Last Activity	Last Mod/Rev
		##CASH ##SMRT														
	1	Tue/5-6A	30S	$10	May12/20	May12/20	1	1	1	$10	NEWS10 GOOD MORN -SA	0.9	2.1	0.9	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -SA														
	2	Wed/5-6A	30S	$10	May13/20	May13/20	1	1	1	$10	NEWS10 GOOD MORN -SA	0.9	2.1	0.9	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -SA														
	3	Thu/5-6A	30S	$10	May14/20	May14/20	1	1	1	$10	NEWS10 GOOD MORN -SA	0.9	2.1	0.9	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -SA														
	4	Mon/5-6A	30S	$10	May18/20	May18/20	1	1	1	$10	NEWS10 GOOD MORN -SA	0.9	2.1	0.9	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -SA														
	5	Wed/6-7A	30S	$15	May13/20	May13/20	1	1	1	$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -6A														
	6	Thu/6-7A	30S	$15	May14/20	May14/20	1	1	1	$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -6A														
	7	Fri/6-7A	30S	$15	May15/20	May15/20	1	1	1	$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -6A														
	8	Mon/6-7A	30S	$15	May18/20	May18/20	1	1	1	$15	NEWS10 GOOD MORN -6A	2.2	5.3	2.2	May04/20	Rev #0: A
		Contract Comment: NEWS10 GOOD MORN -6A														
	9	Tue/7-9A	30S	$20	May12/20	May12/20	1	1	1	$20	CBS THIS MORNING	3.0	7.3	3.0	May04/20	Rev #0: A
		Contract Comment: CBS THIS MORNING														
	10	Thu/7-9A	30S	$20	May14/20	May14/20	1	1	1	$20	CBS THIS MORNING	3.0	7.3	3.0	May04/20	Rev #0: A
		Contract Comment: CBS THIS MORNING														
	11	Mon/7-9A	30S	$20	May18/20	May18/20	1	1	1	$20	CBS THIS MORNING	3.0	7.3	3.0	May04/20	Rev #0: A
		Contract Comment: CBS THIS MORNING														
	12	Tue/9-10A	30S	$10	May12/20	May12/20	1	1	1	$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0	May04/20	Rev #2: NZ
		Contract Comment: FAMILY FEUD/ AMERICA SAYS														
	13	Thu/9-10A	30S	$10	May14/20	May14/20	1	1	1	$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0	May04/20	Rev #2: NZ
		Contract Comment: FAMILY FEUD/ AMERICA SAYS														
	14	Fri/9-10A	30S	$10	May15/20	May15/20	1	1	1	$10	FAMILY FEUD/ AMERICA SAYS	2.0	4.8	2.0	May04/20	Rev #2: NZ

Figure 6: Single page from document in DeepForm.

In addition to the improved 2020 subset, we manually annotated one hundred 2012 documents, as they can pose different challenges (contain different document templates, handwriting, have lower image quality). They were used to extend development and test set. The final dataset consists of 700 training, 100 development, and 300 test set documents.

PWC★. The authors of AxCell relied on PWC Leaderboards and LinkedResults datasets [24]. The original formulation assumes extraction of *(task, dataset, metric, model, score)* tuples from a provided table. In contrast, we reformulate the task as Document Understanding and provide a complete paper as input instead. These are obtained using arXiv identifiers available in the PWC metadata. Consequently, the resulting task is an end-to-end Key Information Extraction from real-world scientific documents.

Whereas LinkedResults was annotated consistently, the PWC is of questionable quality as it was obtained from leaderboards filled by Papers with Code visitors without a clear guideline or annotation rules. The difference between the two is substantial, i.e., the agreement in terms of F1 score between publications present in both PWC and LinkedResults is lower than 0.35. We attribute this mainly to

flaws in the PWC dataset, such as missing records, inconsistent normalization and the difficulty of the task itself.

Consequently, we decided to perform its manual re-annotation assuming that: (1) The best result for a proposed model variant on the single dataset has to be annotated, e.g., if two models with different parameter sizes were present in the table, we report only the best one. (2) Single number is preferred (we take the average over multiple split or parts of the dataset if possible). (3) When results from the test set are available, we prefer them and don't report results from the validation set. (4) We add multiple value variants when possible. (5) We include information on used validation/dev/test split in the dataset description wherever applicable. (6) We don't report results on the train set. (7) We don't annotate results not appearing in the table. (8) We filter out publications that are hard to annotate even for a human.

Interestingly, human scores on PWC are relatively low in terms of F1 value. This can be attributed to unrestricted nature of particular properties, e.g., *accuracy* and *average accuracy* are equally valid metric values. Similarly, *Action Recognition*, *Action Classification*, and *Action Recognition* are eqally valid task names. At the same time, it is impossible to provide all answer variants during the preparation of the gold standard. We decided to keep the dataset in the benchmark as it is extremely demanding, and there is still a large gap between humans' and models' performance (See Table 3).

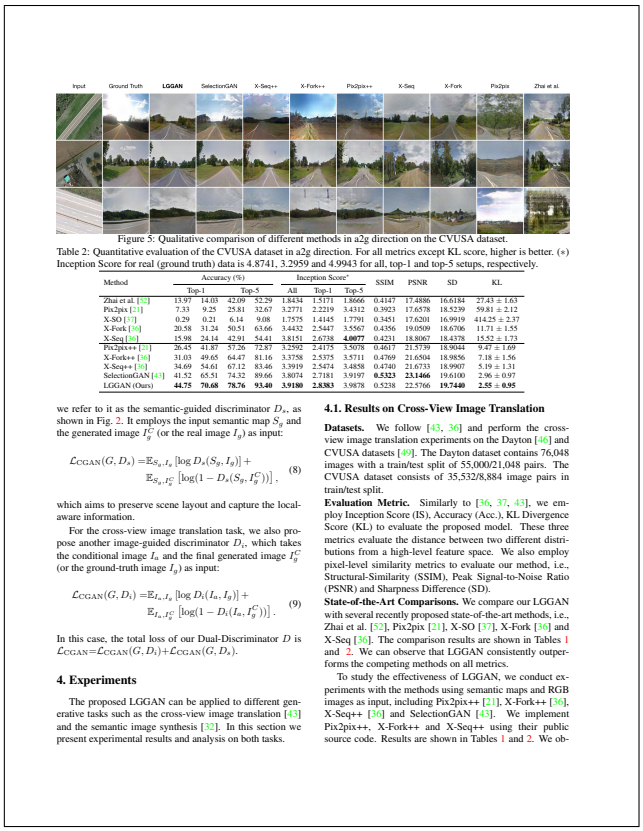


Figure 5: Qualitative comparison of different methods in a2g direction on the CVUSA dataset.

Table 2: Quantitative evaluation of the CVUSA dataset in a2g direction. For all metrics except KL score, higher is better. (*) Inception Score for real (ground truth) data is 4.8741, 3.2959 and 4.9943 for all, top-1 and top-5 setups, respectively.

Method	Accuracy (%)				Inception Score*			SSIM	PSNR	SD	KL
	Top-1		Top-5		All	Top-1	Top-5				
Zhai et al. [52]	13.97	14.03	42.09	52.29	1.8434	1.5171	1.8666	0.4147	17.4886	16.6184	27.43 ± 1.63
Pix2pix [21]	7.33	9.25	25.81	32.67	3.2771	2.2219	3.4312	0.3923	17.6578	18.5239	59.81 ± 2.12
X-SO [37]	0.29	0.21	6.14	9.08	1.7575	1.4145	1.7791	0.3451	17.6201	16.9919	414.25 ± 2.37
X-Fork [36]	20.58	31.24	50.51	63.66	3.4432	2.5447	3.5567	0.4356	19.0509	18.6706	11.71 ± 1.55
X-Seq [36]	15.98	24.14	42.91	54.41	3.8151	2.6738	**4.0077**	0.4231	18.8067	18.4378	15.52 ± 1.73
Pix2pix++ [21]	26.45	41.87	57.26	72.87	3.2592	2.4175	3.5078	0.4617	21.5739	18.9044	9.47 ± 1.69
X-Fork++ [36]	31.03	49.65	64.47	81.16	3.3758	2.5375	3.5711	0.4769	21.6504	18.9856	7.18 ± 1.56
X-Seq++ [36]	34.69	54.61	67.12	83.46	3.3919	2.5474	3.4858	0.4740	21.6733	18.9907	5.19 ± 1.31
SelectionGAN [43]	41.52	65.51	74.32	89.66	3.8074	2.7181	3.9197	**0.5323**	**23.1466**	19.6100	2.96 ± 0.97
LGGAN (Ours)	**44.75**	**70.68**	**78.76**	**93.40**	**3.9180**	**2.8383**	3.9878	0.5238	22.5766	**19.7440**	**2.55 ± 0.95**

we refer to it as the semantic-guided discriminator D_s, as shown in Fig. 2. It employs the input semantic map S_g and the generated image I_g^C (or the real image I_g) as input:

$$\mathcal{L}_{\text{CGAN}}(G, D_s) = \mathbb{E}_{S_g, I_a}\left[\log D_s(S_g, I_g)\right] + \mathbb{E}_{S_g, I_g^C}\left[\log(1 - D_s(S_g, I_g^C))\right], \quad (8)$$

which aims to preserve scene layout and capture the local-aware information.

For the cross-view image translation task, we also propose another image-guided discriminator D_i, which takes the conditional image I_a and the final generated image I_g^C (or the ground-truth image I_g) as input:

$$\mathcal{L}_{\text{CGAN}}(G, D_i) = \mathbb{E}_{I_a, I_g}\left[\log D_i(I_a, I_g)\right] + \mathbb{E}_{I_a, I_g^C}\left[\log(1 - D_i(I_a, I_g^C))\right]. \quad (9)$$

In this case, the total loss of our Dual-Discriminator D is $\mathcal{L}_{\text{CGAN}} = \mathcal{L}_{\text{CGAN}}(G, D_i) + \mathcal{L}_{\text{CGAN}}(G, D_s)$.

4. Experiments

The proposed LGGAN can be applied to different generative tasks such as the cross-view image translation [43] and the semantic image synthesis [32]. In this section we present experimental results and analysis on both tasks.

4.1. Results on Cross-View Image Translation

Datasets. We follow [43, 36] and perform the cross-view image translation experiments on the Dayton [46] and CVUSA datasets [49]. The Dayton dataset contains 76,048 images with a train/test split of 55,000/21,048 pairs. The CVUSA dataset consists of 35,532/8,884 image pairs in train/test split.

Evaluation Metric. Similarly to [36, 37, 43], we employ Inception Score (IS), Accuracy (Acc.), KL Divergence Score (KL) to evaluate the proposed model. These three metrics evaluate the distance between two different distributions from a high-level feature space. We also employ pixel-level similarity metrics to evaluate our method, i.e., Structural-Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness Difference (SD).

State-of-the-Art Comparisons. We compare our LGGAN with several recently proposed state-of-the-art methods, i.e., Zhai et al. [52], Pix2pix [21], X-SO [37], X-Fork [36] and X-Seq [36]. The comparison results are shown in Tables 1 and 2. We can observe that LGGAN consistently outperforms the competing methods on all metrics.

To study the effectiveness of LGGAN, we conduct experiments with the methods using semantic maps and RGB images as input, including Pix2pix++ [21], X-Fork++ [36], X-Seq++ [36] and SelectionGAN [43]. We implement Pix2pix++, X-Fork++ and X-Seq++ using their public source code. Results are shown in Tables 1 and 2. We ob-

Figure 7: Single page from document in PWC.

C Dataset statistics

Chosen datasets represent the plethora of domains, lengths, and document types. This appendix covers the critical aspects of particular tasks at the population level.

Though part of the datasets is limited to one-pagers, the remaining documents range from a few to few hundred pages (Figure 8). At the same time, there is a great variety in how much text is present on a single page – we have both densely packed scientific documents and concise document excerpts

or infographics. This diversity allows us to measure the ability to comprehend documents depending on their length.

D Details of human performance estimation

Estimation of human performance for PWC, WikiTableQuestions, DeepForm was performed in-house by professional annotators who are full-time employees of Applica.ai. Before approaching the process, each of them has to participate in the task-specific training described below.

Number of annotated samples depended on task difficulty and the variance of the resulting scores. We relied on 50 fully annotated papers for the PWC dataset (approx. 150 tuples with five values each), 109 DeepForm documents (532 values), and 300 questions asked to different WikiTableQuestion tables.

Each dataset was approached with two annotators in the LabelStudio tool. Human performance is the average of their scores when validated against the gold standard.

Training. Each person participating in the annotation process completed the training consisting of four stages: (1) Annotation of five random documents from the task-specific development set. (2) Comparative analysis of differences between their annotations and the gold standard. (3) Annotation of ten random documents from the task-specific development set and subsequent comparative analysis. (4) Discussion between annotators aimed at agreeing on the shared, coherent annotation rules.

E Annotation of diagnostic subsets

In order to analyze the prepared benchmark and the results of individual models, diagnostic sets were prepared. These diagnostic sets are subsets of examples selected from the testset for all datasets.

When building a taxonomy for diagnostic sets, we adopted two basic assumptions: (1) It must be consistent across all selected tasks so that at least two tasks can be noted with a given category (2) It should include as many aspects as possible that are relevant from the perspective of document understanding problem.

Initially, we adopted the taxonomies proposed in DocVQA, Infographics, and TabFact as potential categories [29, 28, 6]. In the next step, we adjusted our taxonomy to all datasets following the previously adopted assumptions, distinguishing seven main categories with 25 subcategories (for a more detailed description of the category (see the section E.1). Then, for each dataset, we prepared an annotation task in the LabelStudio tool [5] (see example 9) along with an annotation instruction. Finally, to determine Human performance, the annotation was carried out by a team of specialists from Applica.ai, where the selected example was noted only by one person.

E.1 Taxonomy description

The taxonomy is based on multiple aspects of documents, inputs, and answers and was designed to be sufficiently generic for future adaptation to other tasks. Here, in each category, we describe the predicates that annotators followed when classified an example into specific subcategories.

Answer source. This category is based on the relation between answer and text in the document.

- Extractive – after lowercasing and white-characters removing, the answer can be exact-matched in the document.
- Inferred – other non-extractive cases.

Output format This category is based on the shape of an output.

- Single value – the answer consists of only one item.
- List – multiple outputs are to be provided.

[5] https://labelstud.io/

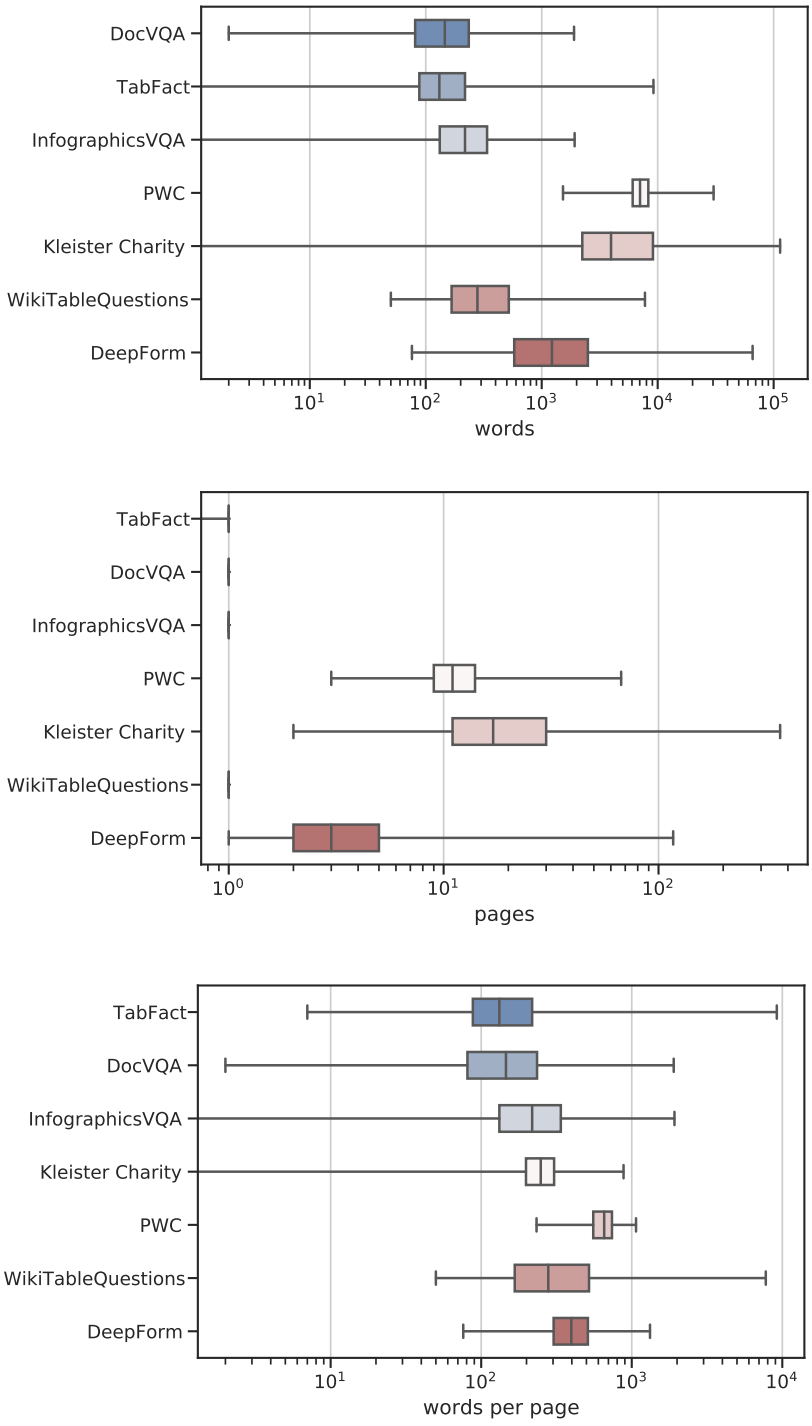


Figure 8: Number of words, pages, and words per page in particular datasets (log scale). Part of the datasets consist only of one-pagers.

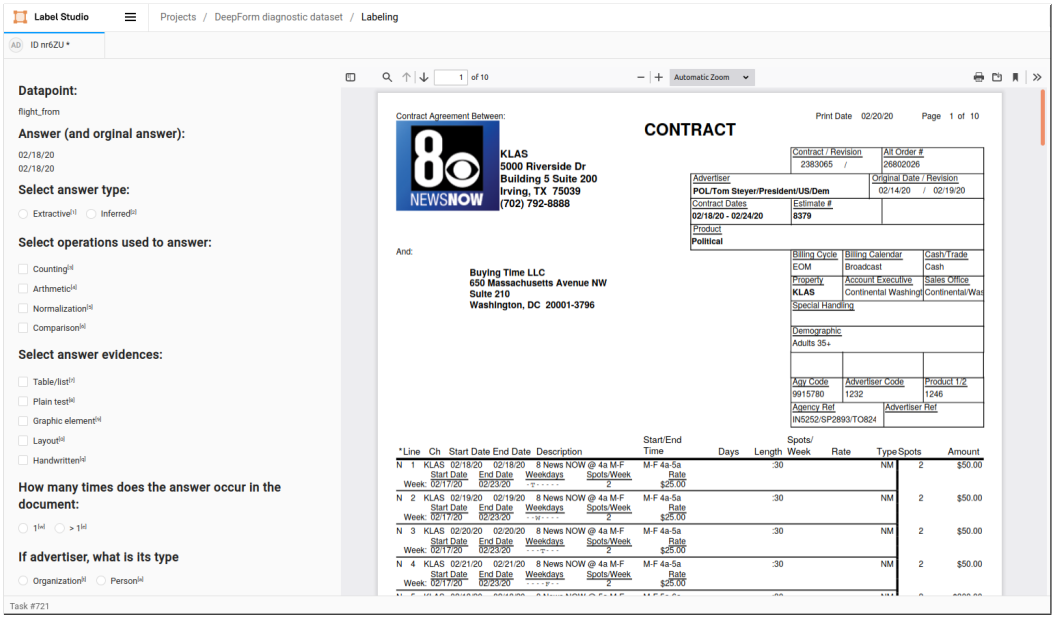


Figure 9: An example of an interface for annotating diagnostic subsets based on document from DeepForm dataset.

Output type. This category is based on the semantic of an output.

- Organization – the answer is a name of an organization or institution.
- Location – the answer is a geographic location globally (e.g., a country, continent, city) or locally (building or street, among others).
- Person – the answer is a personal identifier(name, surname, pseudonym) or its composition. It can have a title prefix or suffix (e.g., Mrs., Mr., Ph.D.) or have a shortened or informal version.
- Number – numerical values given with the unit or percent. Values written in the free text do not comply with this class's definition.
- Date/Time/Duration – the answer represents the date, time, or the difference between two dates or times.
- Yes/No – the answer is a textual output of binary classification, such as Yes/No pairs, and Positive/Negative, 0/1 among others.

Evidence. This category is based on the source of information that allows the correct answer to be generated. When there are multiple justifications based on different pieces of evidence (for example, the address is in a table and block text), it is required to select all the pieces of evidence.

- Table or List – a *table* is a fragment of the document organized into columns and rows. The distinguishing feature of the table is consistency within rows and columns (usually the same data type). Moreover, it may have a header. In that sense, the form is not a table (or at least it does not have to be). A *list* is a table degenerated into one column or row containing a header.
- Plain text – the answer is based on plain text if there is an immediate need to understand a longer fragment of the text while answering.
- Graphic element – the answer is based on graphic evidence when understanding graphically rich, non-text fragments of documents (e.g., graphics, photos, logos (non-text)) are necessary for generating a correct answer.
- Layout – it is evidence when comprehending the placement of text on the page (e.g., titles, headers, footers, forms) is needed to generate the correct answer. This type does not include tables.
- Handwritten – when the text written by hand is crucial for an answer.

Operation. This category is based on the type of operations that are to be performed on the document before reaching to the correct answer.

- Counting – when there is a need to count the occurrences or determine the position on the list.
- Arithmetic – when there is an arithmetic operation applied before answering, or a sequence of arithmetic operations (e.g., averaging).
- Comparison – a comparison in the sense of lesser/greater. Other procedures that a comparison operation can express (e.g., approximation) may be chosen. Here, the operation "is equal" is not a comparison since it is sufficient to match sequences without a semantic understanding.
- Normalization – when we are to return something in the document but in a different form. It may only apply to the output; we do not acknowledge this operation when it is required to normalize a question fragment to match it in the document.

Answer number. This category is based on the number of occurrences of an answer in the document.

- 1 – when there is one path of logical reasoning to find the correct answer in the document. We treat it as one justification for two different reasoning paths based on the same data from the document.
- > 1 – the other cases.

F Training details

The experiments were carried out in an environment with NVIDIA A100-40Gb cards, PyTorch version 1.8.1, and huggingface-transformers in version 4.2.2.

The parameters were selected through empirical experiments with T5-Base model on DocVQA and InfographicsVQA collections. The T5-Large model was used as the basis for finetuning.

The training lasted up to 30 epochs at batch 64 in training, the default optimizer AdamW (lr = 0.0002), and warmup set to 100 updates. Validation was performed five times per epoch, and when no improvement was seen for 20 validation steps (4 epochs), the training was stopped. The length of the input documents has been truncated to 1024 tokens and the responses to 256 tokens. Dropout was set to 0.15, gradient clipping to 1.0, and weight decay to 1e-05.

The complete source code is attached as the supplementary material.

G Considered datasets

The review protocol consisted of a manual search in specific databases, repositories and distribution services. The scientific resources included in the search were:

- `https://paperswithcode.com/datasets/`
- `https://datasetsearch.research.google.com/`
- `https://data.mendeley.com/`
- `https://arxiv.org/search/`
- `https://github.com/`
- `https://allenai.org/data/`
- `https://www.semanticscholar.org/`
- `https://scholar.google.com/`

Results were reviewed by one of authors of the present paper and the resources related to classification, KIE, QA, MRC, and NLI over complex documents, figures, and tables were identified as potentially relevant (in accordance with inclusion criteria described in Section 3.1).

The initial search assumed use of the following keywords: *Question Answering, Visual Question Answering, Document Question Answering, Document Classification, Document Dataset, Information*

Extraction. Additionally, we used *Machine Reading Comprehension, Question Answering, VQA* in combination with *Document*, and *Visual, Document, Table, Figure, Plot, Chart, Hybrid* in combination with *Question Answering* or *Information Extraction*.

Table below lists potentially relevant tasks and results of their assessment according to the criteria of quality, difficulty, and licensing.

Name	Type	Comment	
Kleister NDA	KIE	Dominated by extraction from free text (layout is not important).	[41]
SROIE	KIE	No room for improvement.	[17]
CORD	KIE	No room for improvement.	[35]
Wildreceipt	KIE	Very similar to SROIE and CORD. No room for improvement as the main problem is poor OCR quality.	[43]
FUNSD	KIE	Small dataset size (measured in number of data points) and known disadvantages [50].	[22]
TextbookQA	Document VQA	Source files are not available (only images and free text). In the meantime, the used online textbook has changed.	[25]
PlotQA	Figure QA	Synthetic	[31]
WebSRC	KIE	Templated input data (both questions and websites).	[5]
MultiModalQA	QA over Tables, Images and Text	Automatically generated questions.	[45]
WikiOPS	Table QA	No room for improvement.	[8]
FeTaQA	Table QA	Wikipedia Tables. Answers as a free-form text	[32]
TabMCQ	Table QA	Low number of tables used.	[20]
LEAF-QA	Figure QA	Templated questions.	[4]
VisualMRC	Document VQA	No room for improvement (human performance reached)'.	[47]
DocFigure	Classification	No room for improvement.	[23]
Tabacco3482	Classification	No room for improvement.	
RVL-CDIP	Classification	No room for improvement.	[15]
EURLEX57K	Classification	Dominated by extraction from free text (layout is not important).	[3]
MELINDA	Classification	The dataset was annotated in semi-automatic manner.	[53]
DWIE	IE	Dominated by extraction from free text (layout is not important).	[61]
HybridQA	Table QA	Multihop Question Anwering.	[7]

H Benchmark datasheet

Following Gebru et al. [13] we fill the datasheet for the proposed benchmark. As it was originally designed for datasets, part of the questions might not apply and were skipped.

H.1 Motivation for datasheet creation

Why was the benchmark created? Despite its importance for digital transformation, the problem of measuring how well available models obtain information from a wide range of document types and how suitable they are for freeing workers from paperwork through process automation is not yet addressed. We intend to bridge this major gap by introducing the first Document Understanding benchmark.

Has the benchmark been used already? If so, where are the results so others can compare (e.g., links to published papers)? No, the paper describes the first version of the benchmark.

Who funded the creation dataset? Applica.ai

H.2 Benchmark composition

What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges) Single instance is a PDF document such as report, scientific publication, form, infographic or table excerpted from websites. For each instance in train and dev split we provide associated question-answer or property-value pairs.

How many instances are there in total (of each type, if appropriate)? DocVQA totals $12.8k$ examples, InfographicsVQA totals $5.5k$, Kleister Charity totals $2.7k$, PWC totals $0.4k$m DeepForm totals $1.1k$ WikiTableQuestins totals $2.1k$ and TabFact totals $16.6k$

What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution? OCR layer from scanned PDF, textual question (or property) and textual answer (value), meta-data and diagnostic information.

Is there a label or target associated with each instance? If so, please provide a description. Yes, there is an answer or multiple allowed answers specified for each instance.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. For each instance we provide output form OCR tools (Tesseract, Microsoft Computer Vision API, djvu). However, few documents are problematic for OCR engines and for them we were not able to generate text and layout layer.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Yes, they contain metadata that informs about the id of the document that was used for the instance. Different instances may share the same underlying document.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. For five out of seven tasks from our benchmark we used original datasets splits. The two datasets in which we changed splits are:

DeepForm. The original DeepForm dataset consists of 2012, 2014, and 2020 subsets differing in terms of annotation quality and documents' diversity. We decided to use only the 2020 subset as for 2014, and 2020 annotations were prepared either automatically or by volunteers, leading to questionable quality. The selected subset was randomly divided into training, development and test set. In addition to the improved 2020 subset, we manually annotated one hundred 2012 documents, as they can pose different challenges (contain different document templates, handwriting, have lower image quality). They were used to extend development and test set. The final dataset consists of 700 training, 100 development, and 300 test set documents.

WikiTableQuestions. The original WTQ dataset consists of training, pristine-seen-tables, and pristine-unseen-tables subsets. We treat pristine-unseen-tables as a test set and create new training and development sets by rearranging data from training and pristine-seen-tables. The latter operation is dictated by the leakage of documents in the original formulation, i.e., we consider it undesirable for a document to appear in different splits, even if the question differs. The resulting dataset consists of approximately 2100 documents divided in the proportion of 65%, 15%, 20% into training, development, and test sets.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. In our benchmark we have two sources of errors:

Annotations. For each task we provided human performance estimation which shows how often the annotators were in agreement with each other (what is the level of annotation noise).

OCR output. As an input for all tasks we used PDF files. Therefore, we used OCR tools (which is no perfect) to retrieve text and layout layer (token bounding boxes).

Is the benchmark self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. Despite the fact that the benchmark aggregates dataset published in various sources it is self-contained. To eliminate some of the barriers in future experiments, we proposed a format to unify varied Document Understanding tasks and convert all of the datasets included in the benchmark. Additionally, we provide versioned OCR layers for scanned documents to make models evaluated in the future directly comparable.

All of these resources are provided on the benchmark website, without a need to download them from external sources.

H.3 Collection Process

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? For WikiTableQuestions, we prepare input documents by rendering table-related HTML distributed by authors in *wkhtmltopdf* and crop the resulting files with *pdfcrop*. As these code excerpts do not contain *head* tag with JavaScript and stylesheet references, we use the header from the present version of the Wikipedia website.

As the authors of TabFact distribute only CSV files, we resorted to HTML from the WikiTables dump their CSV were presumably generated from.[6] As Chen et al. [6] dropped some of the columns present in used WikiTable tables, we remove them too, to ensure compatibility with the original TabFact. Rendered files are used analogously to the case of WTQ.

The remaining datasets had their data kept in the original form. Used procedures were designed and validated in an iterative manner by: (1) validating all generated documents against original source (CSV) and (2) checking a random sample of 200 documents manually looking for anomalies. If any errors were detected, the processing software was fixed and the validation procedure started again.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Estimation of human performance for PWC, WikiTableQuestions, DeepForm and annotation of diagnostic subsets was performed in-house (at Applica.ai) by professional annotators in their work time.

H.4 Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. We provide OCR layers for PDF documents to make models evaluated in the future directly comparable.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. The original PDF files are hosted on the https://duebenchmark.com/data and can be downloaded from there.

[6] `http://websail-fe.cs.northwestern.edu/TabEL/tables.json.gz`

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. To preprocess all documents we have used two OCR tools:

1. Tesseract in version 4.1.1[7]
2. Microsoft Azure Computer Vision API (Azure CV) in version 3.0.0[8]

To estimate human performance and for annotation diagnostic datasets we used open source Label-Studio[9] software (screenshots is provided in the paper appendix for reference).

H.5 Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?) All datasets from our benchmark are available on the https://duebenchmark.com/data and can be downloaded from there. Moreover, for each dataset we provide JSON-LD file[10] with detailed description.

When will the dataset be released/first distributed? What license (if any) is it distributed under? We released all datasets already. We used original license for all datasets that we selected to our benchmark.

Are there any fees or access/export restrictions? No.

H.6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? Applica.ai

Will the dataset be updated? If so, how often and by whom? No.

If the dataset becomes obsolete how will this be communicated? We will notify users on benchmark site: https://duebenchmark.com/

Is there a repository to link to any/all papers/systems that use this dataset? Everyone who want to use our benchmark should submit their results via site https://duebenchmark.com/. The submission should also contain reference to the paper.

Any other comments? We are not planning to update prepared datasets in our benchmark but we consider to prepare second version of our benchmark in the future (with updated list of datasets).

H.7 Legal and Ethical Considerations

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. In our benchmark we are using datasets which were collected by other researchers and therefore we do not conduct any ethical review processes. Moreover, all datasets are already available.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description. No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why No.

[7] https://github.com/tesseract-ocr/tesseract/releases/tag/4.1.1

[8] https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr

[9] https://labelstud.io/

[10] https://developers.google.com/search/docs/data-types/dataset

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
DocVQA. No.
InfographicsVQA. No.
Kleister Charity. No.
PWC. Yes.
DeepForm. Yes.
WikiTableQuestions. Yes.
TabFact. Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
DocVQA. No.
InfographicsVQA. No.
Kleister Charity. No.
PWC. Yes — we can check what are the authors of the publications.
DeepForm. Yes — in this dataset we are processing receipts from political campaign ads bought around US elections. Sometimes on these forms we could find politician person names.
WikiTableQuestions. Yes - data comes from Wikipedia so we can check person indirectly by going to Wikipedia page from which table was extracted.
TabFact. Yes — data comes from Wikipedia so we can check person indirectly by going to Wikipedia page from which table was extracted.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. *DocVQA.* No.
InfographicsVQA. No.
Kleister Charity. No.
PWC. No.
DeepForm. Yes — we have information on how much money a given person donated to support the presidency campaign (but this information is publicly available).
WikiTableQuestions. No.
TabFact. No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? We used data collected by other researchers.