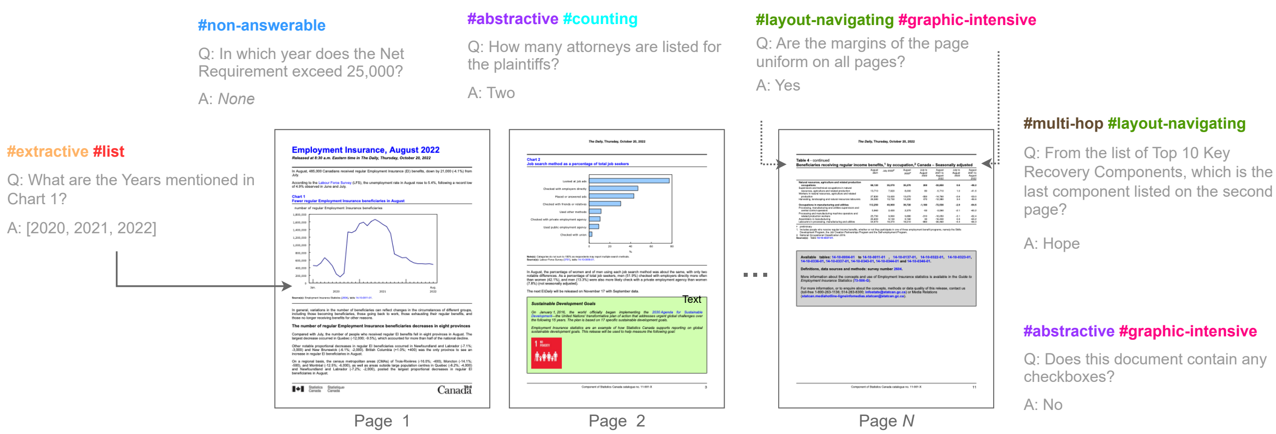
Document Understanding Dataset and Evaluation (DUDE😎)

Jordy Van Landeghem[1,2] Rubèn Tito[5] Łukasz Borchmann[3] Michał Pietruszka[3,6] Paweł Józiak[3,4]
Rafał Powalski[8] Dawid Jurkiewicz[3,7] Mickaël Coustaty[9] Bertrand Ackaert[2] Ernest Valveny[5]
Matthew Blaschko[1] Sien Moens[1] Tomasz Stanisławek[3]

[1]KU Leuven [2]Contract.fit [3]Snowflake [4]Warsaw University of Technology [5]Computer Vision Center, Universitat Autònoma de Barcelona
[6]Jagiellonian University [7]Adam Mickiewicz University [8]Instabase [9]University of La Rochelle



Abstract

*We call on the Document AI (DocAI) community to re-evaluate current methodologies and embrace the challenge of creating more practically-oriented benchmarks. Document Understanding Dataset and Evaluation (**DUDE**) seeks to remediate the halted research progress in understanding visually-rich documents (VRDs). We present a new dataset[1] with novelties related to types of questions, answers, and document layouts based on **multi-industry**, **multi-domain**, and **multi-page** VRDs of various origins, and dates. Moreover, we are pushing the boundaries of current methods by creating multi-task and multi-domain evaluation setups that more accurately simulate real-world situations where powerful generalization and adaptation under low-resource settings are desired. **DUDE** aims to set a new standard as a more practical, long-standing benchmark for the community, and we hope that it will lead to future extensions and contributions that address real-world challenges. Finally, our work illustrates the importance of finding more efficient ways to model language, images, and layout in DocAI.*

[1]huggingface.co/datasets/jordyvl/DUDE_loader

1. Introduction

Early stages of research and growth in any field are characterized by enacting proof-of-concept and demonstrating the feasibility of the proposed solution. In the Deep Learning era, this is often echoed by building narrow and simplified datasets that do not reflect real-world complexity, leading to models that may not be suitable for practical use.

The field of Document Understanding (DU) is not an exception to the recent proliferation of deep architectures, which in this case are predominantly used for classification and information extraction from documents. However, the wide and complex nature of documents presents many challenges that remain unsolved or not yet addressed. One such challenge is domain generalization, where a model trained on medical documents may not be directly applicable to financial or tabular content. Another challenge concerns task-agnostic architectures, where a model must be able to adapt to various DU subtasks such as document classification, key information extraction (KIE), and question answering (QA). Lastly, the high variability of document contents and layouts often leads to highly imbalanced samples within document types, resulting in a long-tailed distribution with few or almost no samples to train a model.

Despite the importance of these challenges, there is currently no DU benchmark dataset that simultaneously addresses all of these issues. This paper proposes a novel dataset formulated as an instance of Document Visual Question Answering (DocVQA) to evaluate how well current DU solutions deal with multi-page documents, if they can navigate and reason over visual layouts, and if they can generalize their skills to different document types and domains.

The data collection and evaluation design of **DUDE** naturally motivates targeting models that can answer natural yet highly diverse questions (e.g., regarding document elements, their properties, and compositions) for any VRD (e.g., drawn from potentially unseen distributions of layouts, domains, and types). The presented problem setting relates to Multi-Domain Long-Tailed Recognition (MDLT) [97], which concerns learning from multi-domain imbalanced data whilst addressing label imbalance, divergent label distributions across domains, and possible train-test domain shift. Put plainly, since we cannot provide ground truth QA pairs for, e.g., stamps, on every document type (domain), we expect a solution to transfer the subtask 'stamp detection' learned on document types where stamps naturally occur (and thus training QA pairs were created organically) to other domains. The DocVQA and MDLT formulations of **DUDE** allow us to create a longstanding, challenging benchmark that in the future can be easily extended with more subtasks formulated as QA pairs, and domains relating to document types (see Limitations).

The contribution of this work is twofold. First, we have created **DUDE**, a novel large-scale, multi-paged, multi-domain, multi-industry DocVQA benchmark for evaluating DU progress. Second, we show that the zero-shot and fine-tuned performance of current state-of-the-art models applied to DU lags far behind human baselines, explained in part by the need for more holistic and efficient modeling of language, vision, and richly structured layouts.

2. Related Work

Document Understanding encompasses datasets related to various subtasks like document layout analysis [110, 49], classification [30], key information extraction [85, 35], table extraction [83, 109, 108], and visual question answering [57, 59, 91]. These benchmarks lead to end-to-end DU architectures that have transformed common DocAI practices [72, 5, 33, 23, 25, 50, 71]. These task-specific benchmarks, however, are often tailored to a single domain, limiting the ability to create and assess how well DU models generalize to other document types and domains. To fill this gap, we adopt a visual question answering (VQA) approach, which has been crucial in the growth of the DU field.

The VQA paradigm provides a natural language interface for various tasks from both computer vision and natural language processing. In the latter, the question-answering approach has been successfully used in several domains, including medicine [67, 39, 64, 36, 48, 76, 61], open-domain knowledge [98, 54, 58, 53], emotions [26, 9], code [2, 51], logical reasoning [52, 101, 107, 96], claim verification [88, 32, 104], and math [105, 31, 16, 60, 4]. As a result of its ability to function as a natural language interface for various forms of data, this paradigm has been applied to other domains. For example, the question-answering approach is combined with modalities such as videos [44, 13, 14, 28, 17], images [99, 3, 29, 68, 7, 8], speech [100, 43], knowledge graphs [93, 84, 80, 22, 37], and maps [70, 15].

Overall, the convergence of computer vision and NLP through the emergence of VQA tasks has also opened up new avenues for research in the DU field, with many DU datasets now including rich visual content alongside questions. Yet, prior study on document VQA has mainly focused on single-page documents [57, 89, 56] with rare exceptions such as MP-DocVQA [90]. However, [57, 89] pose only extractive questions where the answer follows the context on which the question is defined as in other question answering benchmarks [78, 92, 42]. Moreover, these datasets do not contain *non-answerable* questions as in established (natural language) QA datasets like [77, 42]. To the best of our knowledge, there are no VQA datasets containing questions requiring lists as an answer. There are however few text-only QA datasets that contain such answer types [69, 46, 18]. Other datasets mainly related to our work are rather domain-specific like [112, 87, 56, 86, 73]. We give a detailed comparison of most related document VQA datasets in Table 1 highlighting the major contributions.

3. DUDE Dataset

While **DUDE** shares some similarities with existing VQA datasets, a closer comparison (see Table 1) highlights its unique features. We are confident that the model's proficiency in the areas introduced in this work will showcase its capability to handle the intricacy and diversity of document understanding tasks in real-world scenarios.

Documents. The dataset covers a wide range of document types, sources and dates, as shown in Table 1 and Figure 1 where its diverse nature is confirmed by the spread of document content representations.[2] Moreover, it covers a broad range of domains, including medical, legal, technical, and financial, among others, to evaluate models' ability to handle diverse topics and the specific knowledge each requires. Furthermore, the dataset contains documents with varying layouts: diverse text arrangements, font sizes, and styles, to

[2]This holds not only when textual content is considered but also for document images (Figure 9 in the Appendix).

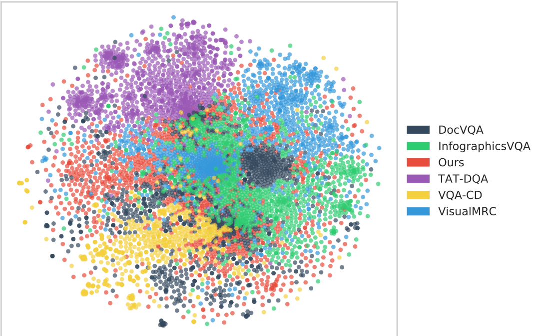


Figure 1: Visualization of inter-document similarities between samples from different datasets (t-SNE over TF-IDF representations of 1k passages from each source).

ensure that models can handle visually diverse documents.

In contrast to our proposal, current VQA datasets often focus on homogeneous documents, such as invoices in VQA-CD [55] or financial reports in TAT-DQA [112]. Even when not restricted to a single domain or layout, these datasets share essential characteristics. For example, InfographicsVQA [56] demonstrates significant diversity in topics and designs, but still embodies a preference for visual aids over complex tables or long text passages. Moreover, VQA datasets are commonly restricted to either born-digital or scanned documents, which limits their ability to measure the robustness to mixed-origin files that one usually finds in real-world applications. In particular, this restriction makes it uncertain whether state-of-the-art performers on website fragments from VisualMRC [87] can be efficient on multi-column layouts and documents with OCR errors or incorrectly-detected reading orders. Finally, a typical dataset for document visual question answering contains documents from a limited period, i.e., a few years (Table 1).

Considering the properties mentioned above, the most diverse dataset to date is Single Page DocVQA (SP-DocVQA) [57], which contains mixed-origin documents of different types created over several decades. However, it is built exclusively on single-page document excerpts and is limited to several domains represented in the Industry Documents Library. As a result, it complements rather than serves as a touchstone for general-purpose DU systems. MP-DocVQA [90] extends this including previous and posterior pages of the documents. However, the questions are kept the same which makes the extra pages mere distractors.

Questions. We use VQA as a natural language interface to VRDs, challenging the DU model with diverse questions, advanced operations, and multi-step reasoning to achieve real-world success.

Firstly, we assert that various layouts and visual elements must be comprehended semantically. As such, we introduce complex questions targeting these document elements, requiring comprehension beyond the document content, such as '*how many text columns are there?*', '*does the document contain words with diacritics?*' or '*which page contains the largest table in the document?*'. These Layout-navigating questions bridge the gap between Document Layout Analysis and Question Answering paradigms.

Our unique and detailed compositional questions demand a model that comprehends semantics and generalizes to new questions in a zero-shot setting. For example, >90% of our questions are unique, while we target questions whose answer scope is much more diverse than in previous works.[3] Since neural networks are known to perform poorly at mathematical reasoning and symbolical processing, we provide training and evaluation questions demanding arithmetic and comparison operations on numbers and dates.

Moreover, we feature multi-hop questions that indicate a model's robustness to sequential reasoning and mimic how humans ask questions. They may be useful in real-world tasks such as '*If the checkbox on page 1 section 3a indicates that the company is incorporated, how much yearly revenue did it generate in 2022 (given the table on page 5)?*'

Answers. Even though some VQA datasets are deliberately limited to questions of exclusively extractive (SP-DocVQA) or abstractive (VisualMRC) nature, others do not obey such restrictions and include both question types (see Table 1). The dataset we provide includes both abstractive and extractive answers, covering various types such as *textual, numerical, dates, yes/no, lists, or no answer*.

This allows us to cover all possible business use cases and reveal major deficiencies of existing DU systems beyond typical textual answers. For instance, no existing VQA dataset includes not answerable questions and questions answered with a list. In turn, the models considered to date supposedly tend to make unreliable guesses on questions with an answer not entailed by the content [77]. Our dataset is designed to cover answers beyond plain extractive text such as a list of items or even 'None'.

The 'None' answer type demands that the model correctly identifies that the answer cannot be provided, as the question needs to be better formed, e.g., it asks about the value of an empty cell in the table. In addition, list generation problems pose challenges to the model, as (1) more tokens need to be generated, (2) they may be sourced from different places in the document, and (3) OCR reading order may influence the element ordering.

3.1. Gathering Documents

A fundamental difficulty in gathering raw source files was ensuring dataset diversity while fulfilling strict licens-

[3] Answer type comparison is included in supplementary materials.

ing requirements. Therefore, rather than depending on initial sources of files, e.g., libraries that originally published digitized materials, we resorted to aggregate websites.

The document collection process was manual and assumed formulating queries to archive.org (containing 36M books and texts), commons.wikimedia.org (with 86M media types of various types), and documentcloud.org (with around 5M public documents). The queries consisted of keywords relevant to some category of interest, e.g., the *resume* category of our proposal consists of 'resume', 'cv', 'curriculum', and 'biography' keywords). Where necessary, a separate query parameter ensured that the resulting files belonged to the public domain or were released under a permissive license. Information on keywords and the search procedure is distributed as a part of the DUDE dataset.

From the resulting documents, we selected those representing the requested category and visually distinctive from the ones already gathered. Special care was put into removing examples that visibly expose controversial content or may be subject to privacy or legal concerns, despite the declared license. We collected five thousand, typically multi-page, English documents using this methodology.

3.2. Annotation Process

The annotation process involved in-house annotators and Amazon Mechanical Turk freelancers. For the latter, there is limited control over the expertise, and where justified, we resorted to limiting task availability depending on the number of completed tasks and historical acceptance rate.[4] The former are five highly qualified people with a Ph.D. in Linguistics. These three annotation scenarios will be referred to as *All MTurkers*, *Best MTurkers*, and *Qualified Linguists*.

We estimate the total cost of annotation involving both *Linguists* and *MTurkers* as $20,000.

Phase 1. We started by providing *All MTurkers* documents described in Section 3.1 in separate batches aimed at collecting abstractive, extractive, and list QA pairs. Each freelancer was asked to propose up to five questions of a particular type, and in the case of extractive ones to provide an evidence bounding box. The exception to this process is the annotation of non-answerable questions previously shown to be particularly challenging [77]. These are predominantly annotated by *Qualified Linguists* and because of their quality promoted without passing through Phases 2-3.

Candidate QA pairs are semi-automatically filtered to exclude annotations that cannot be valid due to the length, use of non-typical character combinations, or type-specific criteria, such as non-list answers for list batches. Additionally, we cluster duplicate and near-duplicate question-answer pairs to ensure dataset diversity and promote them directly to Phase 3 after a manual review (the same QA pairs provided independently by several annotators indicate their validity).

[4]Approval above 97% over at least 5k HITs.

Phase 2. The rest of the annotations promoted from Phase 1 were directed to *All MTurkers*, but this time instead of providing complete QA pairs, they were asked to answer the question from the previous round. Obtained triples of questions and two answer variants (one from each phase) were evaluated using inter-answer ANLS (defined in Section 3.5) promoted to the final dataset if the agreement was >0.8. Otherwise, QA triples were directed to Phase 3.

Phase 3. *Best MTurkers* were provided with document, question, and answer variants to decide the correctness of each answer and optionally overrule both variants if they are not correct. Outliers from decisions in this phase, such as repealing without a judgment on previous answers, were reviewed by *Qualified Linguists* and corrected if needed.

Optional Phase 4. Annotations of the test set were reviewed by *Qualified Linguists*. Given data from Phase 3, they corrected questions, answers and created metadata related to diagnostic categories described in Section 3.4.

3.3. Dataset Statistics

We conducted a statistical analysis of our dataset and found that the distribution of document length, question length, and answer type was much more diverse than in other datasets in the same domain. We also used the Simpson diversity coefficient [81] for analysis and summarized the results in Table 1. The following are the statistics for the data split:

	train	val	test (diagnostic)
documents	3,010	749	1,215 (530)
questions	23,728	6,315	11,448 (2,462)

Table 2: Data split counts.

The number of tokens in the document distribution is much more diverse compared to other datasets, a consequence of the more diverse distribution of pages (see Figure 3). Note some of the documents are more visual than textual (or even visual-only), making the left whisker essentially reach 0 ($\log_2$-scaling of x-axis).

The distribution of the number of tokens in answers is heavy-tailed, to some extent this is also the property of the distribution of number of tokens in questions. Furthermore, 90.9% of questions are unique, and so are 70.7% of answers (taking answer variants into account).

Dataset	Ours	SP-DocVQA	VisualMRC	InfographicsVQA	TAT-DQA
		Dataset-level properties			
Sources	Multi	Industry docs	Web pages	Infographics	Finance reports
Origin	BD, Scan	Mostly scans	BD	BD	BD
Period	1860-2022	1960-2000	Jan-Mar 2020	not specified	2018-2020
Documents	5,019	12,767	10,234	5,485	2,758
Pages (*avg±std*)	5.72±6.4	1.0±0.0	1.0±0.0	1.0±0.0	1.11±0.32
Tokens (*avg±std*)	1,831.53±2,545.06	183±149.96	154.19±79.34	287.98±214.57	576.99±290.12
Simpson coeff. (ResNet)	0.82	0.76	0.83	0.86	0.73
Simpson coeff. (Tf-Idf)	0.95	0.93	0.99	0.94	0.15
		Question-level properties			
Questions	41,541	50,000	30,562	30,035	16,558
Unique (%)	90.9	72.34	96.26	99.11	95.65
Length (*avg±std*)	8.65±3.35	8.34±3.04	9.38±4.01	11.57±3.71	12.51±4.18
Semantics	All	T, L, F, Ch	T, L, F, Ch	T, L, F, Ch, M	T, L
		Answer-level properties			
Unique (%)	70.7	64.29	91.82	48.84	77.54
Length (*avg±std*)	3.35±6.1	2.11±1.67	8.38±6.36	1.66±1.43	3.44±7.20
Extractive (%)	42.39	100.0	0.0	71.96	55.72
Abstractive (%)	38.25	0.0	100.0	24.91	44.28
List (%)	6.62	0.0	0.0	5.69	0.0
None	12.74	0.0	0.0	0.0	0.0

Table 1: Summary of the existing English document datasets and our challenge. `BD` stands for born-digital. Layout semantics are abbreviated as (T)able, (L)ist, (F)igure, (Ch)art, and M(ap). Comparison based on Azure Cognitive Services (3.2) OCR.

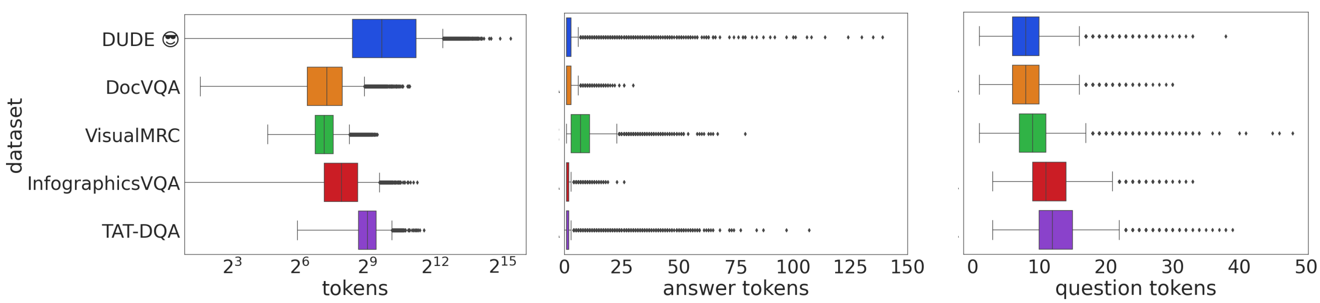


Figure 2: Distribution of the number of tokens in documents, answers, and questions.

We scrutinized the answer types by aggregating possible answers into classes representing the information they conveyed. The study used heuristics to determine if the answers fit into NER labeling scheme [1] or categories we anticipated, such as *yes/no* and *none*, or did not anticipate, such as *color*. This resulted in 25 different groups of answers, with the *other* answer type being the fourth largest group. Cramer's V coefficient was used to check for correlations between question types and answer types, and the results indicated that there were few correlations (see Appendix D.1). The expected correlations, such as *none* answers with *not-answerable* questions or *yes/no* answers with *abstractive* questions, were present, but barely any correlation was significant. This suggests it is hard to guess the answer based on the question solely.

We study relative diversity measure, called Simpson coefficient [111, 81]. To define it, consider a fixed distance function $d(a_1, a_2)$ defined for pair of documents $a_1, a_2 \in A$: the dataset. In our applications, it is the cosine similarity of a document embedding. Further, for an arbitrary number of datasets $A_1, \ldots, A_N$ the diversity of A_1 with respect to $A_2, \ldots, A_N$ is defined as

$$\mathrm{Div}_{A_2,\ldots,A_N}(A_1) = 1 - p\Big(d(a_{11}, a_{12}) < \min_{i=2:N} d(a_{i1}, a_{i2})\Big)$$

where $a_{i1}, a_{i2} \in A_i$, are randomly selected, $i = 2 : N i = 2 : N$. We report relative diversities of each of the datasets, relative to other datasets in the study, based on two embeddings: visual (ResNet-101 embeddings-based) and se-

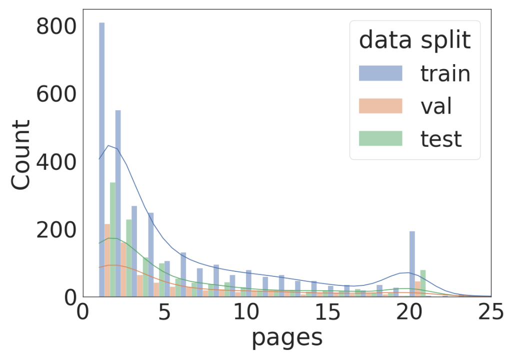


Figure 3: While other datasets are predominantly single-page only, the number of pages featuring in **DUDE** is more diverse, yet still biased towards shorter documents.

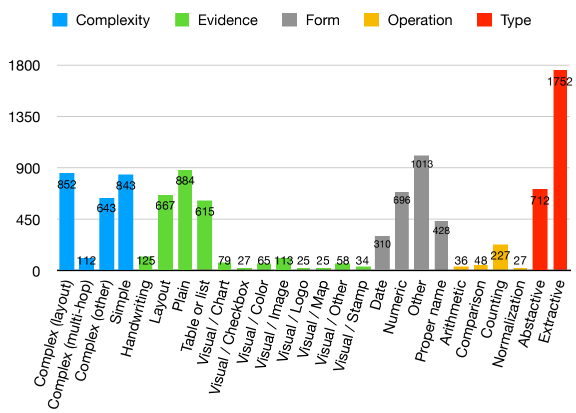


Figure 4: Count of particular diagnostic categories in a subset of 2.5k test set QA pairs annotated in detail to help analyze models' performance.

mantic (Tf-Idf embeddings-based), in Table 1. The results show that the probability that two random documents from **DUDE** are more similar than each random pair of documents from other datasets is small , meaning that documents in our dataset are well-distributed and diverse.

3.4. Diagnostic Subsets

Following previous DU datasets, we gather diagnostic metadata for close to half of the documents and QA pairs in the test set (see Figure 4). These are intended to enable a fine-grained analysis of the models' performance. The taxonomy used is an extension of the one from earlier works [57, 56, 10], covering **DUDE**-specific questions and enables a more detailed examination of visual artifacts under consideration.

Question type and perceived complexity. We distinguish questions perceived as *simple*, i.e., those based on spotting value near a phrase mentioned explicitly as a part of the question. For example, "Who is the Secretary of the U.S. Department of Commerce?" when the document contains "Penny Pritzker, Secretary, U.S. Department of Commerce." Such could be guessed given an approximate string matching algorithm and does not require much comprehension beyond that. The remaining questions are marked as *hard* with distinguished categories of *hard multi-hop questions*, and *hard meta/layout-navigating questions*.

Answer evidence. We provide information on what types of elements have to be comprehended to provide an answer, including *free text*, *handwriting*, *table or list*, and *layout*, i.e., non-tabular spatial understanding of text placement. These follow the ontology established by previous works [57, 56, 10]. In addition, we supply hints on graphical artifacts one needs to consider for particular questions, such as *image/photo*, *plot/chart*, *checkbox*, and *annotation*.

Required operation. We distinguish *arithmetic*, *comparison*, *counting*, and *normalization* operations to provide information on the need for performing, respectively, arithmetic operations on extractable data, comparing numerical values or sizes, counting elements or converting data present in the document to another format (e.g., rounding or date format conversion).

Answer form/shape. Finally, we provide information on the shallow form of the returned answer, including *date*, *numeric*, and *proper name*.

3.5. Evaluation

The evaluation process follows the typical paradigm of separate training, validation, and test splits. We provide both a standalone evaluator and a website[5] [95] to submit test set predictions.

To assess models' performance, we rely on the ANLS metric introduced by authors of the ST-VQA dataset [8]. Roughly speaking, it is a generalization of accuracy that does not penalize the system for an answer whose similarity to the gold standard measured with normalized Levenshtein similarity is above a specified threshold. Moreover, the metric assumes the presence of multiple, equally valid reference answers. The mentioned properties account for possible OCR errors or different phrasings, such as the same numerical answer represented as *two* and *2* by different annotators.

In practice, production DU systems provide an estimation of confidence in order to triage documents that do not need to be manually reviewed by a human. While the reliability of the automation ability of a DU solution is deemed quintessential for generating business value in practice [11], DU research rarely reports any confidence evalu-

[5] rrc.cvc.uab.es/?ch=23

ation. Some exceptions are in closely related task domains like scene text recognition [82] and QA [38, 106].

With DUDE, we want to establish calibration evaluation and confidence ranking as a default evaluation methodology in DU, especially since the field is so close to applications.

To this end, we report (next to ANLS) two additional metrics, Expected Calibration Error (ECE) [65, 63, 27], and Area-Under-Risk-Coverage-Curve (AURC) [24, 34].

Calibration requires that the probability a model assigns to its predictions equals their true likelihood of being correct [19, 20, 102].

ECE approximates top-1 calibration error by a weighted average over the accuracy/confidence difference of histogram bins. Particularly in our evaluation setting, we consider a predicted answer correct if its ANLS to the ground truth answer is above a pre-defined threshold (τ=0.5). For consistency, not-answerable and list-answers both have confidence estimated for the answer as a whole (regardless of the number of answers). Following [66], we apply equal-size binning (with 100 bins, $\mathcal{L}_p norm = 1$), avoiding some pathologies of equal-range binning [41, 94].

AURC is a selective classification metric that evaluates how well an estimator prevents silent failures on an *i.i.d* test set. As an aggregate measure of estimator performance (ANLS) and confidence ranking, it provides a more practically useful estimate of overall performance when the estimator can abstain from (low-confidence) decisions and defer to a human for feedback.

By reporting the above metrics, we hope that in future work there will be contributions (e.g., calibration methods for improved forecasting or metrics for better predictive uncertainty evaluation) that concretely target the empirical observations of overconfidence/miscalibration in DU models.

3.6. Baselines

Human performance. To establish the human baseline, we assign test set questions to *Qualified Linguists*, ensuring none of them will face the same documents as reviewed in Phase 4. The procedure results in an estimation of 74.76 ANLS points (Table 3). At first glance, this result seems low. Still, when analyzing results case by case, it turns out that it's hard to score much better since the answer format can influence the overall results a lot: *Eagle* vs. *an eagle* (0.625 ANLS), *62%* vs. *62* (0.67 ANLS), *1958-04-29* vs. *4-29-58* (0 ANLS), *Clemson University, Clemson South Carolina* vs. *Clemson University* (0 ANLS). We achieved the lowest performance (67.58) on the extractive question type, which confirms our hypothesis since the abstractive answers are shorter (mostly numbers, yes/no, or colors).

We analyzed the maximum score achieved by the best-performing model for each diagnostic test category and plotted that against the human performance in Figure 5.

Reference models. We assessed a group of modelsto determine how their performance is influenced by different factors such as (1) their ability to handle textual, layout, and visual elements, (2) whether they were fine-tuned for the task, (3) their size in (trainable parameters), and (4) the maximum input length they can handle.

To analyze factors (1) and (2), we conducted a zero-shot evaluation of several baseline text-only models. We used three encoder-based models (BERT [21], Longformer [6], and BigBird [103]) that cannot generate text and three that feature a decoder (T5 [75], GPT-3-Davinci [12], and ChatGPT) and have this capability. Next, we extended the T5 architecture with 2D layout embeddings [10, 72] and fine-tuned models with increasing maximum sequence lengths (512 $\rightarrow$ 8192) on **DUDE**. Finally, we evaluated our replication of the hierarchical Hi-VT5 model [90], as this model has the ability to decode text, understand multi-page layouts, and comprehend visual page features using DiT [47].

Regarding factors (2) and (3), we evaluated models of various sizes ranging from 131M (BigBird) to 175B (GPT-3-Davinci) and varied the input context from 512 (BERT) to 20480 (Hi-VT5) tokens. Overall, we thoroughly evaluated multiple models in the different testing setups to determine their performance under various conditions, as seen in Table 3.

3.7. Analysis & Discussion

To summarize, our study reveals that existing advanced language models such as BERT, Longformer, and BigBird struggle with comprehending visual elements and document layouts. To address this issue, we introduced T5, T5-2D, and Hi-VT5 models that incorporate layout and visual information. Still, their performance remains unsatisfactory, as evidenced by the comparison with the human baseline, similar to what has been reported for InfographicsVQA. This indicates that there is still scope for enhancing the visual understanding of **DUDE** models. Moreover, our findings indicate that a large LLM capable of processing long inputs alone is insufficient for achieving strong performance in **DUDE**, especially for the extractive type of answer. Finally, the dataset's length significantly affects the models' scores, as seen by the increase in scores by $4.4 - 5.0$ points when the T5 and T5+2D context length is extended from 512 to 8192. Similarly, the model size has a positive correlation with the final score, but it holds only within a particular model-type and is not the main factor influencing the results. State-of-the-art performance of $46.04\ \text{ANLS}_{\text{all}}$ was achieved on $T5_{large}$ with a 2D layout understanding that consumed 8192 tokens, confirming the observation above.

4. Conclusion

In conclusion, this paper introduces a new large-scale multi-paged, multi-domain, multi-industry Document Vi-

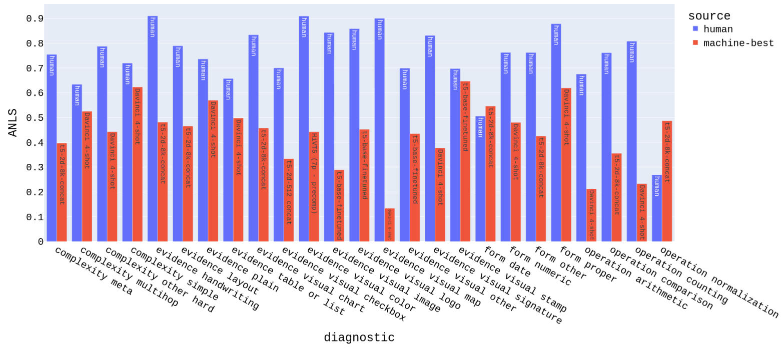


Figure 5: We report the average ANLS for the human expert vs. the best-performing model per diagnostic category as a ceiling analysis.

Model	Init.	Params	Max Seq. Length	Test Setup	$\text{ANLS}_{\text{all}} \uparrow$	$\text{ECE}_{\text{all}} \downarrow$	$\text{AURC}_{\text{all}} \downarrow$	ANLS_{do}	ANLS_{do} Abs	ANLS_{do} Ex	ANLS_{do} NA	ANLS_{do} Li
text-only Encoder-based models												
Big Bird	MPDocVQA	131M	4096	Concat*	26.27	30.14	44.22	30.67	7.11	40.26	12.75	8.46
BERT-Large	MPDocVQA	334M	512	Max Conf.*	25.48	34.06	48.60	32.18	7.28	42.23	5.88	11.13
Longformer	MPDocVQA	148M	4096	Concat*	27.14	27.59	44.59	33.45	8.55	43.58	10.78	10.62
text-only Encoder-Decoder based models												
T5	base	223M	512	Concat-0*	19.65	19.14	48.83	25.62	5.24	33.91	0	7.31
T5	MPDocVQA	223M	512	Max Conf.*	29.48	27.18	43.06	37.56	21.19	44.22	0	10.56
T5	base	223M	512	Concat+FT	37.41	**10.82**	41.09	40.61	42.61	48.20	53.92	16.87
T5	base	223M	8192	Concat+FT	41.80	17.33	49.53	44.95	47.62	50.49	63.72	7.56
text-only Large Language models (LLM)												
ChatGPT	gpt-3.5-turbo	20B	4096	Concat-0	-	-	-	35.07	16.73	42.52	70.59	15.97
				Concat-4	-	-	-	41.89	22.19	49.90	**77.45**	17.74
GPT3	davinci3	175B	4000	Concat-0	-	-	-	43.95	18.16	54.44	73.53	36.32
				Concat-4	-	-	-	47.04	22.37	**57.09**	63.73	**40.01**
text+layout Encoder-Decoder based models												
T5-2D	base	223M	512	Concat+FT	37.10	10.85	41.46	40.50	42.48	48.62	52.94	3.49
T5-2D	base	223M	8192	Concat+FT	42.10	17.00	48.83	45.73	48.37	52.29	63.72	8.02
T5-2D	large	770M	8192	Concat+FT	**46.06**	14.40	**35.70**	**48.14**	**50.81**	55.65	68.62	5.43
text+layout+vision models												
HiVT5		316M	20480	Hierarchical+FT	23.06	11.91	54.35	22.33	33.94	17.60	61.76	6.83
LayoutLMv3	MPDocVQA	125M	512	Max Conf.*	20.31	34.97	47.51	25.27	8.10	32.60	8.82	7.82
Human baseline								74.76	81.95	67.58	83.33	67.74

Table 3: Summary of Baseline performance on the **DUDE** test set ($_{all}$) and diagnostic subset ($_{do}$). Test setups are defined as *Max Conf.*: predict one answer per page and return an answer with the highest probability over all pages, *Concat*: predict on tokens truncated to maximum sequence length, *FT* stands for fine-tuning on **DUDE** training data, and *-0* refers to zero-shot and *-4* few-shot inference. Average ANLS results per question type are abbreviated as (Abs)tractive, (Ex)tractive, (N)ot-(A)nswerable, (Li)st. (*) We report only results for best performing test setup (either *Max Conf.* or *Concat*). All scalars are scaled between 0 and 100 for readability.

sual Question Answering Benchmark for document understanding. Our dataset is adjusted to the real-world environment where we need to process long documents and understand different types of documents. The benchmark includes visual semantics such as *tables, charts, figures, lists, checkboxes, stamps*, and more, which are essential for real-world document understanding. The performance of state-of-the-art textual and multi-modal models still lags behind human performance, indicating the need for further improvement in visual understanding for DU models. Nevertheless, we believe evaluating systems on **DUDE** could inspire new architectures and methods.

Limitations. As our approach is closer to real-world industrial applications, and enables models to recognize and understand new unseen data without the need for retraining, it does come with some limitations and constraining factors, including the use of only English language documents. Future work could address these limitations and expand the benchmark to include other languages. Moreover, although our dataset can be considered large-scale, it still represents a relatively small sample size of the plethora of documents that exist in the real world.

References

[1] SpaCy en_core_web_lg label scheme. https://spacy.io/models. Accessed: 2023-03-08. 5

[2] Rajas Agashe, Srinivasan Iyer, and Luke Zettlemoyer. JuICe: A large scale distantly supervised dataset for open domain context-based code generation. 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 5436–5446, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2015. 2

[4] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms, 2019. 2

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 2

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 7, 1

[7] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019. 2

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 2, 6, 3, 4

[9] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online, Nov. 2020. Association for Computational Linguistics. 2

[10] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 6, 7

[11] Pascal Bornet, Ian Barkin, and Jochen Wirtz. *Intelligent automation: Welcome to the world of hyperautomation: learn how to harness artificial intelligence to boost business & make our world more human*. World Scientific, 2021. 6

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7

[13] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. LifeQA: A real-life dataset for video question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France, May 2020. European Language Resources Association. 2

[14] Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. In-the-wild video question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. 2

[15] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps, 2022. 2

[16] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online, Aug. 2021. Association for Computational Linguistics. 2

[17] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. TutorialVQA: Question answering dataset for tutorial videos. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5450–5455, Marseille, France, May 2020. European Language Resources Association. 2

[18] Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv preprint arXiv:1908.05803*, 2019. 2

[19] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. 7

[20] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 7

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

[22] Ritam Dutt, Kasturi Bhattacharjee, Rashmi Gangadharaiah, Dan Roth, and Carolyn Rose. PerKGQA: Question answering over personalized knowledge graphs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 253–268, Seattle, United States, July 2022. Association for Computational Linguistics. 2

[23] Lukasz Garncarek, Rafal Powalski, Tomasz Stanislawek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip

Graliński. Lambert: Layout-aware language modeling using bert for information extraction. In *ICDAR*, 2021. 2

[24] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 7, 4

[25] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50, 2021. 2

[26] Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. A question answering approach for emotion cause extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2

[27] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330, 2017. 7

[28] Deepak Gupta and Dina Demner-Fushman. Overview of the MedVidQA 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2

[29] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. 2

[30] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 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, 2015. 2

[31] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. 2

[32] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States, July 2022. Association for Computational Linguistics. 2

[33] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 2

[34] Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *International Conference on Learning Representations*, 2023. 7, 1, 4

[35] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 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, 2019. 2

[36] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[37] Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, Jens Lehmann, and Maria Maleshkova. Conversational question answering over knowledge graphs with transformer and graph attention networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 850–862, Online, Apr. 2021. Association for Computational Linguistics. 2

[38] Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, 2020. 7

[39] Sanjay Kamath, Brigitte Grau, and Yue Ma. Verification of the Expected Answer Type for Biomedical Question Answering. In *First International Workshop on Hybrid Question Answering with Structured and Unstructured Knowledge (HQA'18)*, WWW '18 Companion Proceedings of the The Web Conference 2018, pages 1093–1097, Lyon, France, Apr. 2018. ACM Press. 2

[40] Andreas Kirsch. Player of jeopardy: Chatgpt evaluation, 2023. 3

[41] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019. 7, 4

[42] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019. 2

[43] Egor Lakomkin, Sven Magg, Cornelius Weber, and Stefan Wermter. KT-speech-crawler: Automatic dataset construction for speech recognition from YouTube videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 90–95, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. 2

[44] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[45] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966. 3

[46] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, 2022. 2

[47] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022. 7

[48] Jing Li, Shangping Zhong, and Kaizhi Chen. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[49] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis, 2020. 2

[50] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 2

[51] Chenxiao Liu and Xiaojun Wan. CodeQA: A question answering dataset for source code comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2618–2632, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[52] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 2

[53] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[54] Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering, 2020. 2

[55] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. Qalayout: Question answering layout based on multimodal attention for visual question answering on corporate document. In Seiichi Uchida, Elisa Barney, and Véronique Eglin, editors, *Document Analysis Systems*, pages 659–673, Cham, 2022. Springer International Publishing. 3

[56] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 3, 6

[57] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020. 2, 3, 6

[58] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions, 2020. 2

[59] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 2

[60] Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2

[61] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. 2

[62] Muhammad Akhtar Munir, Muhammad Haris Khan, M Saquib Sarfraz, and Mohsen Ali. Towards improving calibration in object detection under domain shift. In *Advances in Neural Information Processing Systems*, 2022. 4

[63] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 7

[64] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview of BioASQ 2022: The tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Lecture Notes in Computer Science*, pages 337–361. Springer International Publishing, 2022. 2

[65] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 625–632, 2005. 7

[66] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019. 7, 4

[67] Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online, July 2020. Association for Computational Linguistics. 2

[68] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image, 2020. 2

[69] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 2

[70] Tzuf Paz-Argaman and Reut Tsarfaty. RUN through the streets: A new dataset and baseline models for realistic urban navigation. 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 6449–6455, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[71] Michał Pietruszka, Michał Turski, Łukasz Borchmann, Tomasz Dwojak, Gabriela Pałka, Karolina Szyndler, Dawid Jurkiewicz, and Łukasz Garncarek. Stable: Table generation framework for encoder-decoder models, 2022. 2

[72] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. In *ICDAR*, 2021. 2, 7

[73] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. DuReader$_{\text{vis}}$: A Chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2

[74] Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. T5score: Discriminative fine-tuning of generative evaluation metrics. *arXiv preprint arXiv:2212.05726*, 2022. 4

[75] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 7

[76] Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online, June 2021. Association for Computational Linguistics. 2

[77] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. 2, 3, 4

[78] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2

[79] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR, 2022. 4

[80] Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676, Online, Aug. 2021. Association for Computational Linguistics. 2

[81] E. H. SIMPSON. Measurement of diversity. *Nature*, 163(4148):688–688, apr 1949. 4, 5

[82] Ron Slossberg, Oron Anschel, Amir Markovitz, Ron Litman, Aviad Aberdam, Shahar Tsiper, Shai Mazor, Jon Wu, and R Manmatha. On calibration of scene-text recognition models. *arXiv preprint arXiv:2012.12643*, 2020. 7

[83] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022. 2

[84] Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3157–3164, New York, NY, USA, 2020. Association for Computing Machinery. 2

[85] Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In *ICDAR*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer, 2021. 2

[86] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images, 2023. 2

[87] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021. 2, 3

[88] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 2

[89] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021. 2, 3, 4

[90] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022. 2, 3, 7, 1

[91] Rubèn Tito, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2021 competition on document visual question answering. In *International Conference on Document Analysis and Recognition*, pages 635–649. Springer, 2021. 2

[92] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016. 2

[93] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. Lc-quad: A corpus for complex question answering over knowledge graphs. In *International Semantic Web Conference*, pages 210–218. Springer, 2017. 2

[94] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019. 7, 4

[95] Jordy Van Landeghem, Lukasz Borchmann, Rubèn Tito, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *Proceedings of ICDAR 2023*, 2023. 6

[96] Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. Towards fine-grained causal reasoning and qa, 2022. 2

[97] Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, generalization and beyond. *arXiv preprint arXiv:2203.09513*, 2022. 2

[98] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. 2

[99] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. 2

[100] Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States, July 2022. Association for Computational Linguistics. 2

[101] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020. 2

[102] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002. 7

[103] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020. 7, 1

[104] Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. ParsFEVER: a dataset for Farsi fact extraction and verification. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 99–104, Online, Aug. 2021. Association for Computational Linguistics. 2

[105] Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. NOAHQA: Numerical reasoning with interpretable graph question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[106] Shujian Zhang, Chengyue Gong, and Eunsol Choi. Knowing more about questions can help: Improving calibration in question answering. *arXiv preprint arXiv:2106.01494*, 2021. 7

[107] Xinbo Zhang, Changzhi Sun, Yue Zhang, Lei Li, and Hao Zhou. NAIL: A challenging benchmark for na\"ive logical reasoning, 2022. 2

[108] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021. 2

[109] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 564–580. Springer, 2020. 2

[110] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 2

[111] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. 5

[112] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022. 2, 3