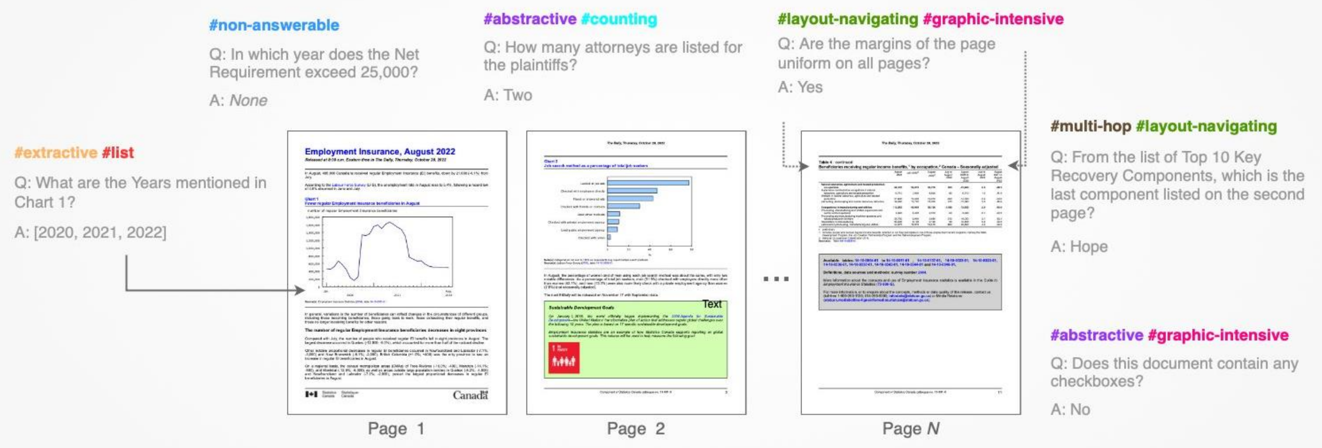
ICDAR 2023 Competition on Document UnderstanDing of Everything

SAN JOSE, CALIFORNIA, USA 2023

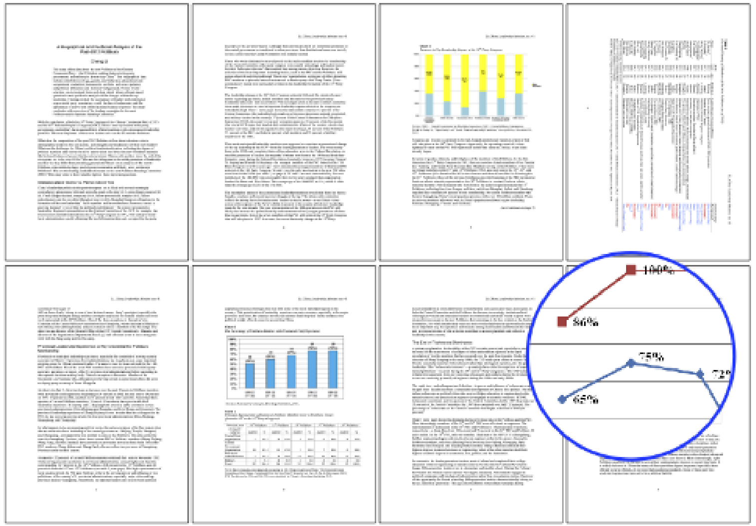
Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek.



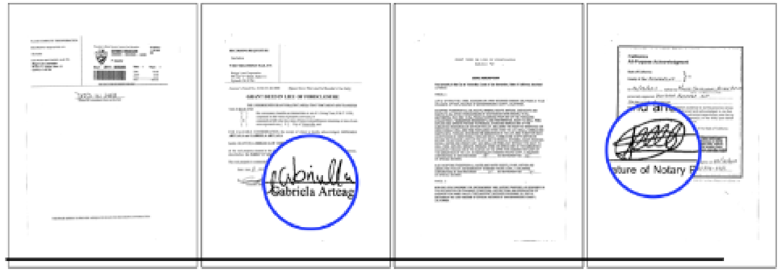
QA as a natural language interface to Visually-Rich Documents

-Everything-, you mean?

Visual evidence (chart). *What is the maximum percentage of the blue graph line on page 8?* A highly demanding question that requires simultaneous competency of visual comprehension (locating chart and line color), navigating through layout (determining adequate page), and numerical comparison (deciding on the highest value).

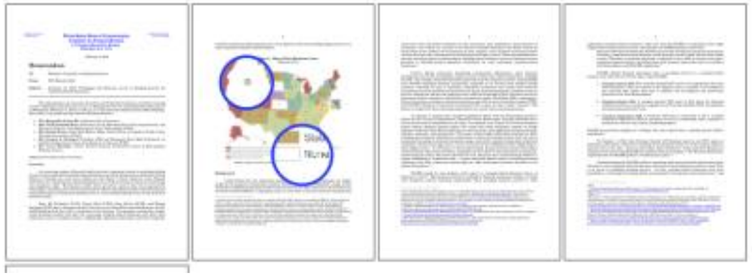


Requires counting. *How many pages have a signature?* The question requires visual comprehension (recognition of signature), knowledge about layout, and counting.

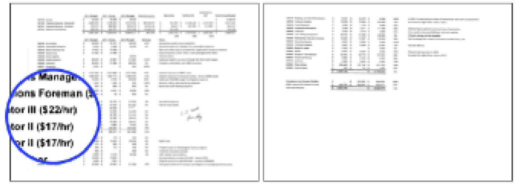


Source	Answer	ANLS	Conf.
Ground truth	*2*		
Human	*2*	1.0	—
T5	*1*	0.0	0.01
ChatGPT	*4*	0.0	—
GPT3	*[Not-answerable]*	0.0	—
T5-2D	*4*	0.0	0.69
HiVT5	*4*	0.0	0.41

Visual evidence (map), multi-hop. *Which states don't have any marijuana laws?* The multi-hop question requires visually comprehending the map and linking knowledge from its legend with depicted regions.



Requires arithmetic. *What is the difference between how much Operator II and Operator III makes per hour?* The question requires table comprehension, determining relevant values, dividing extracted integers, and correcting the subject-verb agreement.



Outline

1. DUDE: the project 😎
    - Scope and objectives
2. DUDE: the dataset
    - Summary and statistics
    - Evaluation and baselines
3. DUDE: the competition
    - Competition protocol
    - Submissions and final ranking
4. DUDE: what’s next?

KU LEUVEN

DUDE **Project** 😎

Building a long-standing document understanding benchmark incorporating real-world complexities

Objective/Scope

- Foster research on *generic* document understanding (DU)

- Adopting task paradigm of **Document Visual Question-Answering** and learning paradigm of **Multi-Domain Long-Tailed Recognition**

→ Handle complexity & variety of *real-world* documents and subtasks
→ Generalization to *any documents* and *any questions*

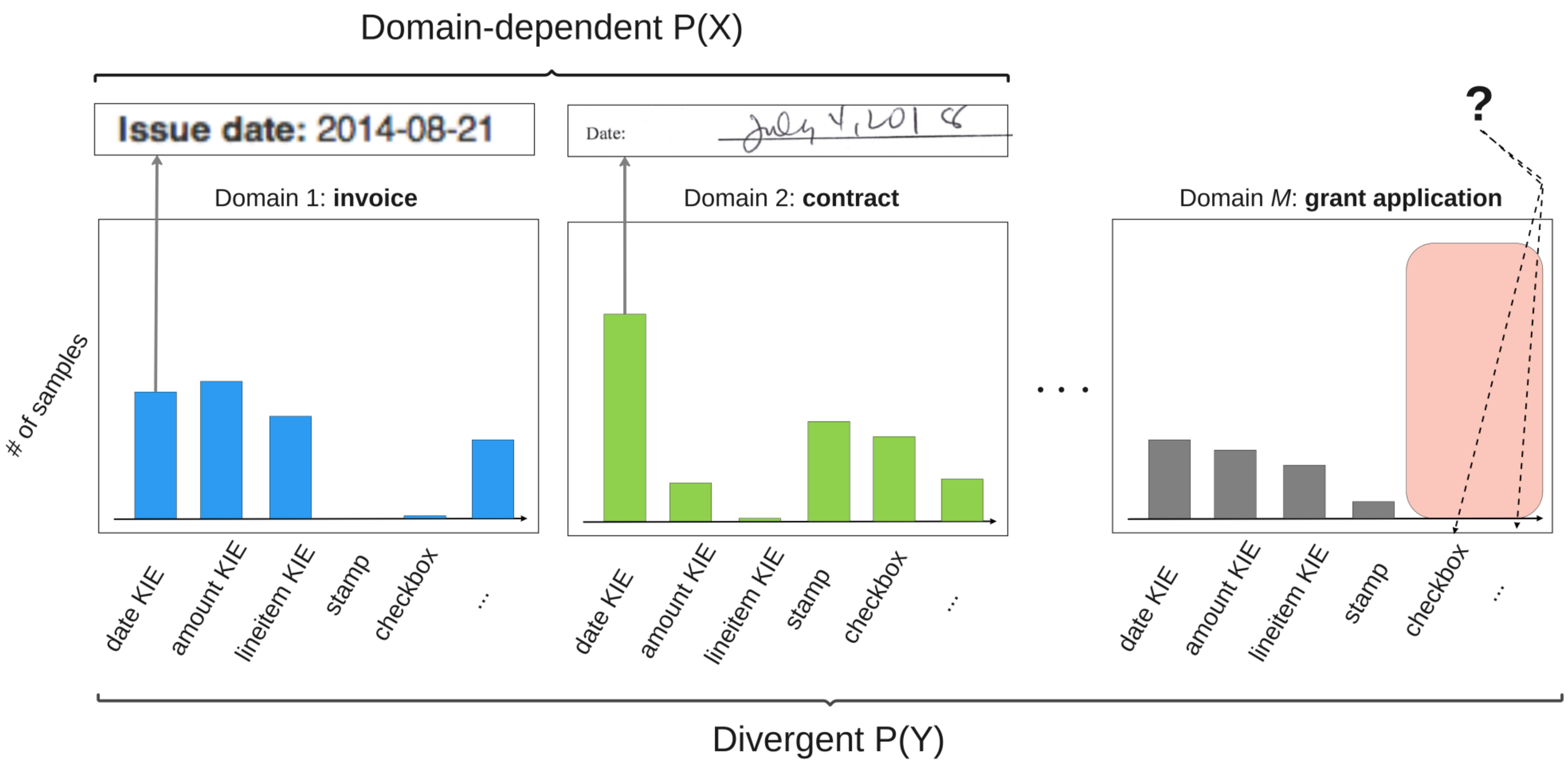
Faculty of Engineering Science, Department of Computer Science, HCI unit, LIIR lab

DocVQA & MDLT

X: documents
Y: QA pairs

Domain: document type

→ Subtask adaptation under low-resource setting
→ Innovation in multi-modal, transfer learning, and zero-shot generalization



Novelty ~ *Why DUDE?*

- The rise of LLMs and their applicability (?) to document understanding
- Publicly available datasets avoid/do not include:
  - **multi-page** documents
  - **multi-industry** documents of sufficiently different types
  - **multi-task** settings
    - CLF, KIE, DLA, HWR, …
- Bridging QA & DLA:
  - Layout semantics (stamp, signature, font style, checkbox)
  - Complex layout-navigating questions demanding multi-step reasoning

Meet the DUDEs










KU LEUVEN
contract.fit
snowflake
CVC Centre de Visió per Computador
La Rochelle Université
INSTABASE

Setting the records straight

- Van Landeghem, J., Borchmann, L., Tito, R., Pietruszka, M., Jurkiewicz, D., Powalski, R., Józiak, P., Biswas, S., Coustaty, M., Stanisławek, T. (2023). **ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE)**. *Proceedings of ICDAR 2023*.

Competition details
Ranked methods
Final ranking

- Van Landeghem, J., **. . .**, *Anckaert, B., Valveny, E., Blaschko, M, Moens, M. F*, & Stanisławek, T. (2023). **Document Understanding Dataset and Evaluation (DUDE)**. *International Conference of Computer Vision 2023*.

Dataset detail stats
Baselines
Evaluation metrics

DUDE **Dataset**

Constructing a multi-faceted resource that challenges the DAR community

Dataset summary

- Sourced a dataset with 40K QA pairs for 5K permissive license documents
  - **multi-source** (*archive, wikimedia, documentcloud*)
  - **multi-domain** (+15 industries)
  - **multi-type** (+- 200 document types)
  - **multi-page** (μ=5 pages)
  - **multi-QA** (extractive, abstractive, list, non-answerable)
  - **multi-origin** (1900-2023)
- Multi-stage annotation process with freelancers and qualified linguists
- Provide three OCR versions (Tesseract – Azure – AWS)

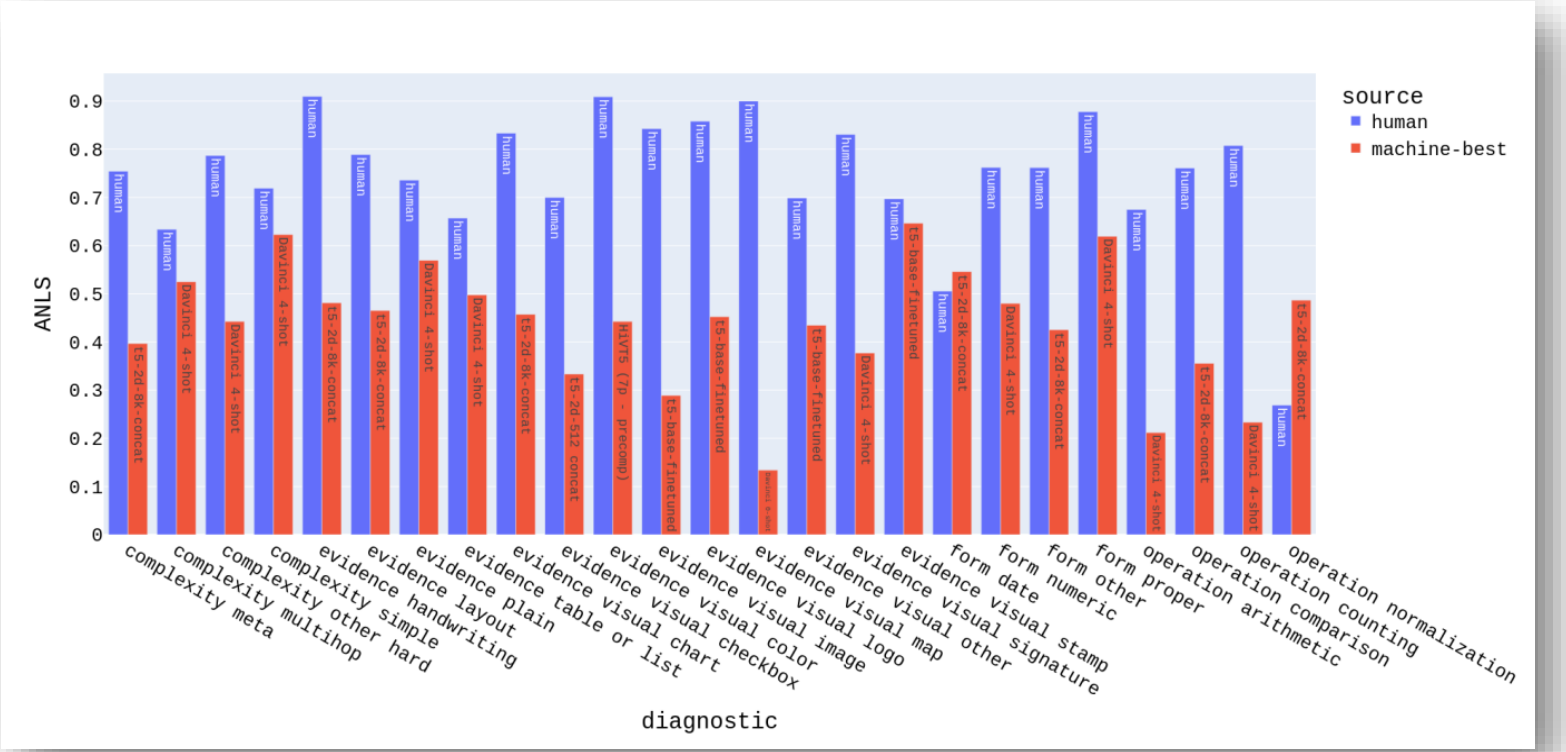
Baselines

Model	Init.	Params	Max Seq. Length	Test Setup	$ANLS_{all}$ ↑	ECE_{all} ↓	$AURC_{all}$ ↓	$ANLS_{do}$	$ANLS_{do}$ 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

I. Generative = must
II. Strong performance of LLMs
III. Stronger performance by models
+layout understanding
++longer sequence length

SOTA ANLS < 50% !

Diagnostic categories performance



Diagnostic categories with

- visual evidence
- reasoning operations

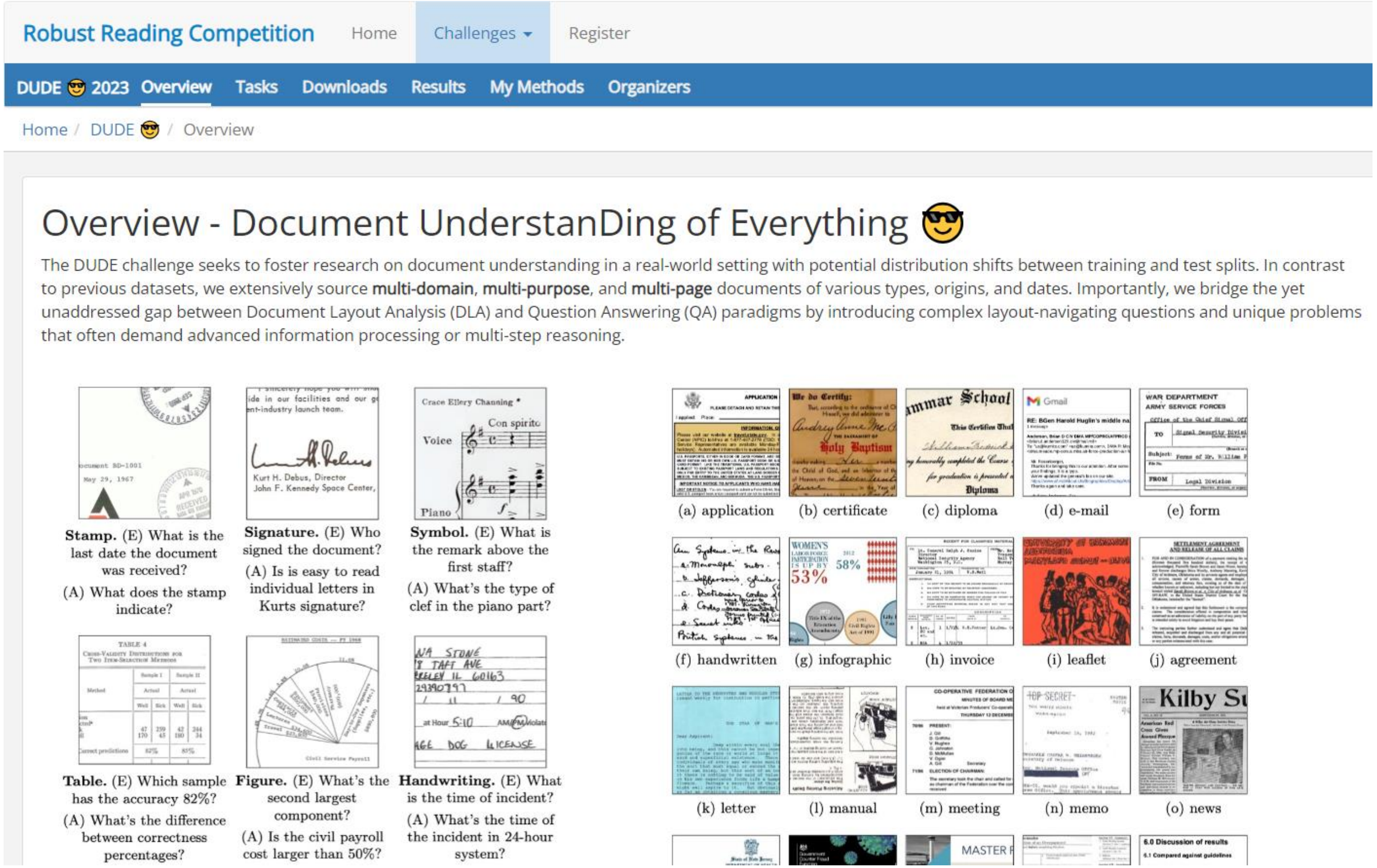
Baselines lagging far behind **human baseline**

KU LEUVEN

DUDE **Competition**

Introducing Document UnderstanDing of Everything

ICDAR 2023 DUDE 😎 Competition



Website:
https://rrc.cvc.uab.es/?ch=23

Timeline: February – May 2023

Protocol:

- *Trainval* (30K-3.7K): February
- *Test* (11.4K-1.3K) March-May

JSON submissions ⇔ model binaries

Incentives

- By design of the dataset and competition → <u>force</u> significant novelty
- Measuring improvements closer to the real-world applicability of DU models

→ **calibrated** and **selective** DocVQA

- Lower answer confidence if unsure about answer correctness
- Refrain from hallucinations on non-answerable questions

Faculty of Engineering Science, Department of Computer Science, HCI unit, LIIR lab

KU LEUVEN

Task formulation

What are the first two behavioral and intellectual disabilities of people with FASDs?

GT: Learning disabilities | Hyperactivity

hyperactivity | speech and language delays
0.9298765
0

- <u>Given</u>:
  - Natural language question (on content, aspect, form, visual/layout)
  - Input document
  - A set of reference answers
- <u>Provide</u>:
  - **Natural language answer**
  - Answer Confidence (float between 0 and 1)
  - Abstention flag (1 for abstaining)

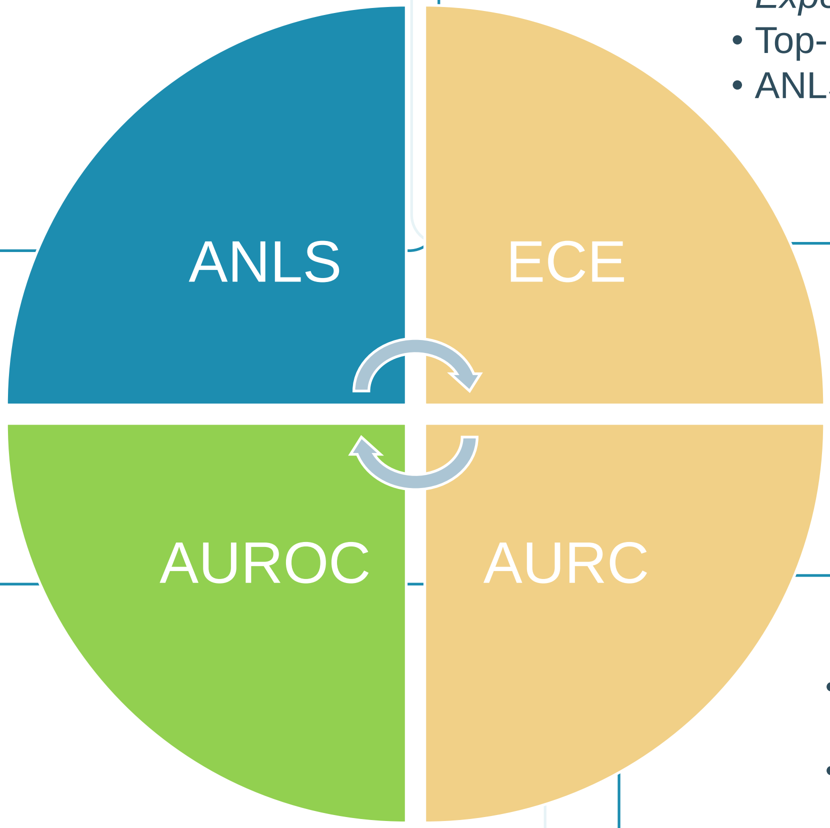
Evaluation methodology

Appendix B.4.

- *Average Normalized Levenshtein Similarity*
- Modified for NA & lists

$$\mathrm{LD}(G,\hat{P}) = \begin{cases} 1 & \text{if } \mathrm{NA}(G) \wedge |\hat{P}| > 0, \\ 0 & \text{if } \mathrm{NA}(G) \wedge |\hat{P}| = 0, \\ |G| & \text{if } |\hat{P}| = 0, \\ \mathrm{LD}(\mathrm{tail}(G), \mathrm{tail}(\hat{P})) & \text{if } G[0] = \hat{P}[0], \\ 1 + \min \begin{cases} \mathrm{LD}(\mathrm{tail}(G), \hat{P}) \\ \mathrm{LD}(G, \mathrm{tail}(\hat{P})) \\ \mathrm{LD}(\mathrm{tail}(G), \mathrm{tail}(\hat{P})) \end{cases} & \text{, otherwise} \end{cases}$$

- *Expected Calibration Error*
- Top-1 prediction miscalibration
- ANLS thresholding discretization



$$\mathbb{I}[\mathrm{ANLS}(y,\hat{y}) > \tau]$$

- *Area-Under-Receiver-Operating-Characteristic*
- Detect out-of-domain test documents

- *Area-Under-Risk-Coverage Curve*
- Selective QA as confidence ranking

Competition Submissions

- Document foundation models
  - UDOP, HiVT5
- LLM or VLMs
  - ChatGPT, BLIP2
- Multi-stage pre-training on VQA data
  - SP/MP-DocVQA, VQAonBD
  - ScienceQA, HotpotQA
- Token embeddings for DU subtasks

Method	Description
LENOVO RESEARCH	
UDOP(M)	Ensemble (M=10) of UDOP [30] (794M each) models without self-supervised pre-training, only fine-tuned in two stages: 1) SP-DocVQA [33] and MP-DocVQA [32], and 2) DUDE (switching between Azure and AWS OCR).
UDOP +BLIP2	UDOP(M=1) with integrated BLIP2 [17] predictions to optimize the image encoder and additional page number features.
UDOP +BLIP2+GPT	UDOP(M=1) and BLIP2 visual encoder with ChatGPT to generate Python-like modular programs to decompose questions for improved predictions [9,6].
UPSTAGE AI	
MMT5	Multimodal T5 pre-trained in two stages: single-page (ScienceQA [28], VQAonBD2023 [27], HotpotQA [35], SP-DocVQA) with objectives (masked language modeling (MLM) and next sentence prediction (NSP)), multi-page (MP-DocVQA and DUDE) with three objectives (MLM, NSP, page order matching). Fine-tuning on DUDE with answers per page combined for final output.
INFRRD.AI	
HiVT5	Hi-VT5 [32] with 20 <PAGE> tokens pre-trained with private document collection (*no information provided*) using span masking objective [14]. Fine-tuned with MP-DocVQA and DUDE.
HiVT5 +modules	Hi-VT5 extended with token/object embeddings for a variety of modular document understanding subtasks (detection: table structure, signatures, logo, stamp, checkbox; KIE: generic named entities; classification: font style).

Competition Final Ranking

	Answer	**Calibration**		**OOD Detection**	ANLS / **answer type**			
Method	ANLS ↑	ECE ↓	AURC ↓	AUROC ↑	*Ex*	*Abs*	*Li*	*NA*
UDOP+BLIP+GPT	**50.02**	**22.40**	**42.10**	**87.44**	**51.86**	**48.32**	**28.22**	**62.04**
MMT5	37.90	59.31	59.31	50.00	41.55	40.24	20.21	34.67
HiVT5+modules	35.59	28.03	46.03	51.24	30.95	35.15	11.76	52.50

Congratulations to **Lenovo Research**
@ Ren Zhou, Qiaoling Deng, Xinfeng Chang, Luyan Wang, Xiaochen Hu, Hui Li, Yaqiang Wu

DUDE: **What’s Next?** ⏭

Future outlook: **the challenge is still on!**

- **Confidence estimation**, **calibration** and **selective generation** is unmined territory, while DUDE offers a proper benchmark for evaluating advances
- Need for **better metrics** than ANLS over multiple references
  - e.g., taking semantic equivalence into account (it’s Paris == the capital of France)
- With the rise of **multi-modal LLMs** (e.g., Kosmos-2, GPT-4), better solutions are coming, yet due to its designed complexity, DUDE might remain “the benchmark to beat” for a long time
- The multi-page aspect is not sufficiently addressed
  - Inefficiency for **long document processing**

Faculty of Engineering Science, Department of Computer Science, HCI unit, LIIR lab

Questions?
Future collaborations?
Ideas for extensions?

SAN JOSE, CALIFORNIA, USA 2023

WEBSITE

https://rrc.cvc.uab.es/?ch=23

DATASET

https://huggingface.co/datasets/jordyvl/DUDE_loader




Dataset statistics

- a broad spectrum of **document** types, domains, sources, and dates
- **questions** beyond document content, including operations and multi-hop
- varied **answer** types such as abstractive, extractive, lists and non-answerable

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