Document Understanding Dataset and Evaluation (DUDE 😎)

Jordy Van Landeghem[1,2], Rubén 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[3,4]

[1]KU Leuven, [2]Contract.fit, [3]Snowflake, [4]Warsaw University of Technology, [5]CVC, [6]Jagiellonian University, [7]Adam Mickiewicz University, [8]Instabase, [9]University of La Rochelle

Dataset | Competition

Overview

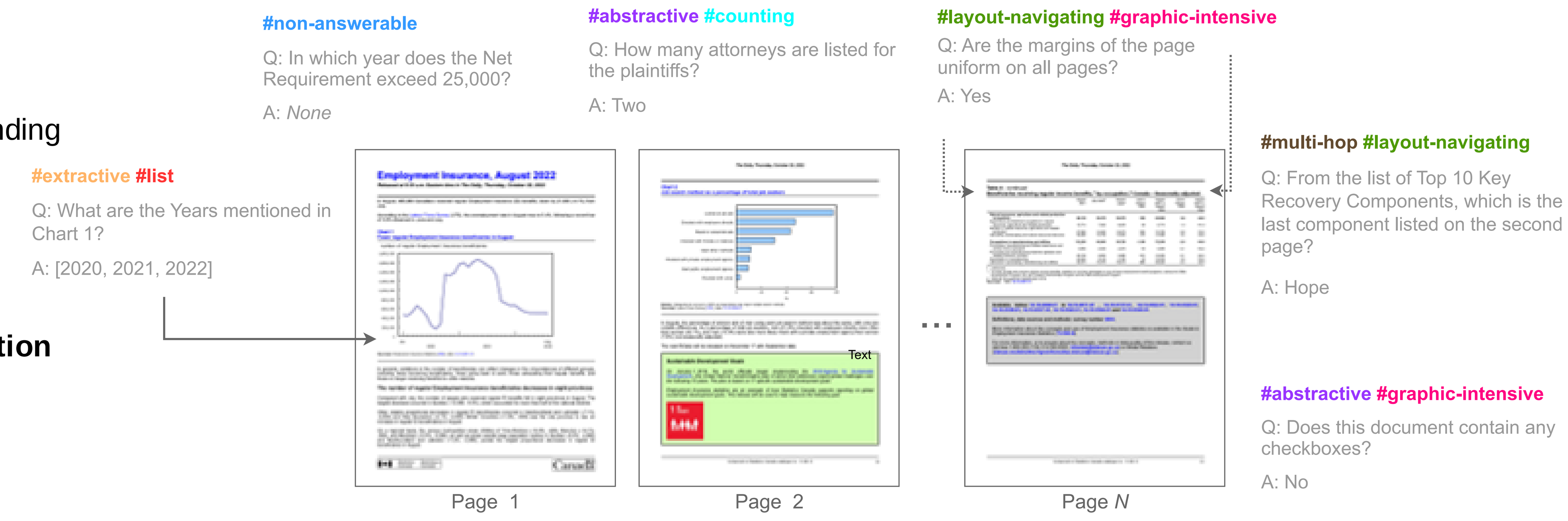
<u>Motivation</u>: *Question Answering as a natural language interface to Visually-Rich Documents*

<u>Objective</u>: Construct a **multi-faceted dataset** to foster research on *generic* Document Understanding

- Handle complexity and variety of ***real-world*** documents and subtasks
- Generalization to ***any* documents and *any* questions**
- Empirically question the ***applicability of LLMs (?)*** to Document Understanding

<u>Approach</u>: **DocVQA** task paradigm & learning paradigm of **Multi-Domain Long-Tailed Recognition**

- Incentivize questions on visual/layout semantics, layout navigation and multi-step reasoning
- Organically obtain questions relevant to the document type and instances



Dataset

Summary

DUDE 😎 collects +40K QA pairs for +5K documents

- **multi-page** (μ=6 pages)
- **multi-source** (*archive, wikimedia, documentcloud*)
- **multi-domain** (+15 industries)
- **multi-type** (+- 200 document types)
- **multi-QA** (extractive, abstractive, list, non-answerable)
- **multi-origin** (1860-2022)

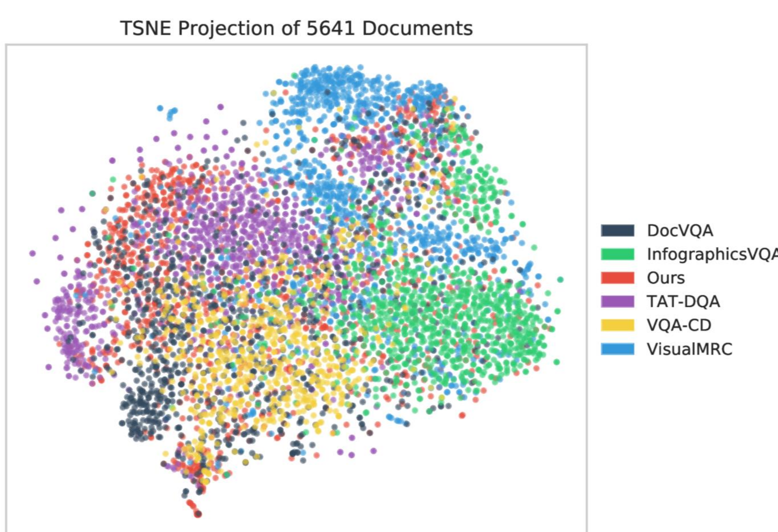
→ Multi-stage annotation process with freelancers and qualified linguists

→ Three OCR versions provided (*Tesseract – Azure – AWS*)

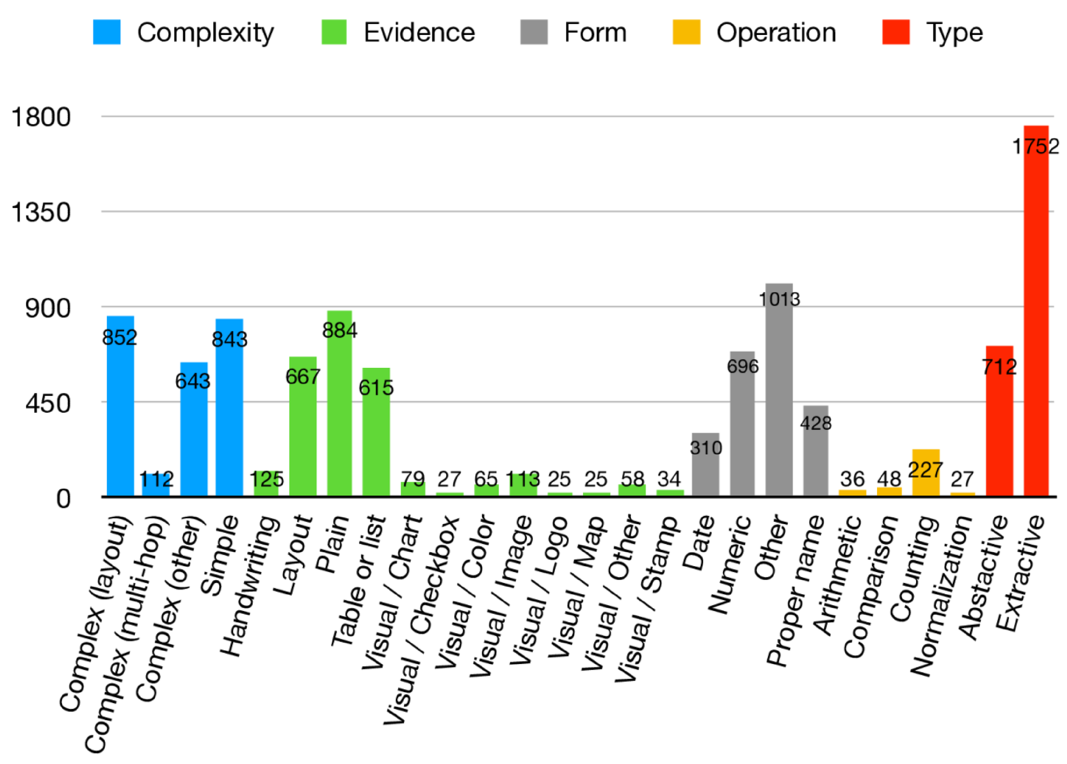
Comparing to existing datasets

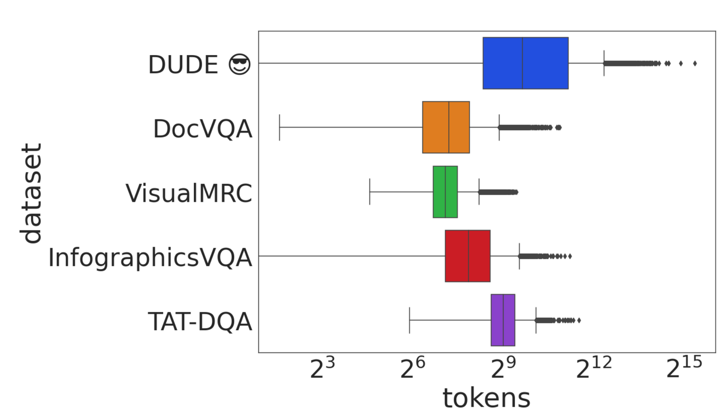
Dataset	Ours	SP-DocVQA	VisualMRC	InfographicsVQA	TAT-DQA
Dataset-level properties					
Sources	Multi	Industry docs	Web pages	Infographics	Finance reports
Origin	BD, Scan	Mostly scans	BD	BD	BD
Period	1860-2022	1960-2000	Jan-Mar 2020	not specified	2018-2020
Documents	5,019	12,767	10,234	5,485	2,758
Pages (*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

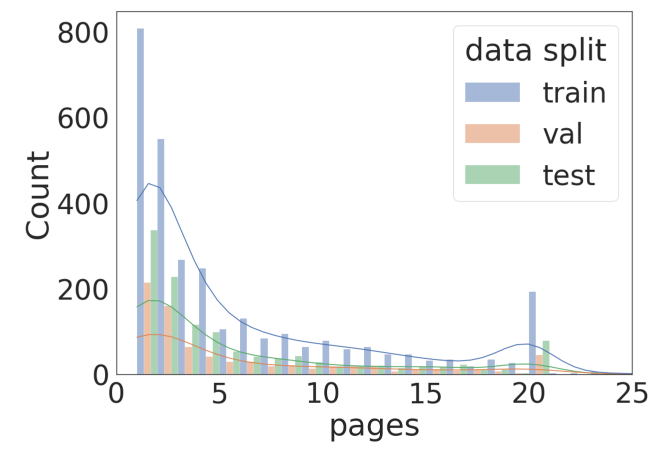
Document diversity



Diagnostic metadata







Evaluation

Task description

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

GT: Learning disabilities | Hyperactivity

hyperactivity | speech and language delays
0.9298765

<u>Given</u>:

- *Natural language question* (on content, aspect, form, visuals, layout)
- Input document
- A set of reference answers

<u>Provide</u>:

- **Natural language answer**
- **Answer Confidence**

Reference Models

1. **Text-only – encoder**
   BERT, Longformer, BigBird
2. **Text-only – +decoder**
   T5, GPT3-Davinci, ChatGPT
3. **Text+Layout – +decoder**
   T5-2D (512 → 8192)
4. **Text+Layout+Vision**
   LayoutLMv3, HiVT5

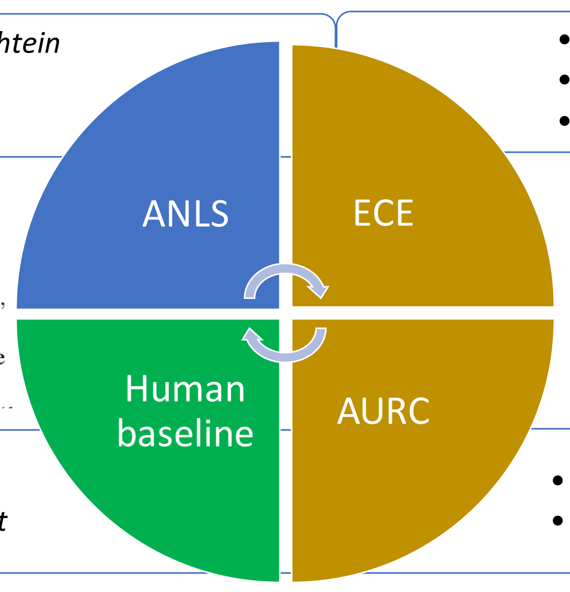
Evaluation methodology

- *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

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

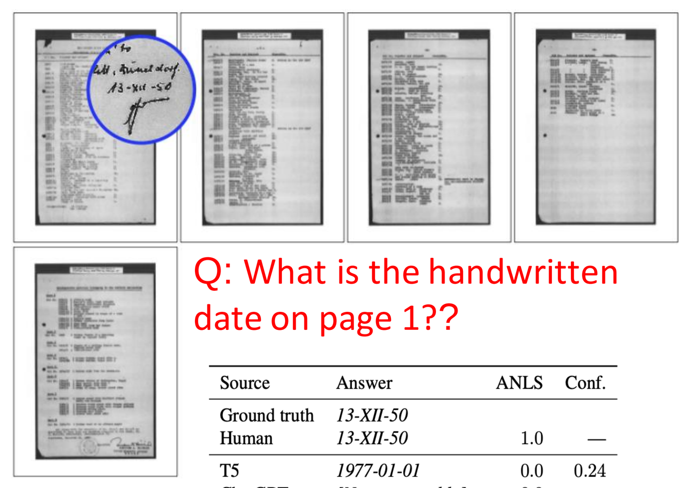


- Diagnostic annotations
- Quantify human *non-expert* performance with ANLS

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

Qualitative examples

Handwritten evidence
Requires arithmetic
Multi-hop visual evidence
Abstract artifacts



Q: What is the handwritten date on page 1??

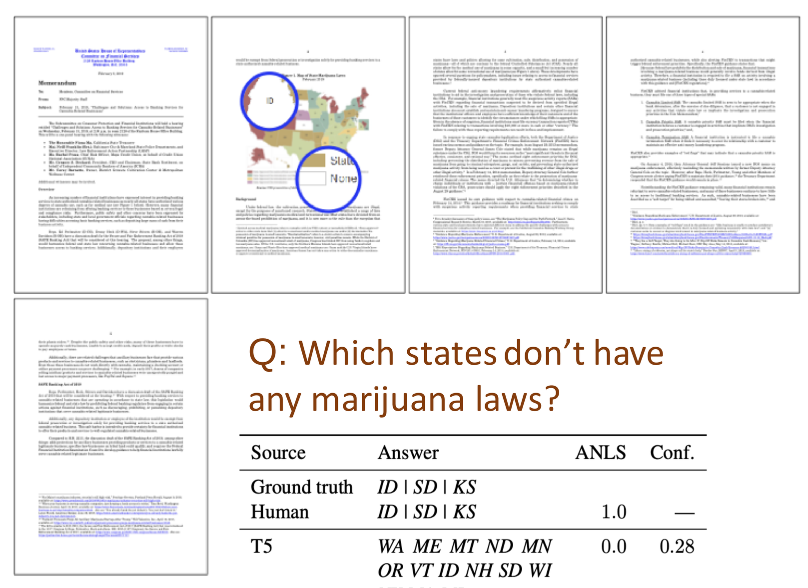
Source	Answer	ANLS	Conf.
Ground truth	*13-XII-50*		
Human	*13-XII-50*	1.0	—
T5	*1977-01-01*	0.0	0.24
ChatGPT	*[Not-answerable]*	0.0	—
GPT3	*15 December 1950*	0.0	—
T5-2D	*1950-12-15*	0.0	0.24
HiVT5	*1977-07-01*	0.0	0.11
BERTQA	*2006 / 1*	0.0	0.5

Q: Is there any redacted section on the document?

Source	Answer	ANLS	Conf.
Ground truth	*No*		
Human	*No*	1.0	—
T5	*yes*	0.0	0.17
ChatGPT	*[Not-answerable]*	0.0	—
GPT3	*[Not-answerable]*	0.0	—
T5-2D	*No*	1.0	0.43
HiVT5	*Yes*	0.0	0.55
LayoutLMv3	*approved for release*	0.0	0.01

Q: What is the difference between how much Operator II and Operator III makes per hour?

Source	Answer	ANLS	Conf.
Ground truth	*$5*		
Human	*$5*	1.0	—
T5	*200*	0.0	0.28
ChatGPT	*$5 per hour.*	0.0	—
GPT3	*Operator II ($17/hr) | Operator III ($22/hr)*	0.0	—
T5-2D	*[Not-answerable]*	0.0	0.31
HiVT5	*[Not-answerable]*	0.0	0.15



Q: Which states don't have any marijuana laws?

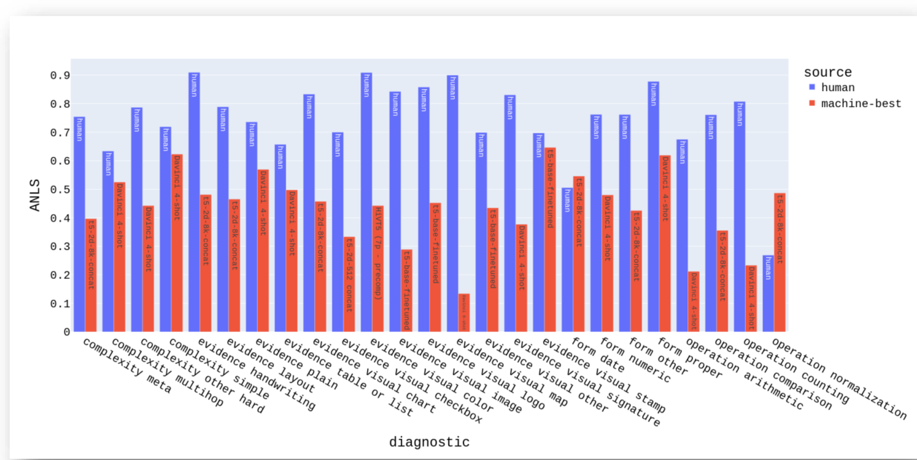
Source	Answer	ANLS	Conf.
Ground truth	*ID | SD | KS*		
Human	*ID | SD | KS*	1.0	—
T5	*WA ME MT ND MN OR VT ID NH SD WI NY MA MI*	0.0	0.28
ChatGPT	*[Not-answerable]*	0.0	—-
GPT3	*American Samoa*	0.0	—-
T5-2D	*i*	0.0	0.03
HiVT5	-	0.0	0.02

Reference model results

Model	Init.	Params	Max Seq. Length	Test Setup	$\mathrm{ANLS}_{\mathrm{all}}\uparrow$	$\mathrm{ECE}_{\mathrm{all}}\downarrow$	$\mathrm{AURC}_{\mathrm{all}}\downarrow$	$\mathrm{ANLS}_{\mathrm{do}}$	$\mathrm{ANLS}_{\mathrm{do}}$ Abs	$\mathrm{ANLS}_{\mathrm{do}}$ Ex	$\mathrm{ANLS}_{\mathrm{do}}$ NA	$\mathrm{ANLS}_{\mathrm{do}}$ 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 with
😎 visual evidence
😎 reasoning operations

Baselines lagging far behind **human baseline**

Future Work

- 😎 <u>Dataset extensions</u>:
  - **multilingual documents** and **cross-lingual questions**
  - **answer grounding annotations** and **question decomposition**
- 😎 **Confidence estimation, calibration** and **selective generation** for DocVQA
- 😎 Need for **better evaluation metrics** than ANLS over multiple references
  - e.g., taking semantic equivalence into account (it's Paris == the capital of France)
- 😎 Investigate solutions for **efficient processing of long, structured documents**