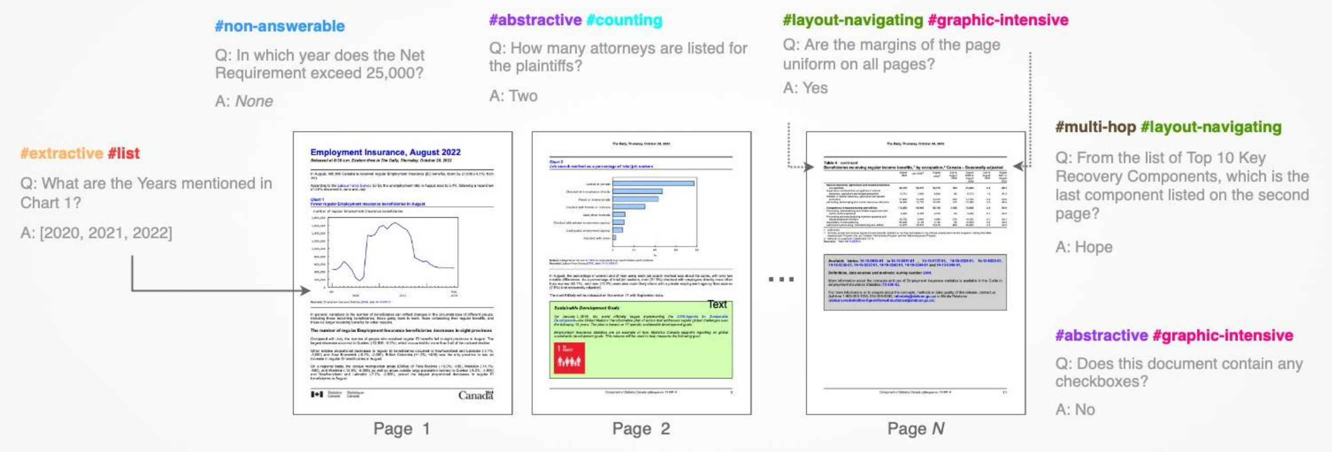
KU LEUVEN

Document UnderstanDing of Everything: **DUDE**, what's next?

Adobe

Jordy Van Landeghem



QA as a natural language interface to Visually-Rich Documents

Outline

1. Intelligent Automation for AI-driven Document Understanding
2. DUDE: the project
    - Scope and objectives
3. DUDE: the dataset
    - Summary and statistics
    - Evaluation and baselines
4. DUDE: the competition
    - Competition protocol – final ranking
5. DUDE: what’s next?

whoami

- Research intern ORACLE (*Seq2Seq for Dialogue Modeling*) and NUANCE (Language Modeling Algorithms)
- AI researcher contract.fit since 2017
- Ongoing Ph.D. project KU LEUVEN on
  Intelligent Automation ***(IA) for*** *Artificial Intelligence* ***(AI)-Driven*** *Document Understanding* ***(DU)***
  - Expected graduation: 02/2024 😌
- Research interests:
  document intelligence 😌
  dataset construction and evaluation methodology
  calibration, predictive uncertainty, failure prediction

More details: https://jordy-vl.github.io/

Lead up to my Ph.D. project

In <u>any business context</u>, where **information transfer** and **inbound communication services** are an important part of the day-to-day processes, a vast number of documents must be handled.

To provide customers with the *expected service levels* (in terms of *speed*, *convenience* and *accuracy*) a lot of time and resources are spent on manually **categorizing documents** and **extracting crucial information**.

contract.fit


(E)mails

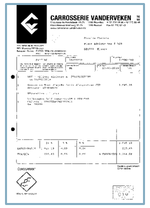

Attachments


Insurance policy


Car order


ID Card


Police report


Repair invoice


Accident form

KU LEUVEN

What makes automation intelligent?

Intelligent Automation (IA) comprises a compelling class of technologies:

- A subset of Artificial Intelligence (AI) for automation of knowledge work
- Robotic Process Automatic (RPA): the macro on steroids
- Workflow & Business Process Management (BPM)

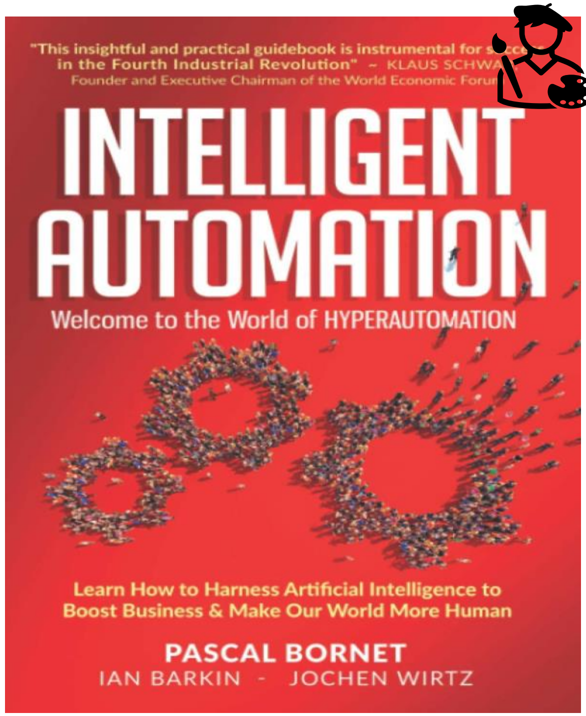
- jointly capable of solving major world problems <u>when combined with people & organizations</u>

- IA allows for the creation of a software-based **digital workforce**, by mimicking four main human capabilities required to perform **knowledge work**:
  1. **V**ision
  2. **L**anguage
  3. **T**hinking **&** **L**earning
  4. **E**xecution

build **straight-through** business processes, which are more efficient (**productivity, processing speed, cost)** and often more effective (**quality and logic**).

<u>Goal</u>: **Taking the robot out of the human**, not replacing human workers

"*AI done right […] will amplify human creativity, productivity and intelligence*" Adobe



<u>Pascal Bornet, Ian Barkin and Jochen Wirtz (2020)</u>

Motivating example: what are the key ingredients for IA?

From: jack.dunn@gmail.com
To: customer_admin@insurco.com
Subject: stop car policy 12-3456-789

In-distribution

Dear,

I would like to end the policy for my vehicle 1-CHA-123. The car has been sold and the license plate returned to the authorities (proof attached).

Kindly refund the unused portion of my premium.

Best,
Jack

From : jeff.smith@gmail.com
To : customer_admin@insurco.com
Subject : Jeff Smith hobby drone coverage

OOD

Dear,

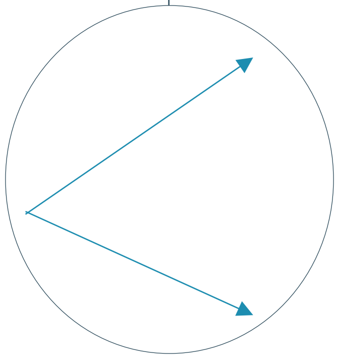
My car is already covered with your insurance company under the policy with number 23-4567-890.
Now, I would like to buy insurance coverage for my new hobby drone "DJI Mavic 2LBZ-548".
See attached the purchase receipt and below an illustration of the device and components.

Will you send me a proposal with monthly coverage fees?
Kind regards,

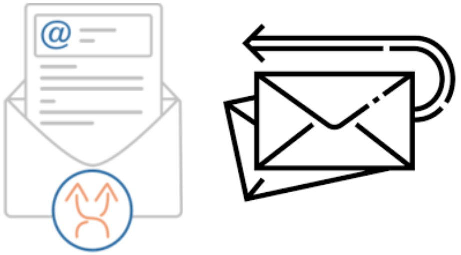
Jeff

Decision-making under Predictive Uncertainty

label	prediction	confidence
process	car policy cancellation	99%
policy number	12-3456-789	95%
license plate	1CHA123	98%



Automate action



label	prediction	confidence
process	**car policy contract start**	**98%**
policy number	23-4567-890	95%
license plate	2LBZ-548	75%

Manual review

! Catastrophically overconfident

Realizing intelligent automation

- Enabling IA involves:
  - Confidence estimation
  - Operational thresholding for automation-risk trade-off
  - Robustness to distribution shifts

- Measuring IA involves:
  - Calibration metrics
  - Confidence ranking

- Improving IA involves:
  - Inducing calibration by post-hoc strategies or designing calibrated loss functions
  - Predictive uncertainty estimation
  - Failure prediction

CVC seminar & calibration-primer 😉

Selected works

- Van Landeghem, J., Blaschko, M., Anckaert, B., & Moens, M. F. (2020). **Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification**. In *Workshop on Uncertainty and Robustness in Deep Learning*. ICML.
- Van Landeghem, J., Blaschko, M., Anckaert, B., & Moens, M. F. (2022). **Benchmarking Scalable Predictive Uncertainty in Text Classification.** In *IEEE Access*, vol. 10, pp. 43703-43737.
- 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)**. In *Proceedings of ICDAR 2023*.
- 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*.
- Van Landeghem, J., Biswas, S., (2023). **Beyond Document Page Classification**. In WACV 2024 (under review).

Ongoing projects:

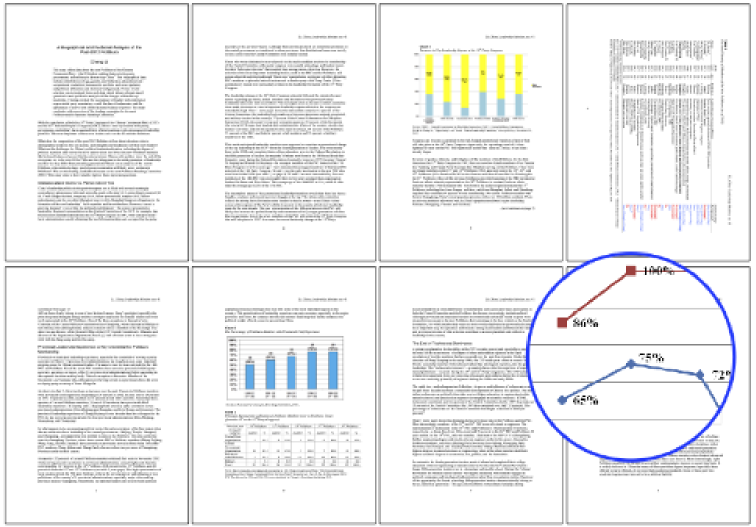
1. Knowledge Distillation for Document Foundation Models
2. A Multi-Modal Multi-Exit Architecture for Efficient Document Classification
3. A Differentiable Surrogate Loss for Selective Classification
4. DaDDa: Building the Dataset of Document Datasets

Document UnderstanDing of Everything: DUDE **Project** 😎

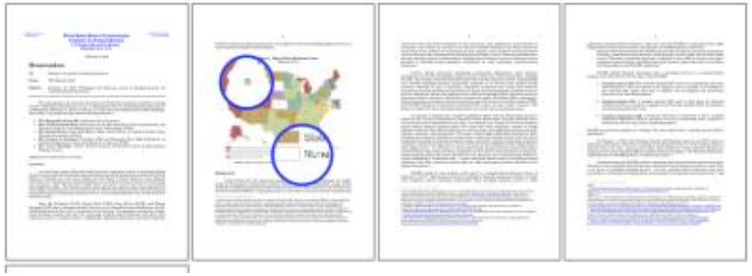
Building a long-standing document understanding benchmark incorporating real-world complexities

-Everything-, you mean?

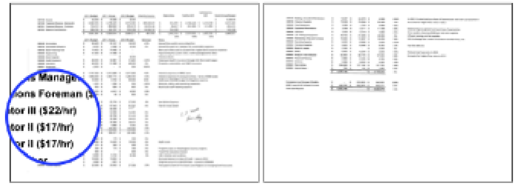
Visual evidence (chart). *What is the maximum percentage of the blue graph line on page 8?* A highly demanding question that requires simultaneous competency of visual comprehension (locating chart and line color), navigating through layout (determining adequate page), and numerical comparison (deciding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



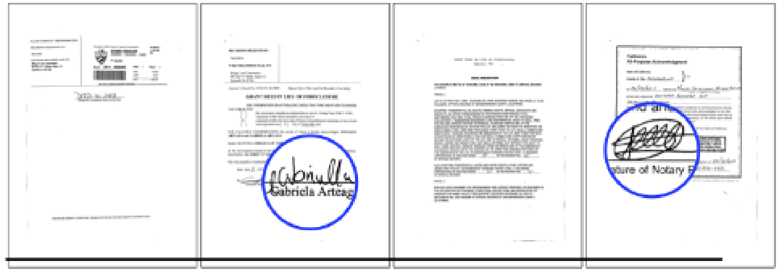
Visual evidence (map), multi-hop. *Which states don't have any marijuana laws?* The multi-hop question requires visually comprehending the map and linking knowledge from its legend with depicted regions.



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.



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

More examples

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**

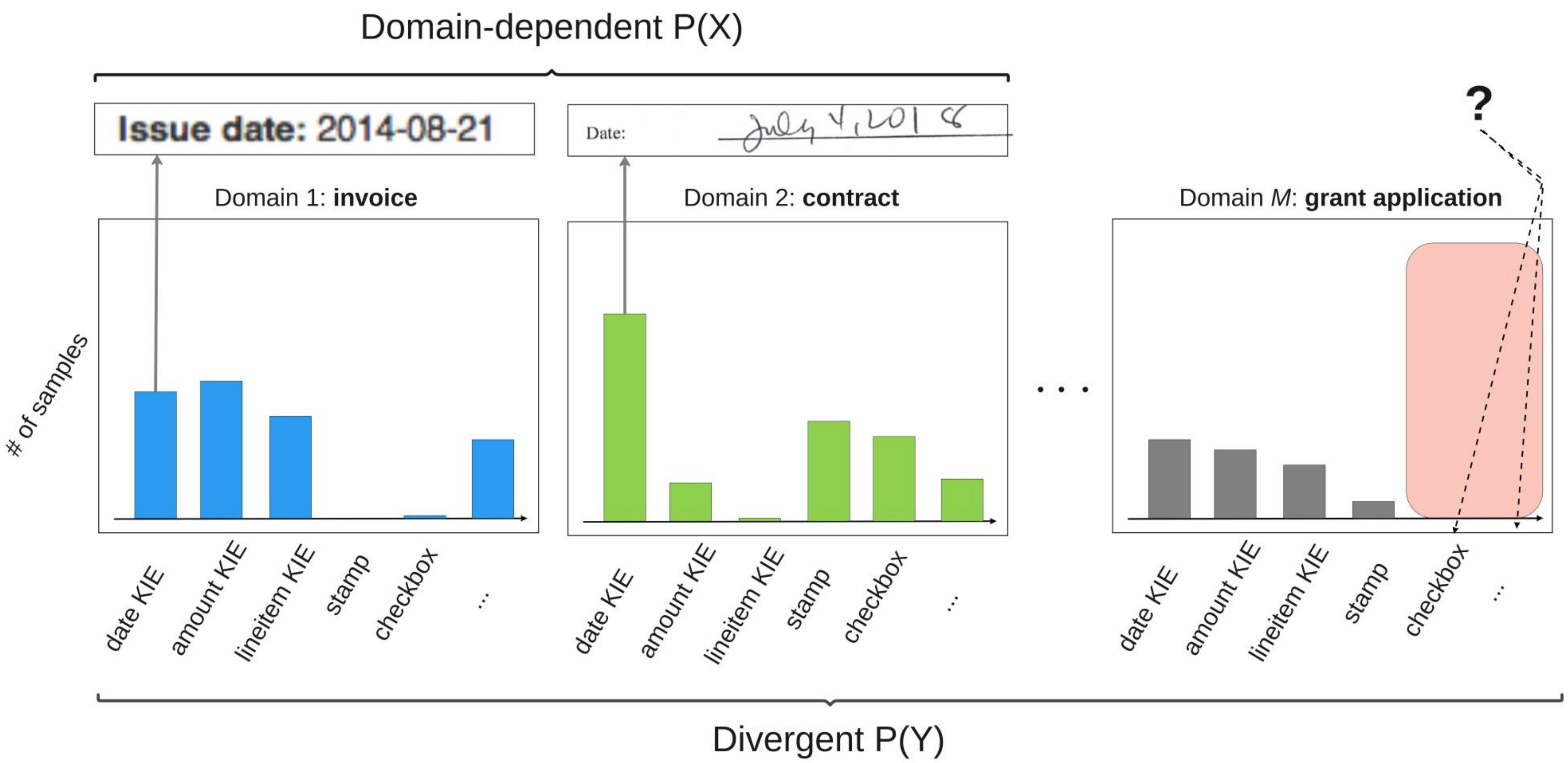
→ Handle complexity & variety of *real-world* documents and subtasks
→ Generalization to *any documents* and *any questions*

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, SV, …
- Bridging QA & DLA:
  - Layout semantics (stamp, signature, font style, checkbox, form fields, ….)
  - 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-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** (1860-2023)
- Multi-stage annotation process with freelancers and qualified linguists
- Provide three OCR versions (Tesseract – Azure – AWS)

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

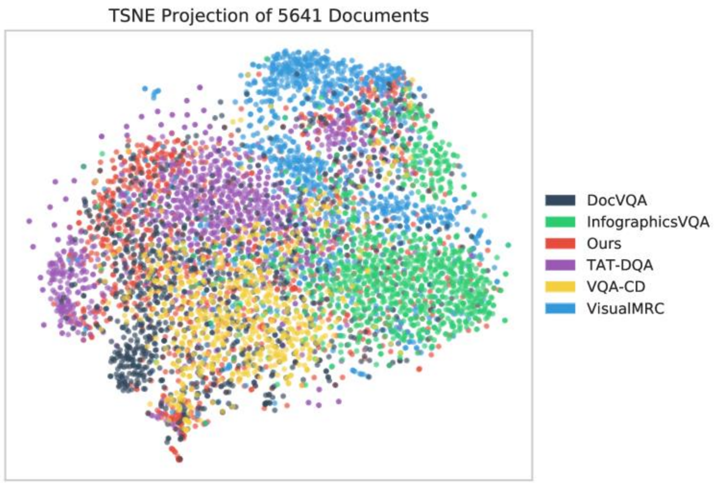
Dataset	Ours	SP-DocVQA	VisualMRC	InfographicsVQA	TAT-DQA
		Dataset-level properties			
Sources	Multi	Industry docs	Web pages	Infographics	Finance reports
Origin	BD, Scan	Mostly scans	BD	BD	BD
Period	1860-2022	1960-2000	Jan-Mar 2020	not specified	2018-2020
Documents	5,019	12,767	10,234	5,485	2,758
Pages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

Document diversity

```
    'banking':
'''
1. Bank statement - statement of account
2. Check - cheque
3. Deposit slip - deposit ticket
4. Credit card statement - statement of account
5. ATM receipt - cash withdrawal slip
6. Payment voucher - payment receipt
7. Transfer receipt - wire transfer receipt
8. Cashier's check - bank draft
9. Direct deposit form - direct deposit authorization form
10. Endorsement stamp - endorsement seal
11. Letter of credit - documentary credit
12. Promissory note - IOU
13. Debit memo - debit note, debit slip
14. Credit memo - credit note, credit slip, refund slip
```

- Approach:
  - Design industry-document taxonomy based on experience
  - Semi-automatically create document type keywords
    - 'Please list 30 common retail document types with their synonyms like Credit memos - {"credit notes", "credit slips", "refund slips"}'

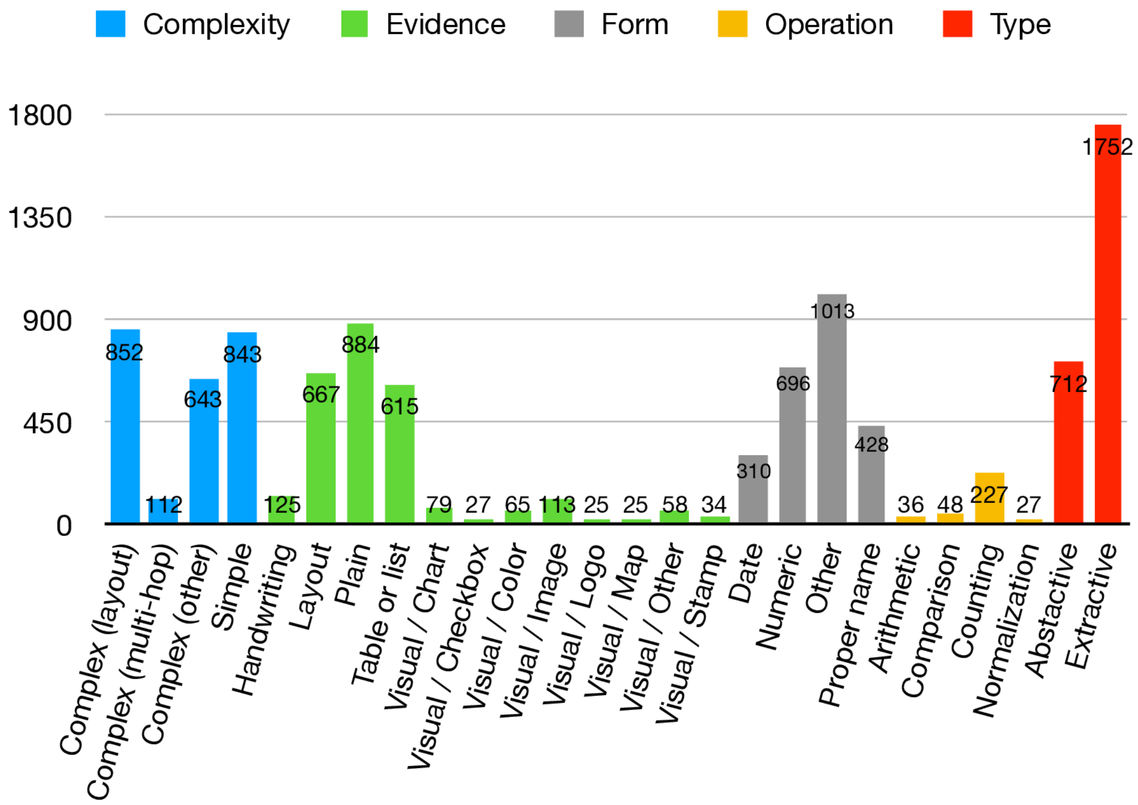
  industry_keywords.py



- Validation:
  - T-SNE plots over TF-IDF and ResNet feature representation
  - Relative diversity metric such as Simpson's coefficient

Question diversity

- Annotation environment design & instructions to obtain different question types
  - Extractive
  - Abstractive
  - List
  - Unanswerable
- Control mechanisms & deduplication
- Verification of question diversity on diagnostic test set (2.5K QA)



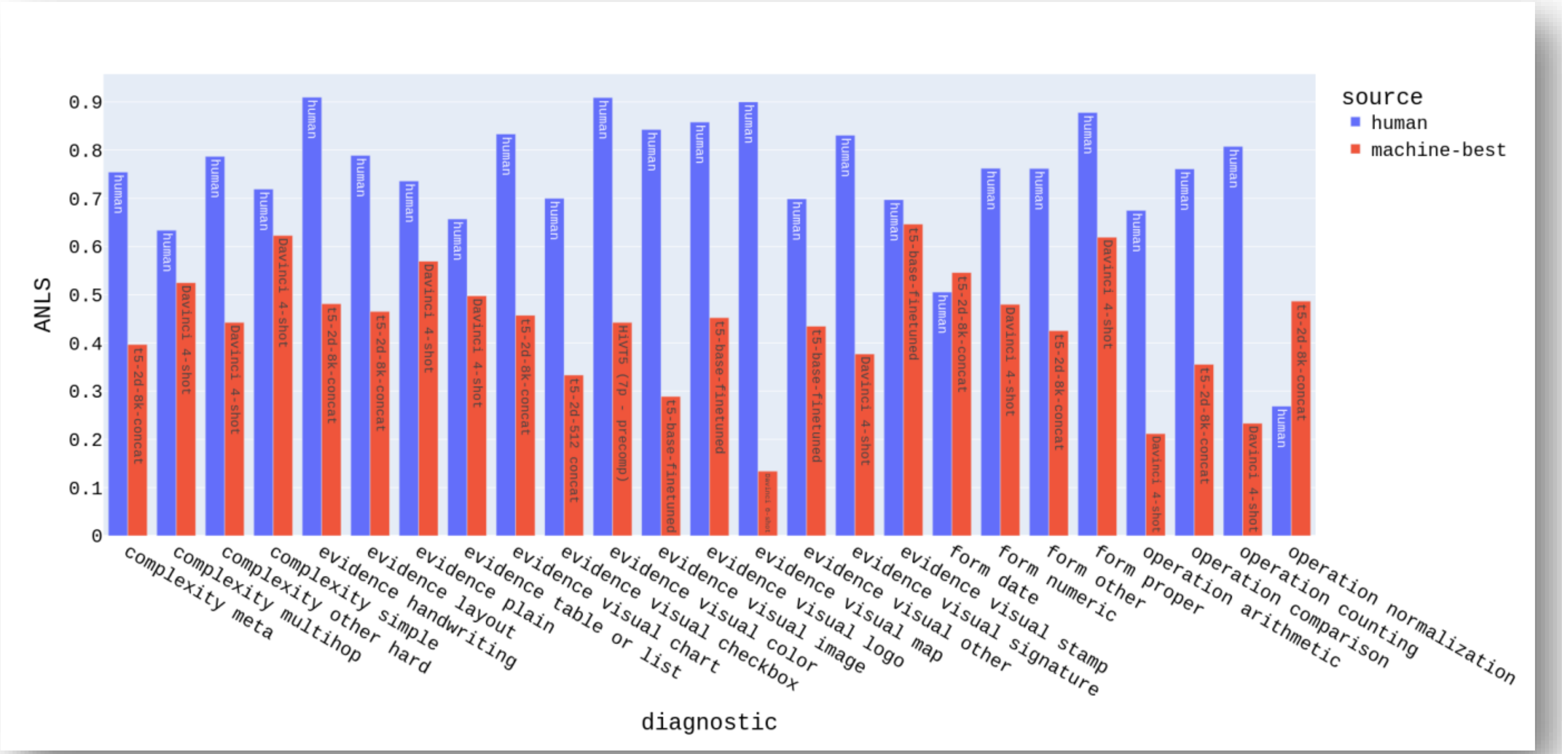
Baselines

Model	Init.	Params	Max Seq. Length	Test Setup	$ANLS_{all}\uparrow$	$ECE_{all}\downarrow$	$AURC_{all}\downarrow$	$ANLS_{do}$	$ANLS_{do}$ 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

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

SOTA ANLS < 50% !

Diagnostic categories performance



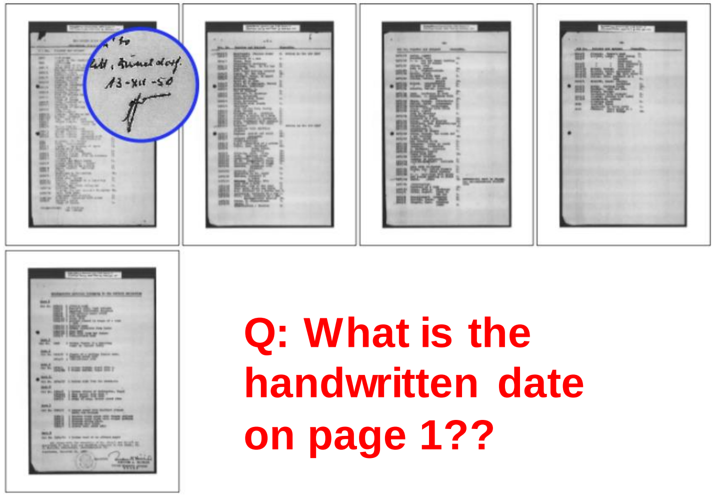
Diagnostic categories with

- visual evidence
- reasoning operations

Baselines lagging far behind **human baseline**

Qualitative examples

Handwritten evidence
Requires arithmetic
Multi-hop visual evidence



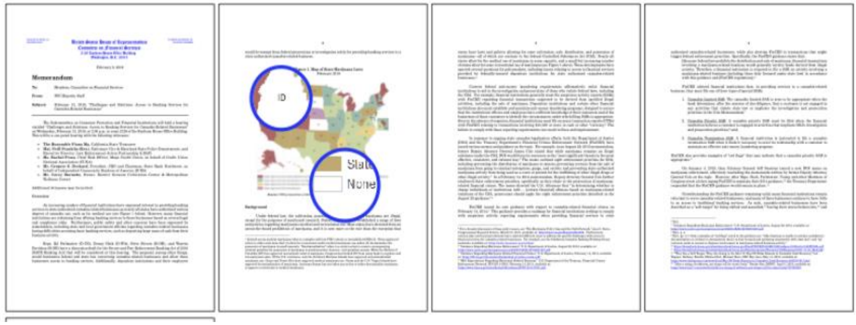
Q: What is the handwritten date on page 1??

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: 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?

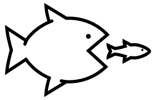


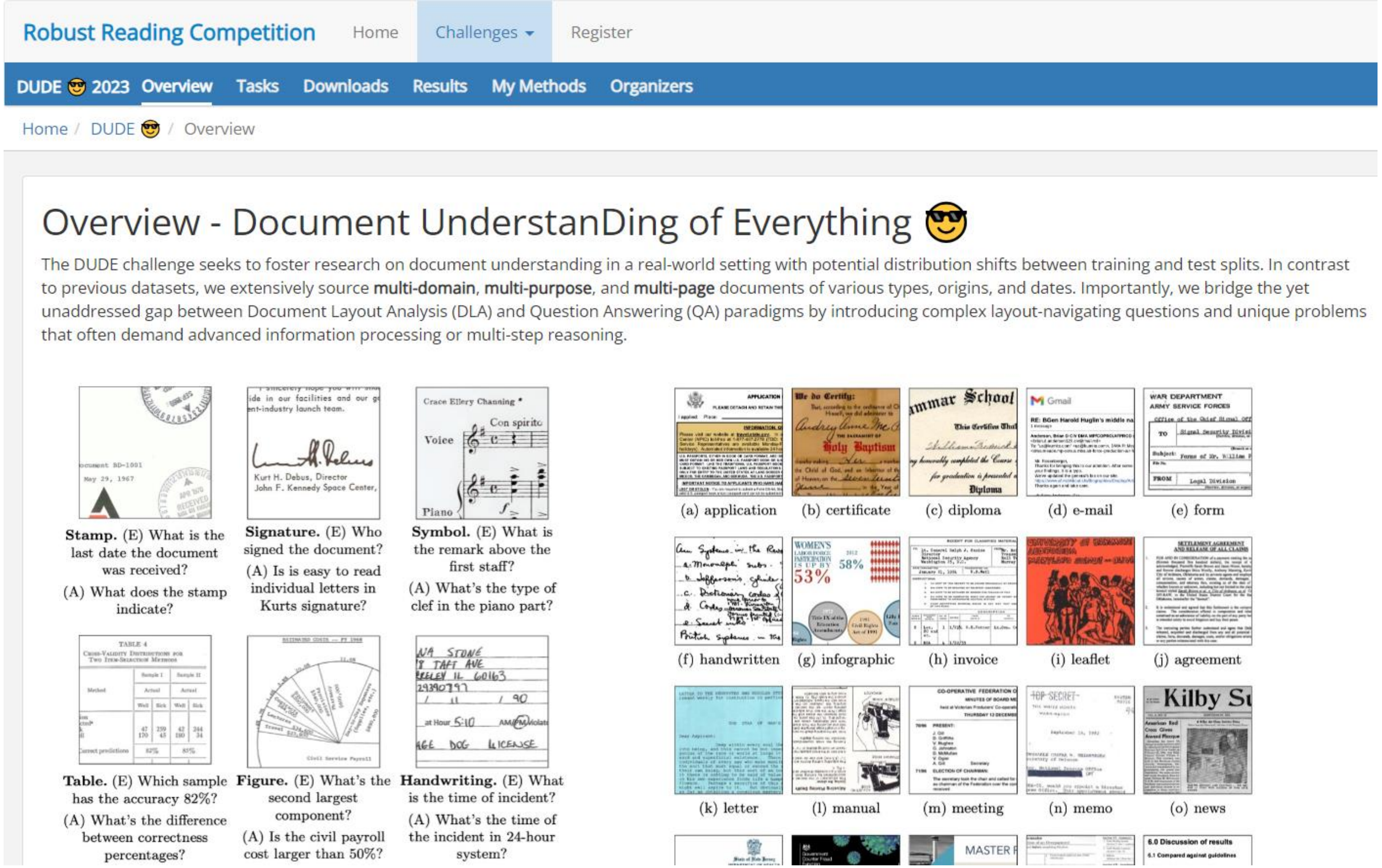
Source	Answer	ANLS	Conf.
Ground truth	*ID | SD | KS*		
Human	*ID | SD | KS*	1.0	—
T5	*WA ME MT ND MN OR VT ID NH SD WI NY MA MI*	0.0	0.28
ChatGPT	*[Not-answerable]*	0.0	—
GPT3	*American Samoa*	0.0	—
T5-2D	*i*	0.0	0.03
HiVT5	-	0.0	0.02

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 → force 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

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

- Given:
  - Natural language question (on content, aspect, form, visual/layout)
  - Input document
  - A set of reference answers

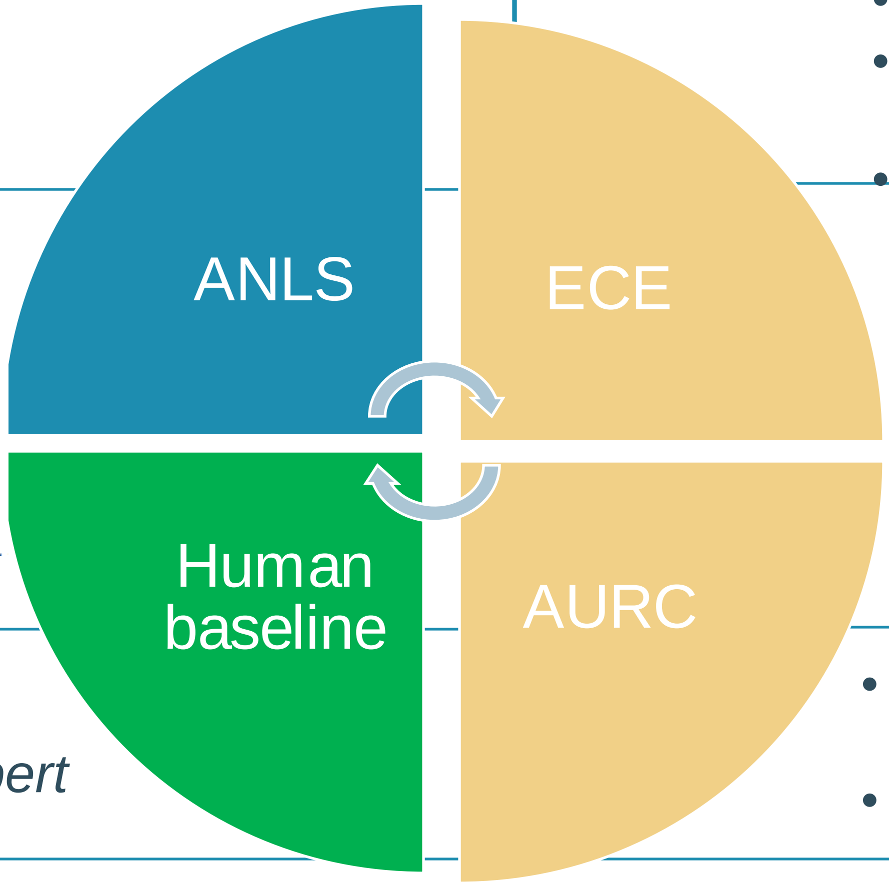
- Provide:
  - **Natural language answer**
  - Answer Confidence (float between 0 and 1)
  - Abstention flag (1 for abstaining)

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

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

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

- *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

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
- The multi-page aspect is not sufficiently addressed
  - Inefficiency for **long document processing**
- 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-4V), better solutions are coming, yet due to its designed complexity, DUDE might remain “the benchmark to beat” for a long time

DUDE: **What's Next?** ⏭

1. Reflections on DUDE
2. Curated research question
   A. Frame of reference
   B. Starting points
   C. Target aspects

Reflections on DUDE

- Benchmark with complexity in design sufficient to counter text-centric LLM approaches
- Launching the call to treat the layout modality as a first-class citizen
- Accentuating the field of document understanding as both separate from NLP and CV
  - Bringing its own set of problems and tasks
  - Requiring solutions beyond what is generated in the predecessor fields

➡ What follow-up **research questions** can we curate from DUDE?

Setting expectations on DUDE

- Finding good answers to covered research questions will be transformative to the technology as we know it
  - Format: RQ, my two cents, literature -> trigger discussion ^^

- RQ:

  How can we obtain the 'ultimate' document understanding dataset?

  A. What is the frame of reference/goal?
  B. What are good starting points?
  C. What aspects should be targeted?

A. Frame of reference

The definition and purpose of a document

- What is a "document"? (Buckland 1997)
  - Any information/evidence serving as a record
  - Communicative intent can be interacted with later

- What is document understanding?

  a complex process that involves holistically processing the layout of a document, as well as the textual and visual elements within it. It also requires the ability to reason with the extracted information, involving multiple skills and concepts, to generate meaningful actions or insights. (*mine*)

- Is it about the intentionality of the document's author or the way a user interacts with it?
  - What questions can be asked by an observer? → *observer's paradox*
    - *situation in which the phenomenon being observed is unwittingly influenced by the presence of the observer/investigator.*

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

Observer’s paradox in data collection for DocVQA

LAC-Verschaeren
Liersesteenweg, 219
2220 HEIST-OP-DEN-BERG

Tel : +32 15 25 80 70
Fax : +32 15 24 24 86
info@verschaeren.audi.be
BTW : BE0407688921
RPR MECHELEN

Offerte

20/04/2017

Ons contacteren

Uw verkoopsraadgever : Brecht Debaere
Tel : +32 15 25 80 78
E-mail : bd@groep-lac-verschaeren.be

Uw Audi A3 Berline
Model 2017

Onafhankelijk van onze wil, kunnen de afbeeldingen enkele verschillen in vergelijking met het geconfigureerde voertuig vertonen.

Personalia

Mr. Jordy VAN LANDEGHEM
Tel : 0471/72.34.01 / Gsm : 0471/72.34.09
E-mail : jordy.vlan@gmail.com
SPRL/BVBA SMART AND EASY
Nieuwgoedlaan 51 Bus 51
9800 PETEGEM-AAN-DE-LEIE, BELGIUM
BTW BE0646697416
Business Customer / BTW 21%

Samenvatting van uw offerte

Codificatie:
20170420 2017
8VMADG E01 0E0EMI
2446282
Opties:
YEI EB4 PNU 1XW
UE4 IT2 0VH
Invoerdersopties:
T926

*Klantenvoordeel :
€ 7.979,99 BTWI

	ZBTW	BTWI
Waarde van het voertuig en toebehoren	€ 27.995,45	**€ 33.874,49**
Voordeel speciale aanbieding *	€ -3.382,63	**€ -4.092,98**
Catalogusprijs	€ 24.612,82	**€ 29.781,51**
Korting *	€ -3.212,40	**€ -3.887,00**
Totaal voertuig en toebehoren (BTW 21% : € 4.494,09)	€ 21.400,42	**€ 25.894,51**
Prijs overname (zie overname overeenkomst)		**€ 0,00**
Te betalen voorschot		**€ 0,00**
Saldo		**€ 25.894,51**

Offerte geldig tot 27/04/2017
- Voor consumenten zal elke verkoop slechts als voltrokken beschouwd worden na ondertekening van een bestelbon (KB 9/7/2000).
- Bij inschrijving van een wagen is een eenmalige belasting op inverkeerstelling (BIV) en jaarlijkse verkeersbelasting te betalen op basis van de kenmerken van de wagen. Vraag meer info aan uw dealer.

Opmerking :
Uw Web Code –ARQ5SK1L– Deze code laat u toe op elk moment de details van uw voertuig terug te vinden in de configurator op internet.
Proefrit :
Laat U overtuigen met een testrit. Met plezier zullen we dit voor U organiseren.

Accountant: Is the VAT rate calculated correctly?
Is the VAT number present?

Legal: What is the chassis identifier to forward?

Insurance: What is the true net value of this car for insuring the risk in an omnium coverage?

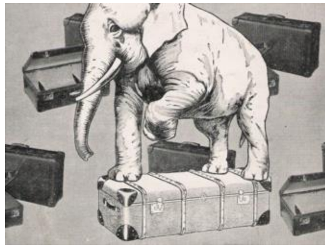
Customer: Why did he get a better quote on this car than me? Does this invoice include delivery?

Neighbor: why did this end up in my mailbox?
Why did he choose a black car? Is this a fast car?

…

Which questions do we expect a model to answer? **Curse/Opportunity?**

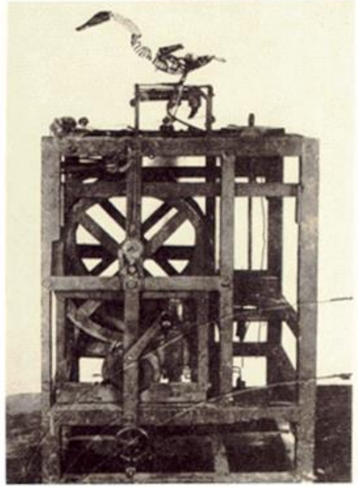
Measuring complexity and generalization



Turing test of document AI?

=> common-sense reasoning on documents from ***real-world interactions***

- Duck test *to test abductive reasoning, yielding a plausible conclusion without verification – inference to best explanation*
- Elephant test *refers to situations in which an idea or thing, "is hard to describe, but instantly recognizable when spotted"*
- Moravec's Paradox *states that it is easy to train computers to perform tasks that humans find difficult, such as mathematics and logic.*
- ...



Relevant caveats throughout dataset construction

- Beware creating Clever Hans effects
- Consider balancing language priors (adversarial Winogrande)

B. Starting points

I. ImageNet and MSCOCO

II. Recent document dataset efforts

¿Replicating? the ImageNet moment

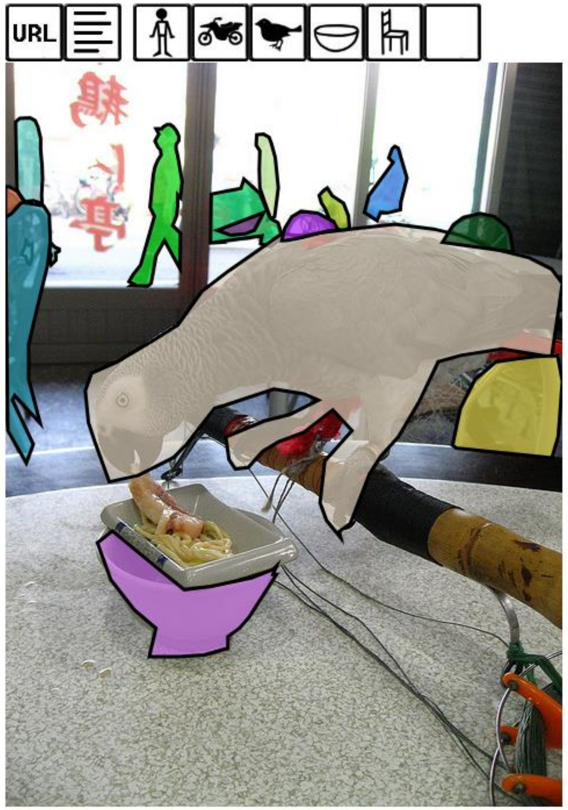
ImageNet

+ Large-scale
+ Established ground truth schema (WordNet nouns)
- Single classification task
- Label noise

MSCOCO

+ Large-scale
+ Common instances in context
+ Multi-task

Detection | DensePose | Keypoints | Stuff | Panoptic | Captions



Drawing inspiration closer to home

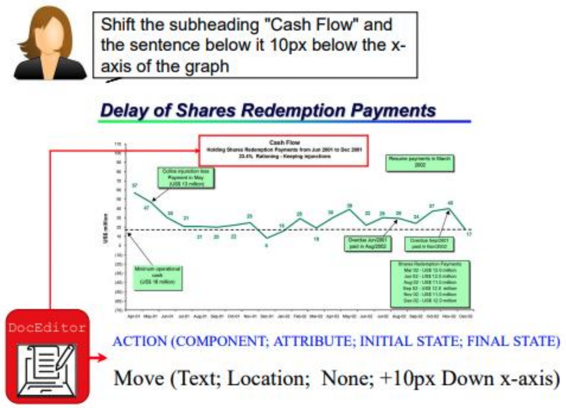
- To build the equivalent of MSCOCO in document understanding, DUDE offers a great starting point, under **some conditions and necessary extensions.**
- Innovations introduced by recent datasets on
  - Long, structured document VQA
  - Language-guided document editing

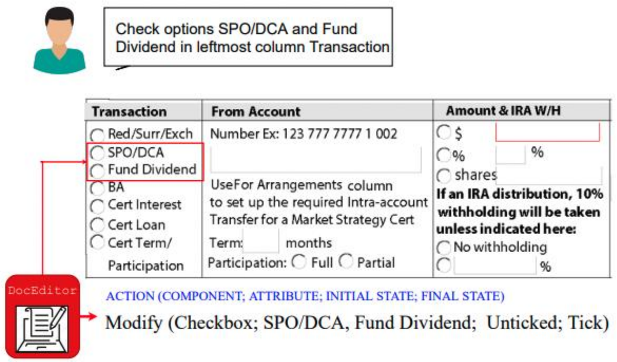
- <u>Ground truth collection</u>:
  - **DUDE**: post-hoc/MDLT, minimally constrained, human-generated questions
  - **PDFTriage**: pre-defined question types, human-generated questions
  - **DocEdit**: pre-defined taxonomy, human-generated questions

PDFTriage: question types

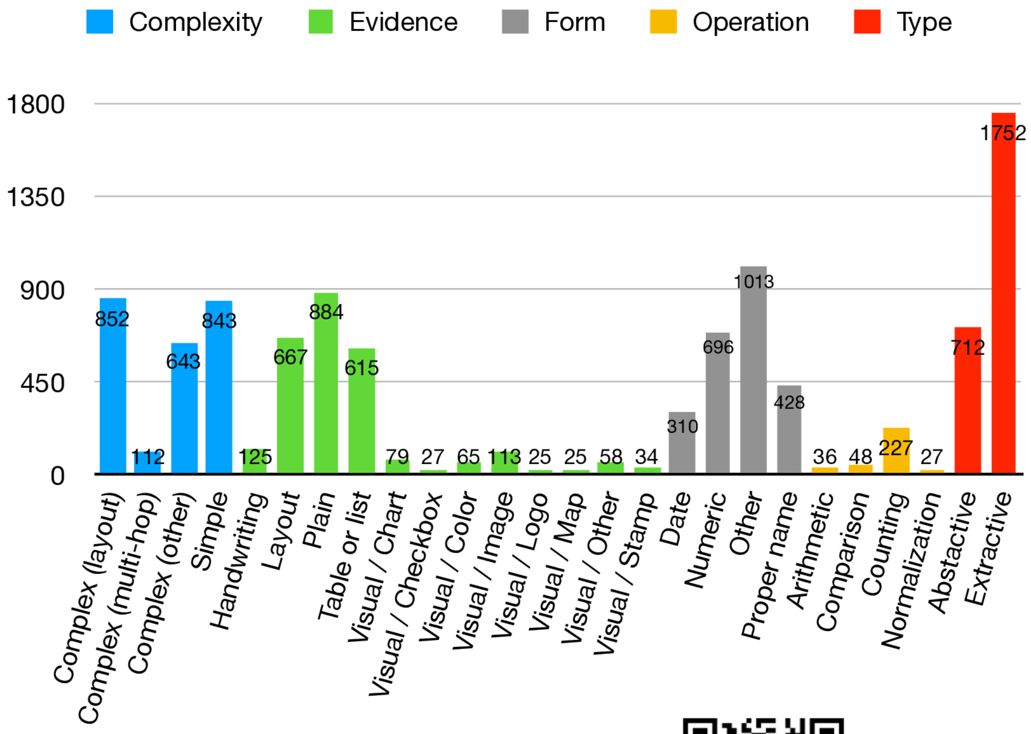
i. Figure Questions (6.5%)
ii. Text Questions (26.2%)
iii. Table Reasoning (7.4%)
iv. Structure Questions (3.7%)
v. *Summarization (16.4%)*
vi. Extraction (21.2%)
vii. Rewrite (5.2%)
viii. Outside Questions (8.6%)
ix. Cross-page Tasks (1.1%)
x. *Classification* (3.7%)

DocEdit: executable





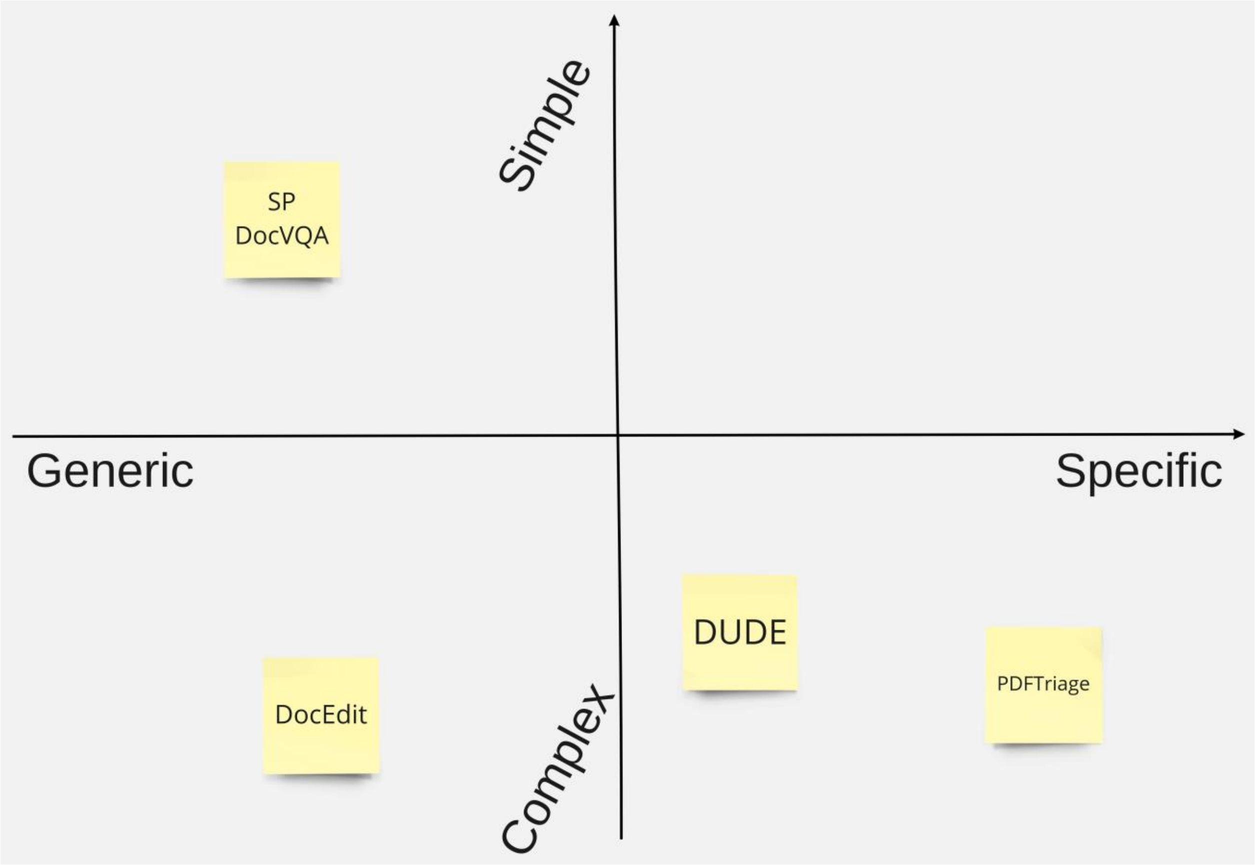
DUDE: diagnostic categories



gradio

https://4e7d0ef5d7d58cb631.gradio.live/

Question complexity and genericity



Interesting extensions:

1. Targeting complexity
2. Targeting specificity
3. Striking balance between question complexity and domain-specificity

Implicit dimensions:

- Accessibility
- Cost
- Scalability

C. Target aspects

I. Scale

II. Richness of supervision

I. Scaling up DUDE: question collection

- Cross-lingual questions to counter reliance on language priors
- Scaling up #Documents and #Questions per document in a balanced way
  - Ideally: scale #Q as a function document complexity
    - Open question: how to quantify document complexity?
  - Straightforward: split questions evenly over pages by chunked annotation
    - Constraining multi-hop and natural question complexity
- Untapped: **machine-generated questions**

1. **Emergent Analogical Reasoning in Large Language Models** [paper] 2023.12
2. **HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation** [paper] 2022.5

I. Scaling up DUDE: document collection

- Document collection approach is manual:
  - keyword-style search -> cluster-based diverse sampling from larger document collections
  - maximizing diversity in terms of all modalities with additional features: language, industry, document concepts, …
  - Open question: how to quantify document data diversity?
  - Need better PDF data exploration tooling
    - e.g., https://vawdataset.com/explore, https://atlas.nomic.ai/
  - Business documents are hard to obtain, backtrack to visually-situated language?
- Where to collect multi-lingual documents?

DocumentCloud

Common Crawl

1. **CCpdf: Building a High Quality Corpus for Visually Rich Documents** [paper] 2023.7
2. **Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding** [paper] 2022.10

I. Question generation

- Question-answer generation – interesting to complement for large-scale dataset
  - Teach current-best DUDE model to generate questions (A|D) -> Q
    - Risk: adding truly diverse and relevant questions?
    - Open question: how to reliably generate unanswerable questions?
      - How to evaluate a system's handling?
  - Gestalt: higher #Qs on heterogenous elements in document [1]
- Multi-step process:
  - Generate document captions alluding to concepts
  - Generate questions based on descriptions and skill templates

1. **Probabilistic homogeneity for document image segmentation [paper] 2022.5**
2. **Open-World Factually Consistent Question Generation** [paper] 2023.7
3. **GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction** [paper] 2023.10
4. **Question-generation-paper-list** [website]

C. Target aspects

I. Scale

II. Richness of supervision

II. Are we not expecting too much with poor stimulus?

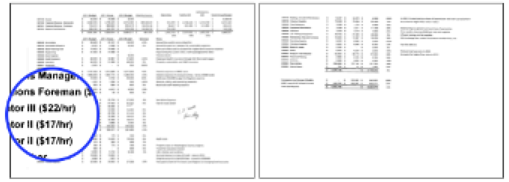
- Currently, expect models to learn how to answer complex questions involving (multiple) manipulation of document-instance and/or domain-specific concepts with a single set of reference answers

→ not providing i) **enough** or ii**) complex enough** or iii) **diverse enough** examples for learning

- Proposed remedy: compositional generalization from ground truth annotated with primitives (skill-concept)
  - → More explicit answer, grounding of answer (if possible) and evidence (attribution), and explanation of relations between skills and concepts
  - ?→ **discrimination** of known and **generalization** to new skill-concept combinations
- Proposed format: full-featured instruction tuning dataset

The difference **[arithmetic]** between the wages of operator 2 **(entity_1)** and 3 **(entity_2)** can be found from page 1, Table 1, column A, row Z **[locating the evidence].** This shows a table **(type of evidence)** over operators' net wages with Operator 1 making $22/hr**[attribute(entity_1)].** and Operator 2 making $17/hr**[attribute(entity_2)].** Thereby, the result is $5/hr **[arithmetic_difference(attribute(entity_1), attribute(entity_2)].**

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.



1. **Otter: A Multi-Modal Model with In-Context Instruction Tuning** [paper] 2023.5
2. **Dynosaur: A Dynamic Growth Paradigm for Instruction-Tuning Data Curation** [paper] 2023.5
3. **Self supervised learning and the poverty of the stimulus** [paper] 2023.09

II. From MDLT toward *skill-concept compositions*

Can each question-answer pair be decomposed into skill-concept compositions?

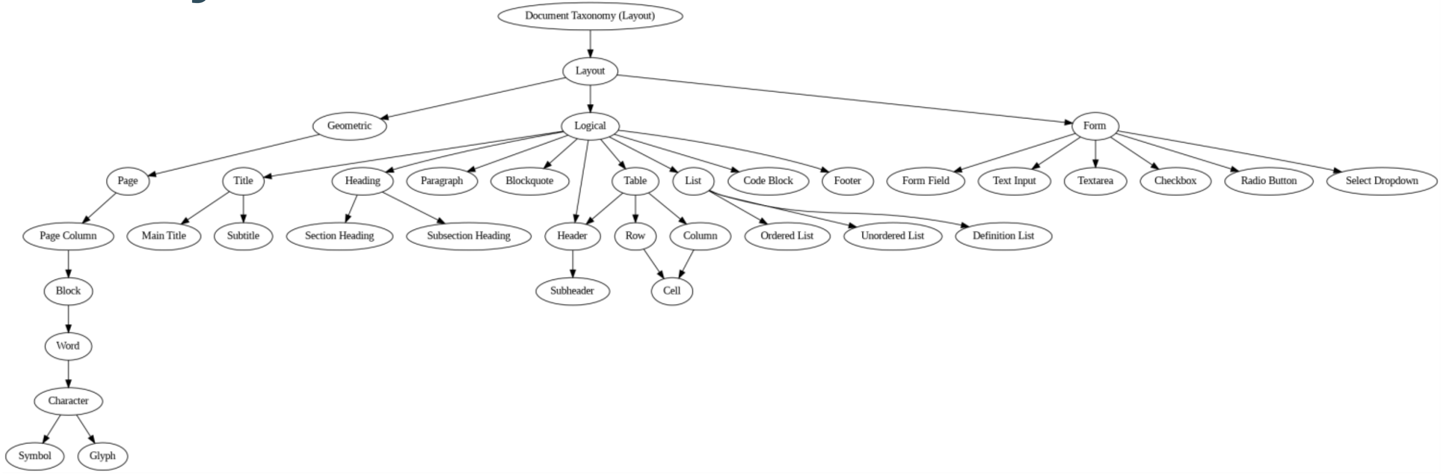
- **Concept**: a generic term to denote document visual objects (atomic [cell, barcode] and molecular [table, chart, form]), entities *(generic [document identifier, person, date] and domain-specific [invoice number, insured, payment date]*
- **Skill**: any manipulation *[existence, counting, relation, hasattribute, ...]* of a concept, or a combination of concepts (*evidence*) involved

document complexity := the expectation over all skill-concept compositions that can be requested for a document

1. **VQACL: A Novel Visual Question Answering Continual Learning Setting** [paper] 2023.5
2. **ViperGPT: Visual Inference via Python Execution for Reasoning** [paper] 2023.8
3. **VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use** [paper] 2023.8
4. **VerbNet;** [website]

II. Prototyping a skill-concepts taxonomy

- Establish generic document concepts
- How to integrate domain-specific concepts?
  - e.g., address block (layout) , invoice number (text)
- **layout**



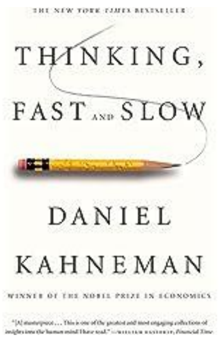
Examples

Table detection -> **Existence(table, document)**
Extract total amount paid from invoice -> **Locate(amount; custom)**

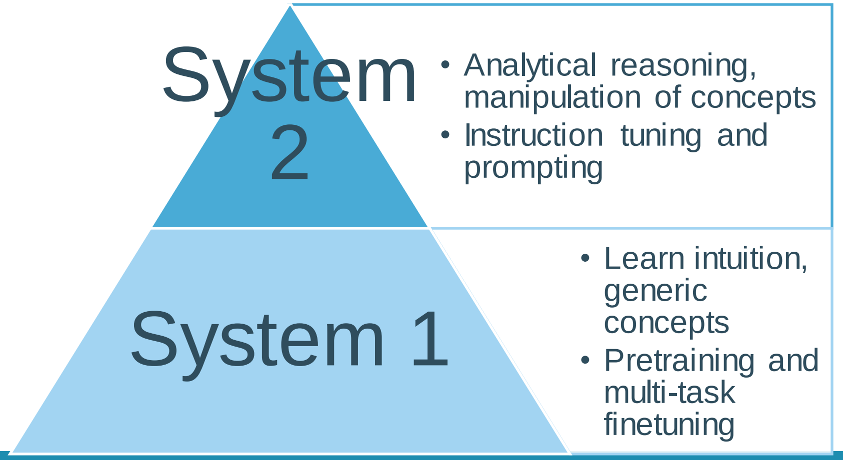
"How many of the contract's pages have signatures?"
Counting([Navigation(document).
Existence(signature, page)])

II. How to use this skills-concepts taxonomy?

“Do we really need thousands of examples of QA pairs to learn a specific skill-concept composition?”



- Build ground truth more applicable to advanced prompting techniques
  - Proven useful with semantic parsing and question decomposition
- Create novel question-answer templates from existing compositions
- Investigate neuro-symbolic architectures that allow for dynamic knowledge graphs



1. **Graph of Thoughts: Solving Elaborate Problems with Large Language Models** [paper]
2. **GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering** [paper] https://cs.stanford.edu/people/dorarad/gqa/
3. **Complex Question Decomposition for Semantic Parsing** [paper] 2019.3
4. **Is a Question Decomposition Unit All We Need?** 2022.5
5. **KOSMOS-2.5: A Multimodal Literate Model** paper
6. **Sample-efficient Learning of Novel Visual Concepts** [paper] 2023

Conclusion and value for Adobe

- Scaling with question generation and document-question diversity
- Designing the most valuable ground truth to learn a complete distribution over skills and document concepts

I hope to have provided some food for thought on making an informed answer.

- I believe that Adobe is the right party to build this *‘ultimate’ document understanding dataset*, targeting both the scale and depth of supervision; establishing Adobe's position as the **document intelligence pioneer**

COMPETITION

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

DATASET

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

Questions?

jordy-vl.github.io/

KU LEUVEN