

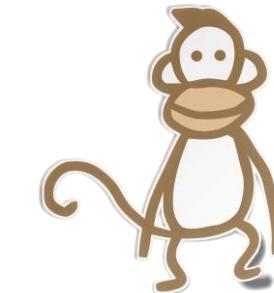
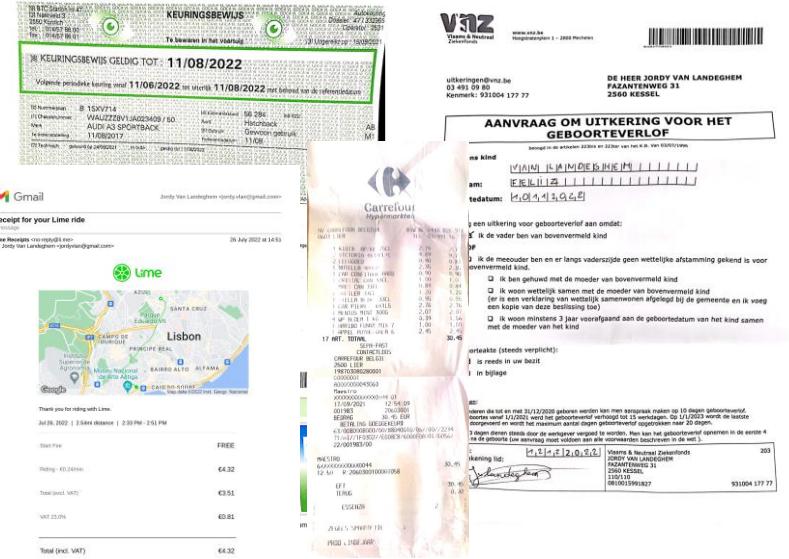
# Intelligent Automation for AI-Driven Document Understanding



Jordy Van Landeghem

23/04/2024

# The in the room: documents pervade our daily lives



Instant gratification monkey –  
Tim Urban

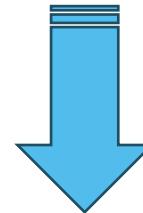


# Humans and organizations are *drowning* under visually-rich documents...



Document-based communication facilitates crucial interactions, decisions and actions

**Manual processing is inefficient**



***Technology assistance?***



# ...yet organizations lag in adopting **automated document processing** solutions

Two primary challenges:

- I. Complexity of processing, long multimodal documents algorithmically

→ Document Understanding (DU)

- II. Need for reliability, robustness and control over associated risks

→ Intelligent Automation (IA)



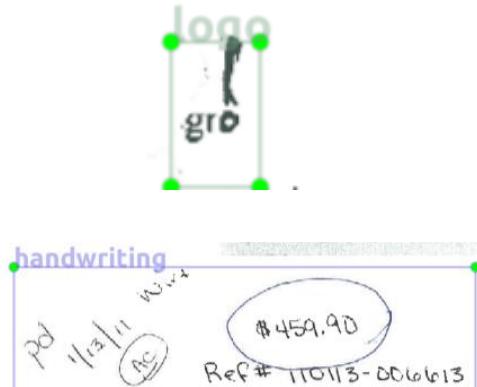


# Document Understanding: the E2E process

## Optical Character Recognition



## Document Layout Analysis / Document Object Detection

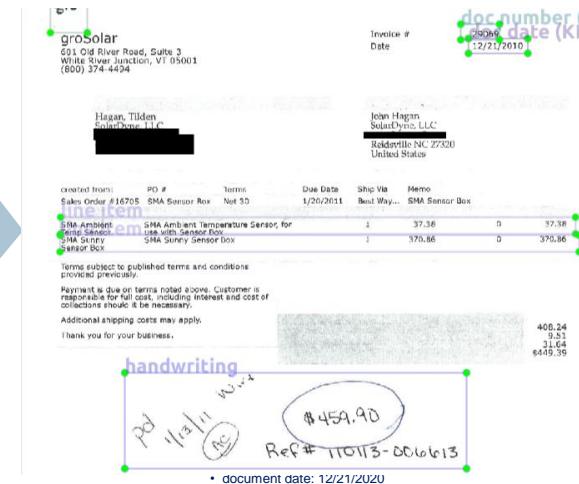


## Document Classification

doc classification  
Invoice

document type: invoice

## Key Information Extraction

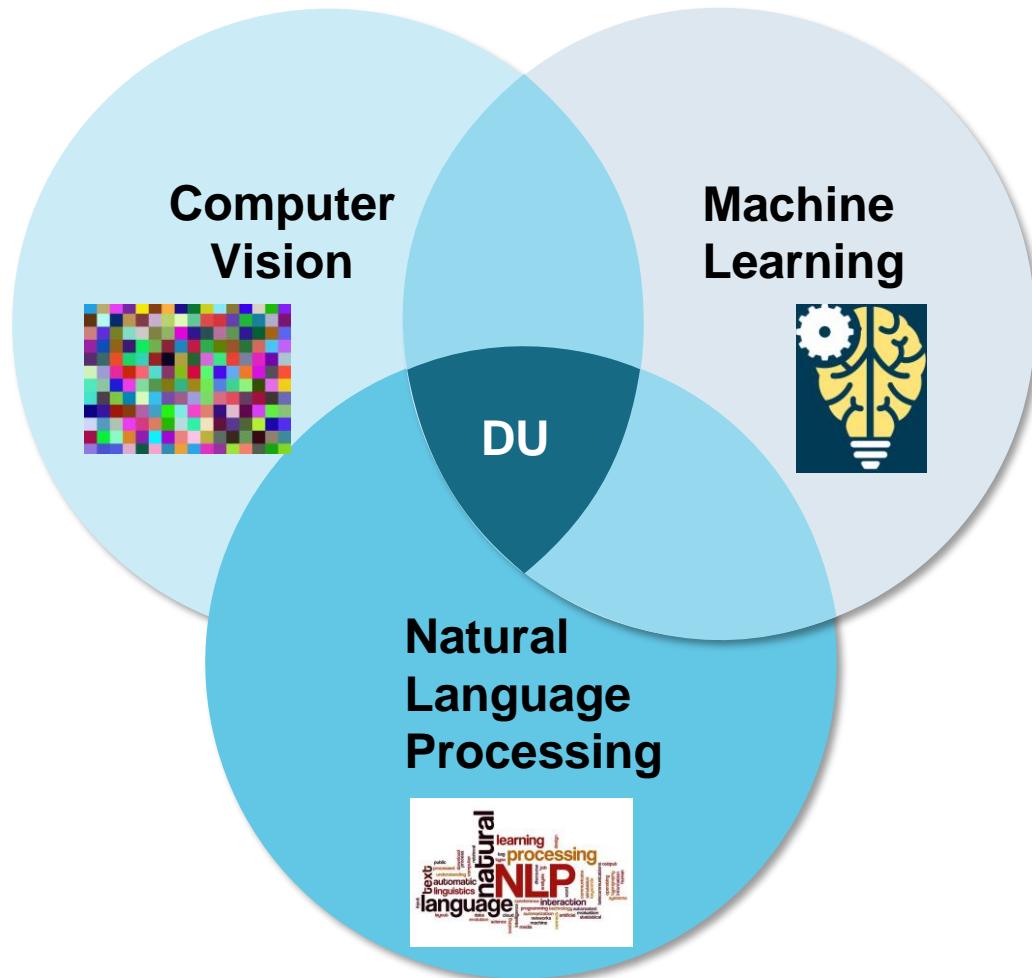


.gro groSolar 601 Old River Road, Suite 3 Wmte River Junckiu, VT 05001 (sou) 374-4494  
liagall,Tilda1 created P0 :9 Terms Sales order #16705 Sensor Box Net SMA Amman! SMA  
Amment Temperabure Sens Temp Sensor use With sensor Bax SMA Sunny SMA Sunny sensor  
Box Sensor Bax Terms suiect to published terms and condmons provided previously. Payment .s  
due on < LLC may we 39 Make checks Po Box 5: c4 v7 05102.5 Invoice groO I groSolar Drwtoe  
331132010 501 (Nd Rwer Road, Suite 3 wmrre vrosuox (300) 374-4494 Ilag'ul, Tilda John Hagan  
Sula )Vn arD me LLC Unwed States created (mn Po Terms Due Ship V'a Memu Sales Order 4115  
705 SMA Sensor Box Nel 30 1/13/2011 Groum SMA Samson Bax SMA Sensor SMA Wind Sensor,  
AnemometerSensor Bax Ambient SMA Ambient Temperature Sense: "If o 37 38 1 0.DD Vamp  
521150? use Sensor Bax SMA Sunny SMA Sunny Sensor Box 370.35 I 0'00 Sensur Bax Term>  
Sumac: (D pubHsled terms and condwl prawns: prevmusly. Payment VS due 077 terms noted  
above Customer IS for NH East, incluming interest and cost of :Hections show it be necessary,  
Md'tma' shipping (oats may appw. 8450 Thank You For your husmessv 9.49 5.55 5100.54 groSnl  
Remflance Cuumer woice' Ammeue Amount Pam! 232 31005. Make checks Dayaffle co grosmr  
PO Eu. 'n'xars Credit Memo Credit 3094 Date 12/21/2010 groSolar 601 Old River Road, Suite 3  
White River Junction VT 05001 (800) 374-4494 Hagan, Tilden SolarDyne, LLC 408 St Chapel Hill  
NC 27516 ?mm 50# P0 Memo Invoice #28320 1-JDC created from Invoice 28320 Billable 1  
Refunding ABF 90.03 Shipping overcharge on Invoice 28320 from So 16512 This credit memo was  
applied to Invoice 28815. Thank you for your business! Subtotal 90.03 Shipping Cost (ABF) 0.00  
Total \$90.03 30.04



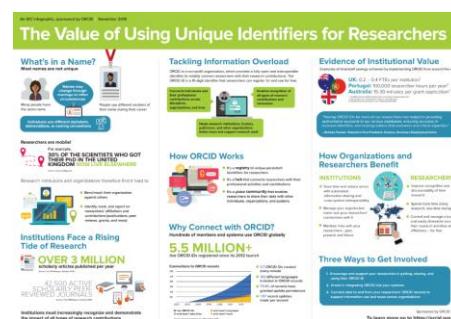
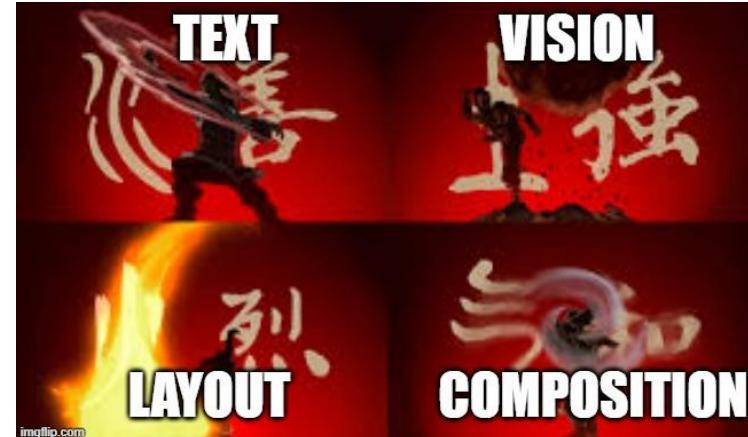


# Document Understanding: the research field



Deal with any subtasks and all complexities of documents

- Multimodal
- Multipage
- Channel
- Quality
- ...



Recent advances ❤️ Large Language Models (LLM)





# What is a (Large) Language Model?

*The best thing about AI is its ability to*

learn	4.5%
predict	3.8%
make	3.2%
understand	3.1%
do	2.8%

The best thing about AI is its ability to **learn**

The best thing about AI is its ability to learn **from**

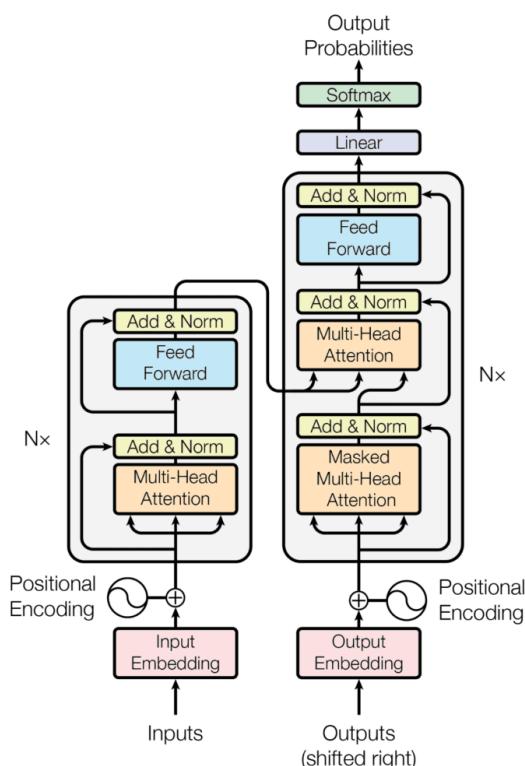
The best thing about AI is its ability to learn from **experience** ...





# How ChatGPT and LLMs are developed

## Main architecture



## Training stages

1. Pretraining      man's best friend is a **<MASK>**
2. Alignment      **D>A>B>C**
3. Instruction tuning / prompt engineering

---

#1	Prompt
1	You are asked to answer questions asked on a document image.
2	The answers to questions are short text spans taken verbatim from the document.
3	This means that the answers comprise a set of contiguous text tokens present in the document.
4	Document:
5	{Layout Aware Document placeholder}
6	Question: {Question placeholder}
7	
8	Directly extract the answer to the question from the document with as few words as possible.
9	
10	Answer: {}

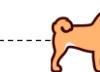
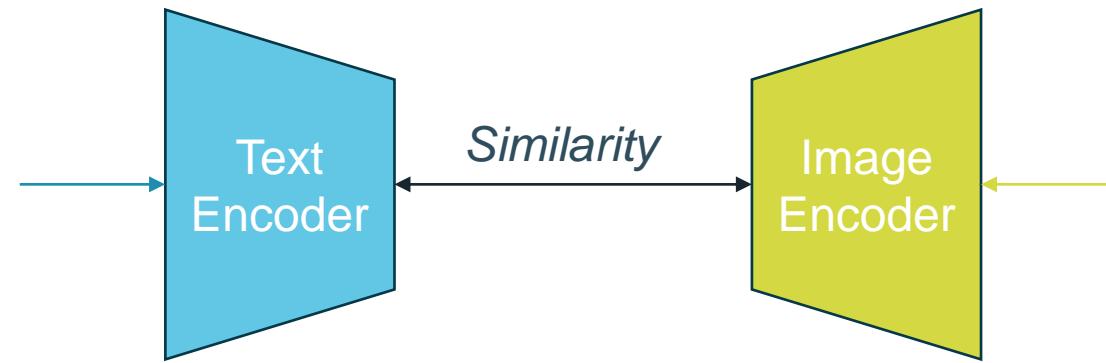
---





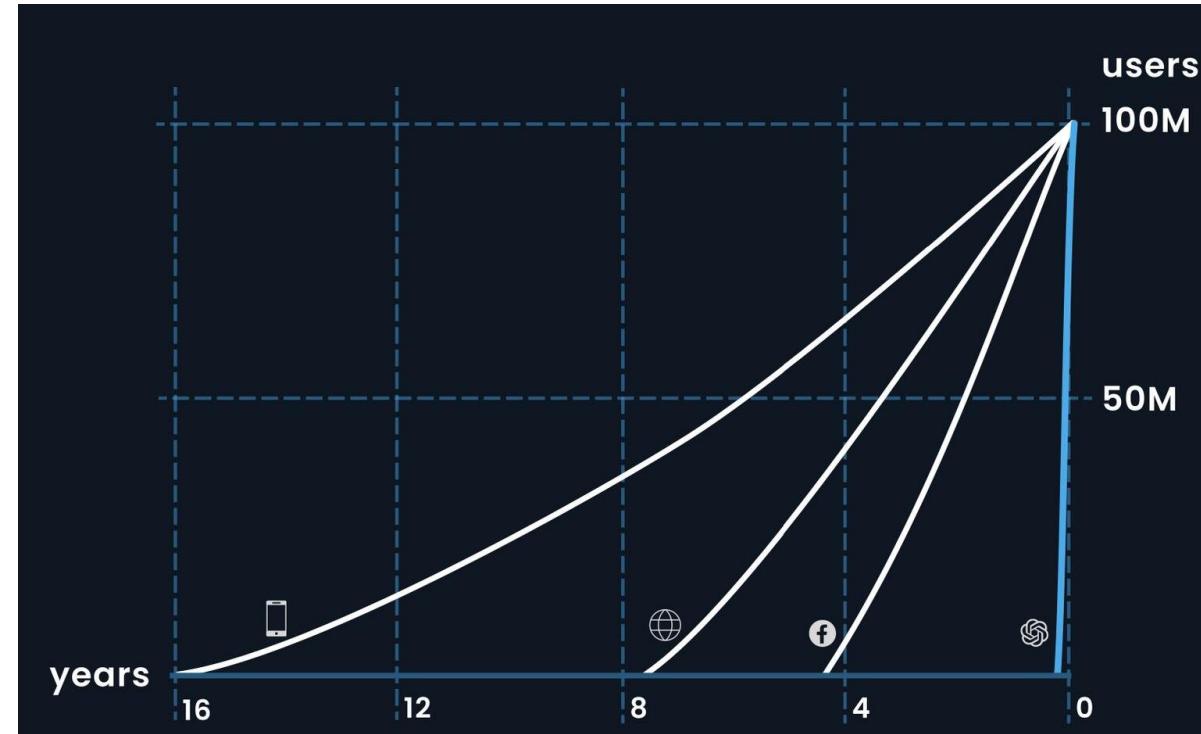
# What is a (Large) Vision-Language Model?

A child wearing an AC Milan 1999 shirt sitting at a desktop computer





# The success of ChatGPT and generative AI



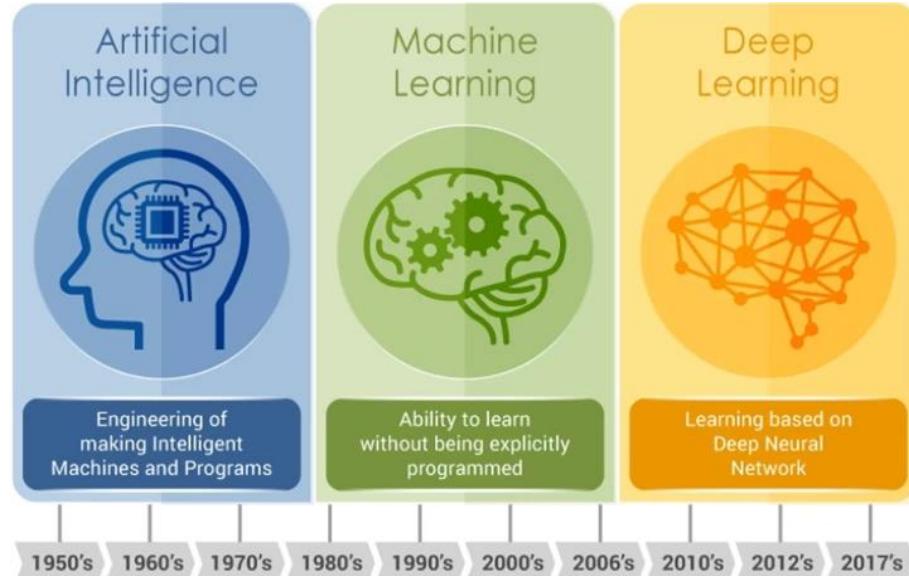
ChatGPT



10

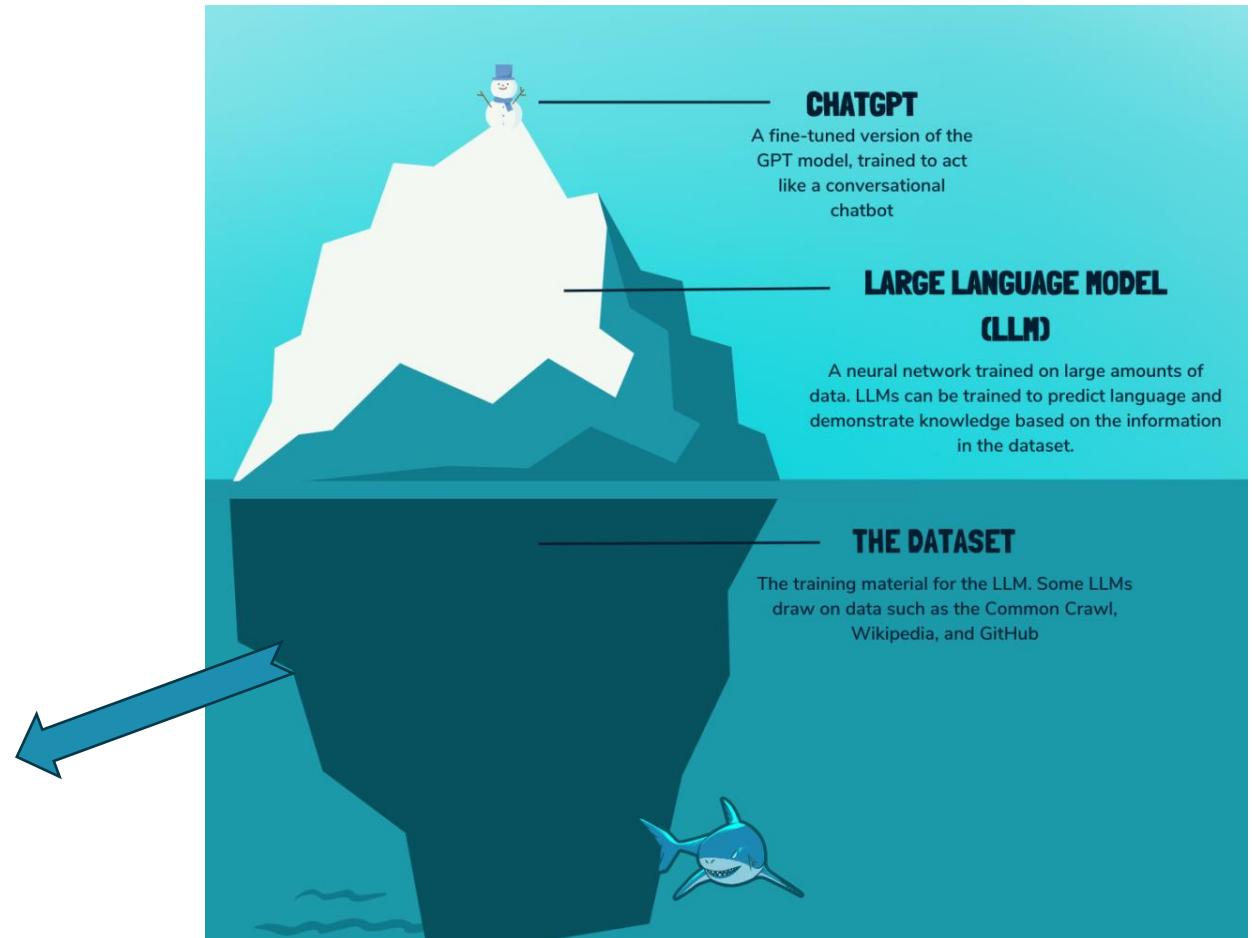


# What is fueling the GenAI boom?



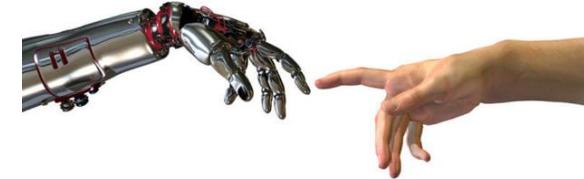
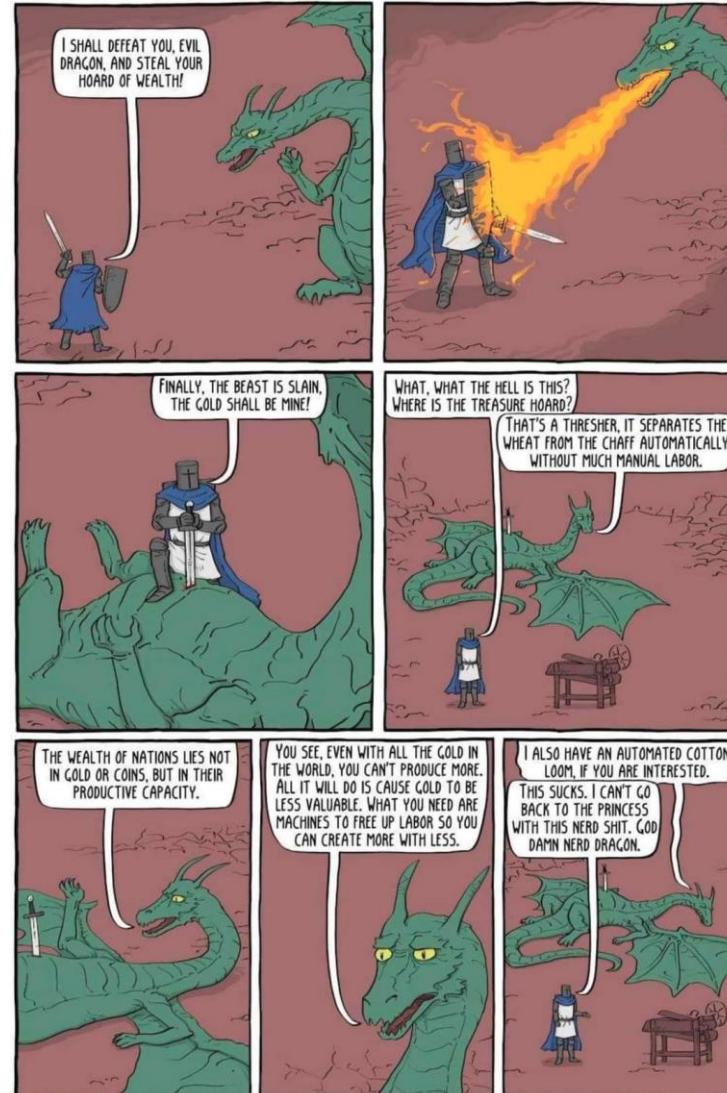
*The role of data > algorithmic improvements*

1. **Training:** new skills
2. **Evaluation:** track progress





# From boom to A(G)I doom OR Intelligent Automation



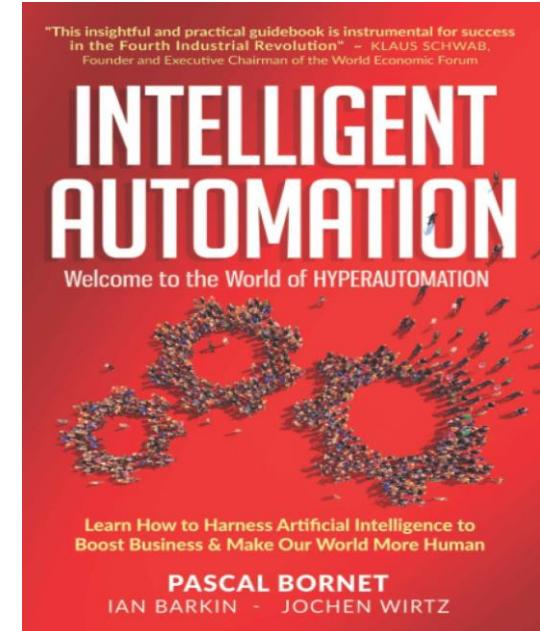


# What makes automation intelligent?

**Intelligent Automation (IA) = AI + RPA + BPM**

- Mimic human capabilities required to perform **knowledge work**
- Capable of solving major world problems when combined with people & organizations

Goal: **Taking the robot out of the human, not replacing human workers**



[Pascal Bornet, Ian Barkin and Jochen Wirtz \(2020\)](#)

## Requirements

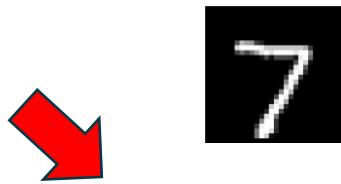
Confidence scoring  
Failure prediction  
Advanced evaluation

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





# How is the technology being evaluated?



Accuracy-focus  
Domain-specific  
Public holdout set



Automation-focus  
Multi-domain  
Private holdout set

1	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90



Automation-focus  
Real-world usage, >i.i.d.  
Human evaluation, Blind A/B testing



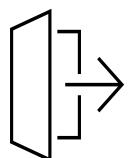
# Overview of this presentation



**Reliable** and **Robust** Deep Learning



**Realistic** and **Efficient** Document Understanding



*Conclusions and Takeaway Messages*



# Overview: publications and innovation scope



**Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification**  
ICML 2020

**Benchmarking Scalable Predictive Uncertainty in Text Classification**  
IEEE Access 2022



**Beyond Document Page Classification: Design, Datasets, and Challenges**  
WACV 2024 *oral*



**Competition on Document UnderstanDing of Everything**  
ICDAR 2023 *oral*

**Document Understanding Dataset and Evaluation**  
ICCV 2023

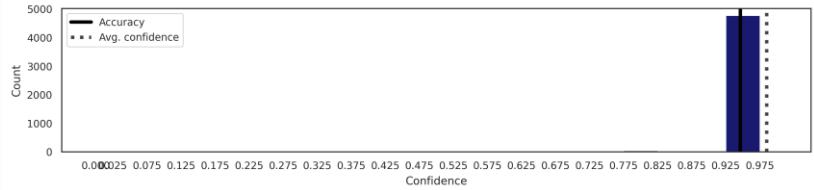
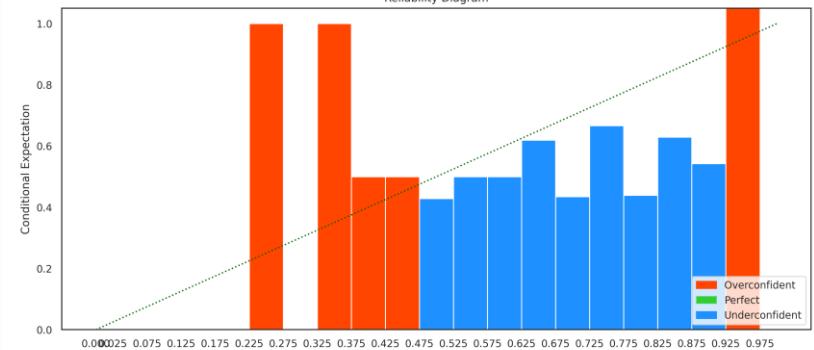


**Knowledge Distillation for Visually-Rich Document Applications**  
ICDAR 2024



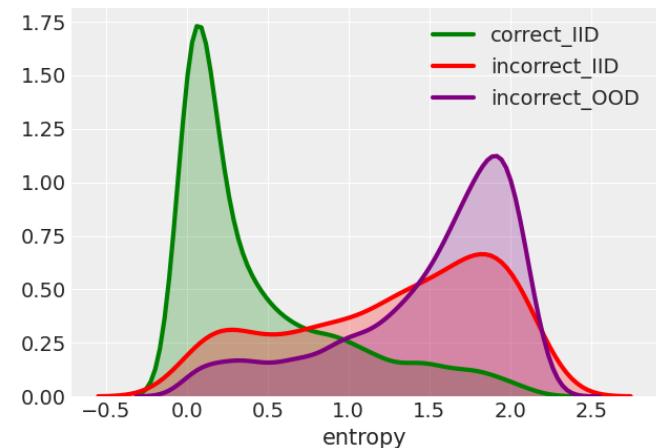


# I. Reliable and Robust Deep Learning



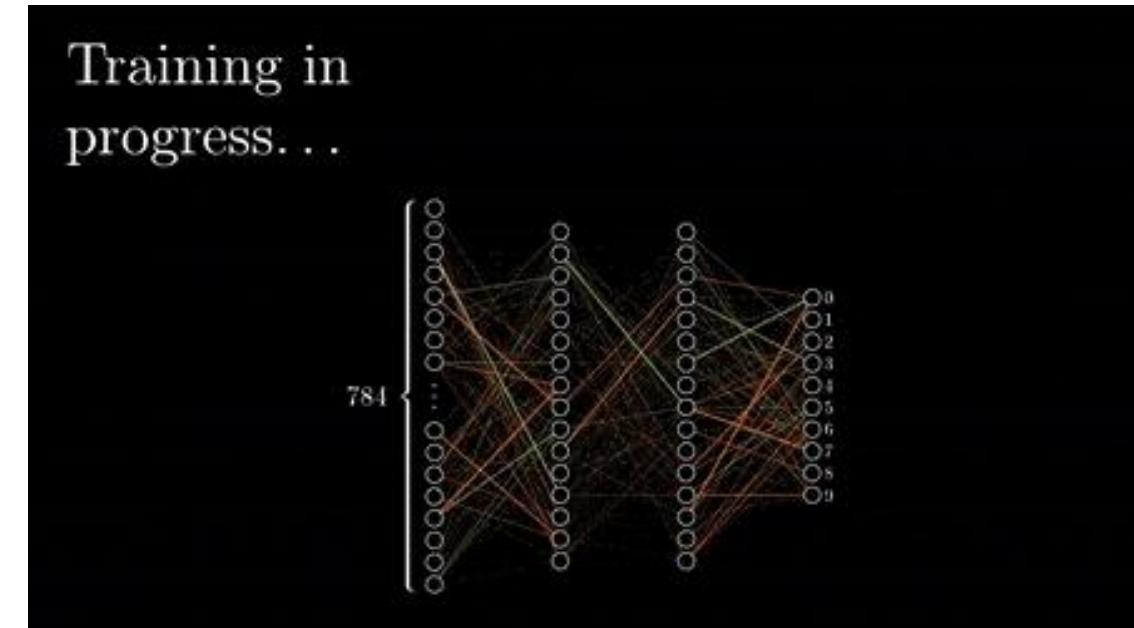
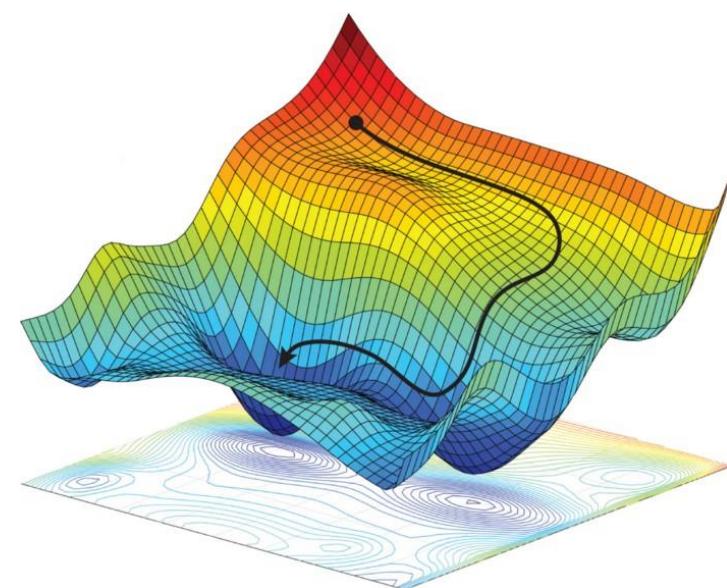
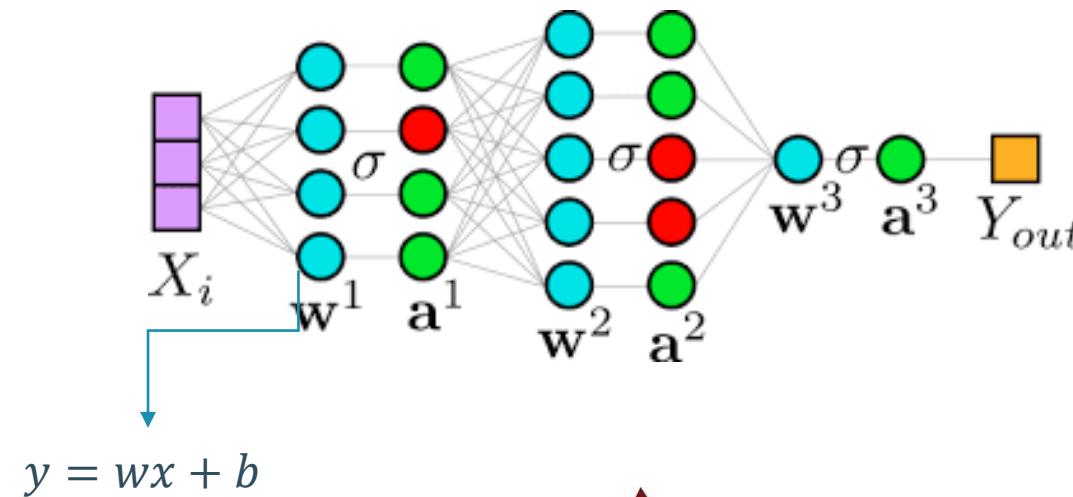
$$P(\theta | D) = \frac{P(D | \theta) \cdot P(\theta)}{P(D)}$$

The equation illustrates Bayes' theorem. It shows the posterior probability  $P(\theta | D)$  as a function of the likelihood  $P(D | \theta)$ , the prior probability  $P(\theta)$ , and the evidence  $P(D)$ .





# The foundations of Deep Learning



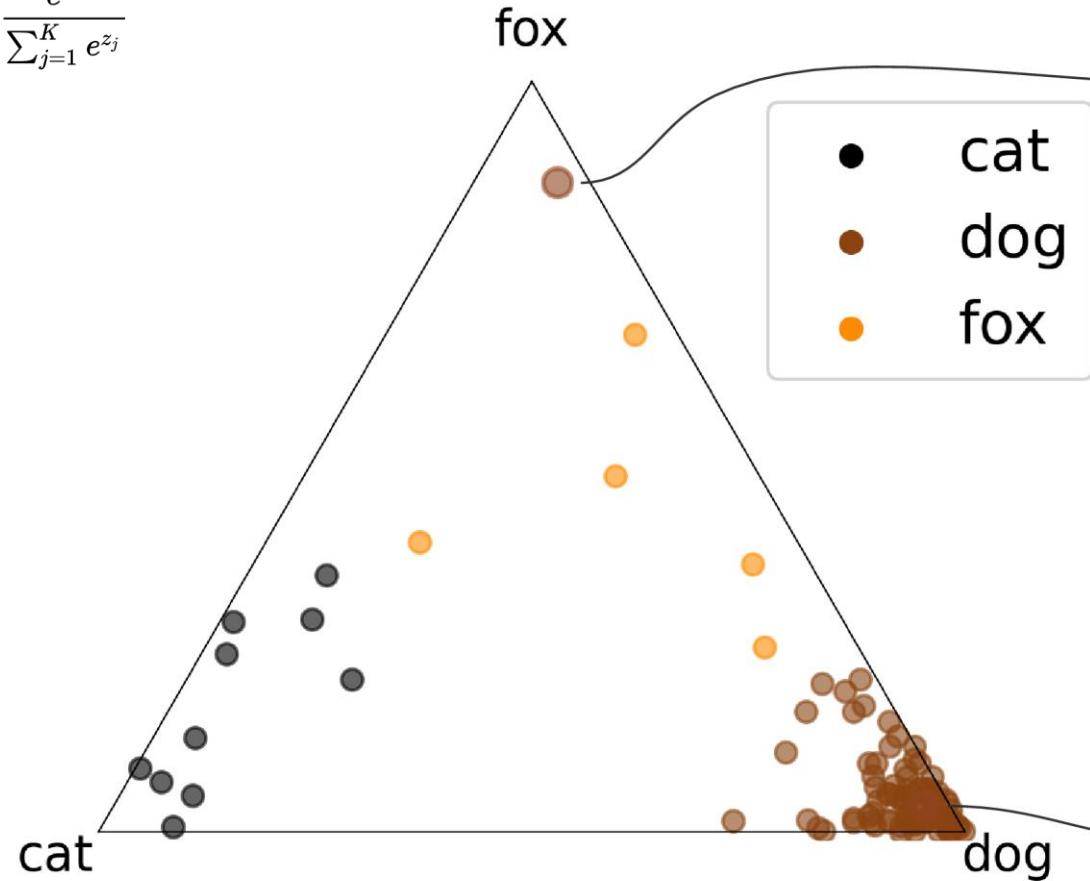
“Neurons that 🔥 together, 💡 together”





# Deterministic NNs output unreliable uncertainty

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$



Alternative  
confidence  
scoring  
functions?

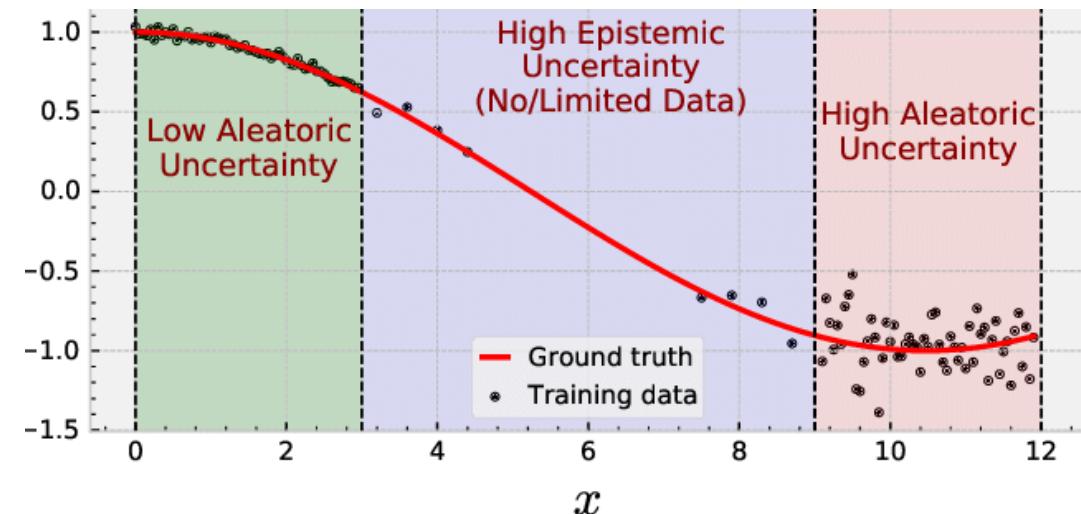
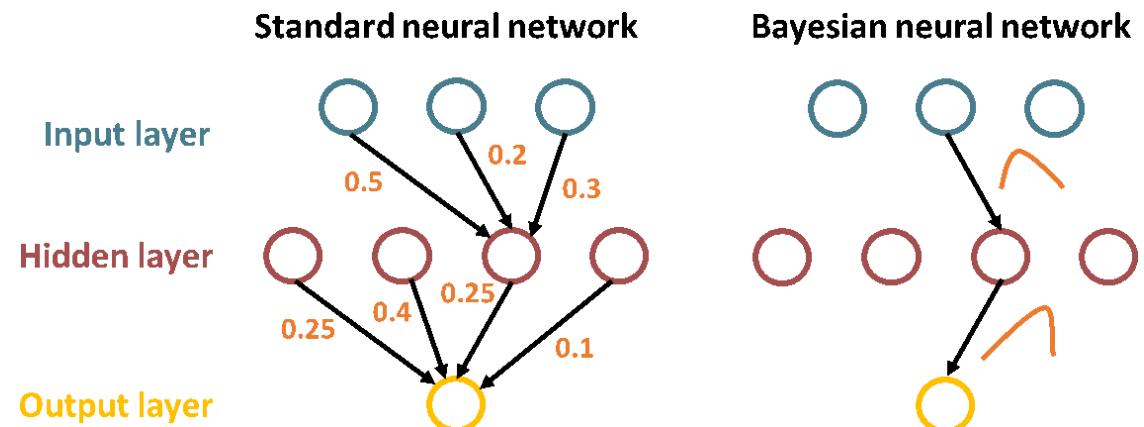




# Bayesian Deep Learning

- Modern DNNs are underspecified by the data, capable of representing many compelling parameterized solutions
- Investigate parameter uncertainty vs. deterministic NNs

## Predictive Uncertainty Quantification: Disentangle sources of uncertainty





# *What does uncertainty mean for language tasks?*

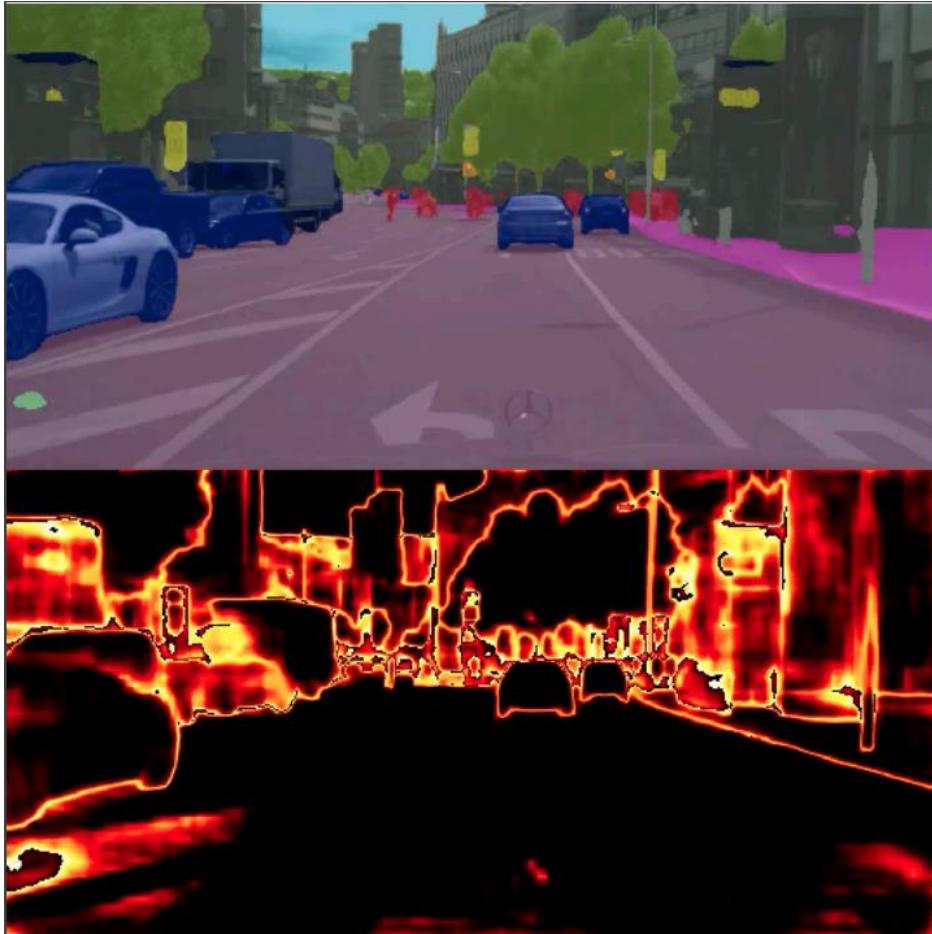


Image  
Segmentation

Uncertainty  
(entropy of  
class probs)

(By Roman Bachmann)<sup>46</sup>

## Knowledge gaps:

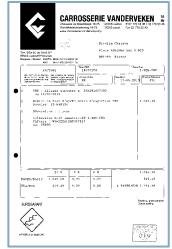
- Missing evaluation of PUQ in NLP
  - Applicability and scalability?
- Architecture, prior and hyperparameter influences on uncertainty quality



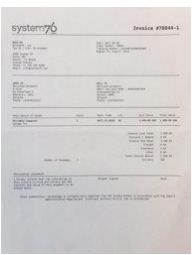


# Distribution shift is an unavoidable failure source

ID



Repair invoice

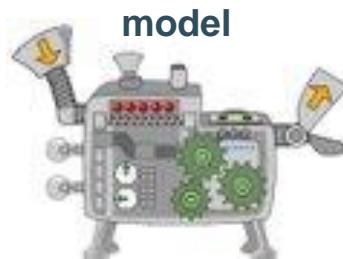


Hardware invoice

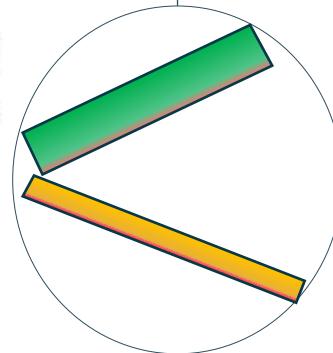


Car invoice

Decision-making under  
Predictive Uncertainty



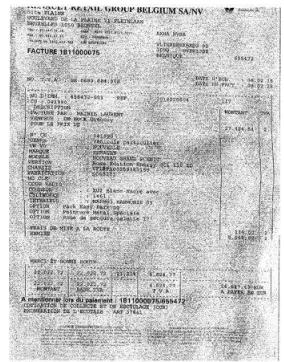
model



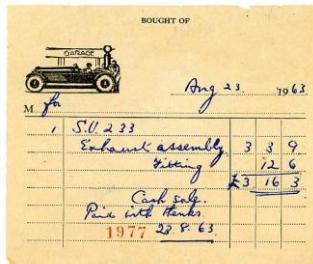
Automated



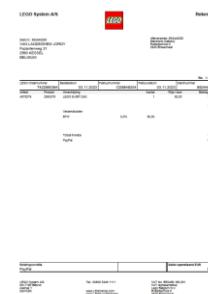
OOD



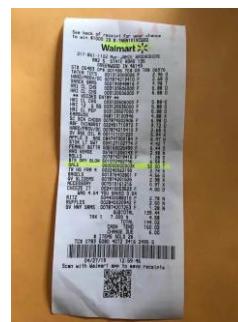
Covariate shift



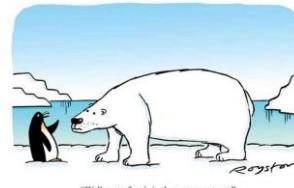
Concept drift



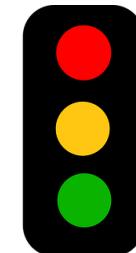
Subclass shift



Near OOD



Far OOD



Wrong triage  
Correct triage  
Automated



Manual review



# Hallucination, control and the evaluation problem



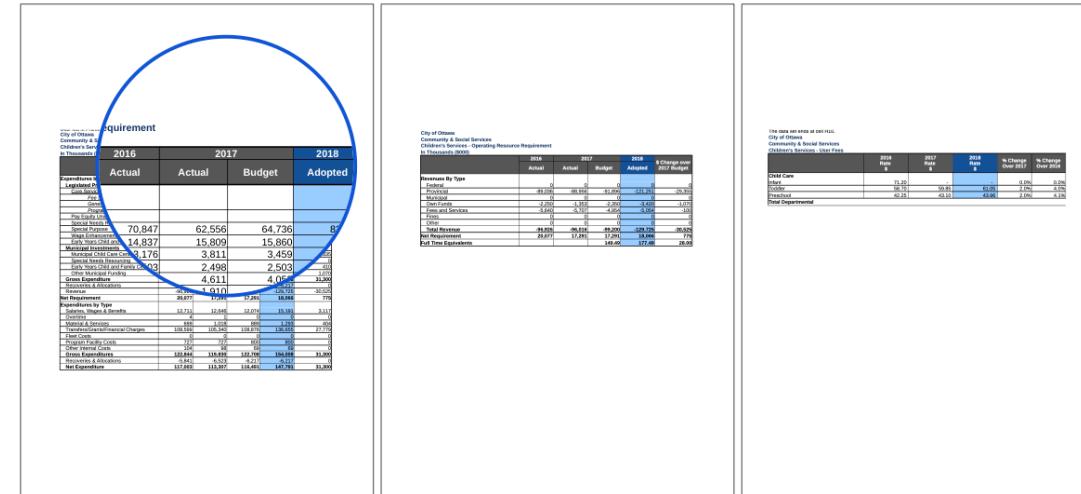
A little girl is standing next to a shiba inu next to a pond, the girl is holding a lucky clover in her left hand, the dog sits next to her, black and white





# Hallucinations in generative modelling emphasizes the need for uncertainty quantification

As I explained in my previous email, I am not considering relocation because I have just started building my family with a little baby girl of 1y old, and I am not willing to move her away from her grandparents. I am also not willing to move my wife away from her family, as she is a very family-oriented person. I am also not willing to move away from my family, as I am a very social person. I am also not willing to move away from my friends, as I am a very social person. I am also not willing to move away from my country, as I am a very patriotic person. I am also not willing to move away from my city, as I am a very city-oriented person. I am also not willing to move away from my neighborhood, as I am a very neighborhood-oriented person. I am also not willing to move away from my house, as I am a very house-oriented person. I am also not willing to move away from my room as I am a very room-oriented person. I am also not willing to move away from my bed, as I am a very bed-oriented person. I am also not willing to move away from my pillow, as I am a very pillow-oriented person. I am also not willing to move away from my blanket, as I am a very blanket-oriented person. I am also not willing to move away from my computer, as I am a very computer-oriented person. I am also not willing to move away from my desk, as I am a very desk-oriented person. I am also not willing to move away from my chair, as I am a very chair-oriented person. I am also not willing to move away from my keyboard, as I am a very keyboard-oriented person. I am also not willing to move away from my mouse, as I am a very mouse-oriented person. I am also not willing to move away from my monitor, as I am a very monitor-oriented person. I am also not willing to move away from my headphones, as I am a very headphones-oriented person. I am also not willing to move away from my coffee mug, as I am a very coffee mug-oriented person. I am also not willing to move away from my coffee machine, as I am a very coffee machine-oriented person. I am also not willing to move away from my coffee beans, as I am a very coffee beans-oriented person. I am also not willing to move away from my coffee grinder, as I am a very coffee grinder-oriented person. I am also not willing to move away from my coffee filter, as I am a



Non-Answerable Question: In what year does the Net Requirement exceed 25,000?

ChatGPT: 2016/2017/2018/...



# Contributions: Reliable and Robust



Predictive Uncertainty for Probabilistic  
Novelty Detection in Text Classification

ICML UDL 2020

Benchmarking Scalable Predictive  
Uncertainty in Text Classification

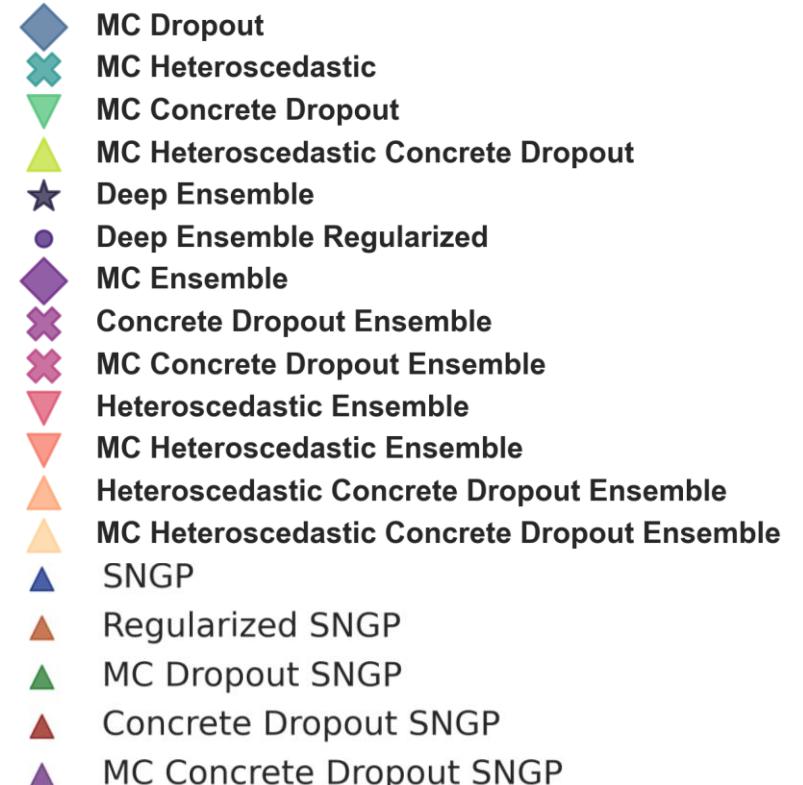
IEEE Access 2022

- BDL survey and literature review
  - PUQ methods NLP benchmark
- 
- Novel hybrid PUQ methods
  - Real-world evaluation setups
  - Take-home guidelines for PUQ



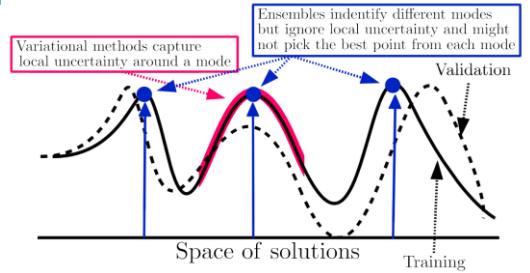
# Presenting the first, comprehensive benchmark for scalable PUQ in NLP

- ✓ 6 text classification datasets
- ✓ 2 neural network architectures
- ✓ 6 unique, 28 total uncertainty methods
- ✓ 5 uncertainty measures
- ✓ 3 experiment setups
- ✓ 5 random seeds
- ✓ 4 hyperparameter ablations

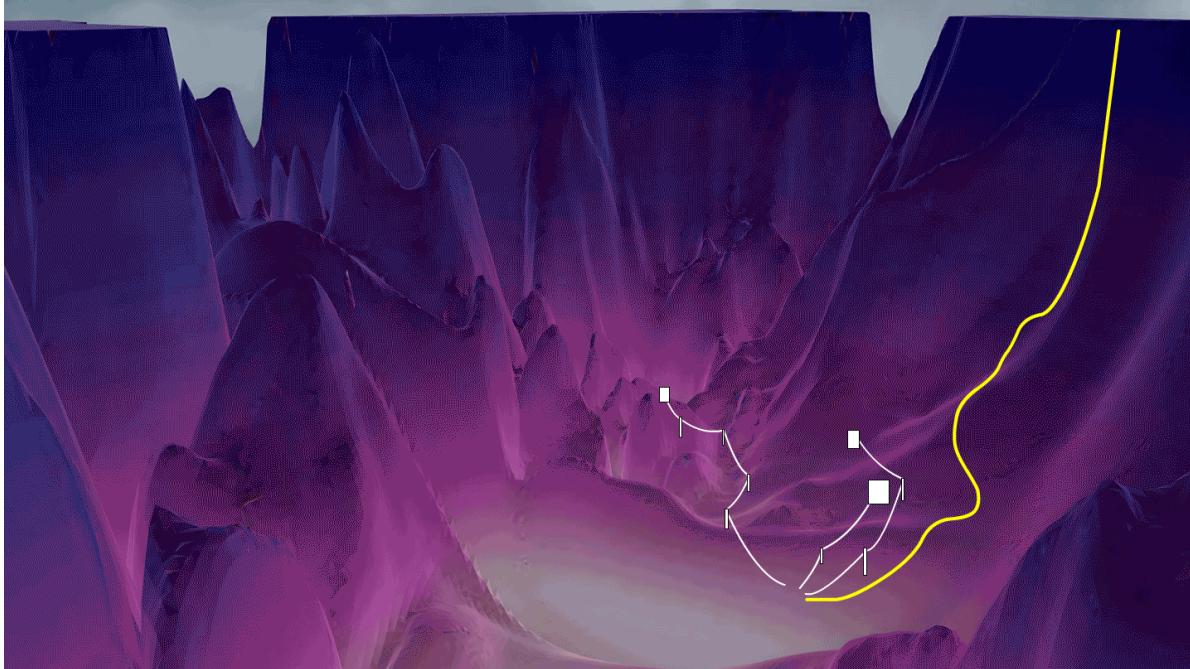




# Proposing novel, hybrid PUQ methods from complementarity in function space



Variational Inference (MC Dropout (*Gal 2016*))



Deep Ensemble (*Lakshminarayanan 2017*)



Credit: Bryan Van Hauwaert

Source: <https://losslandscape.com/explorer>

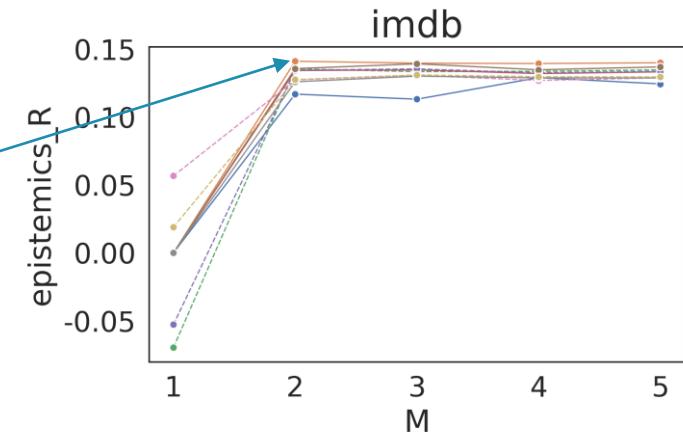
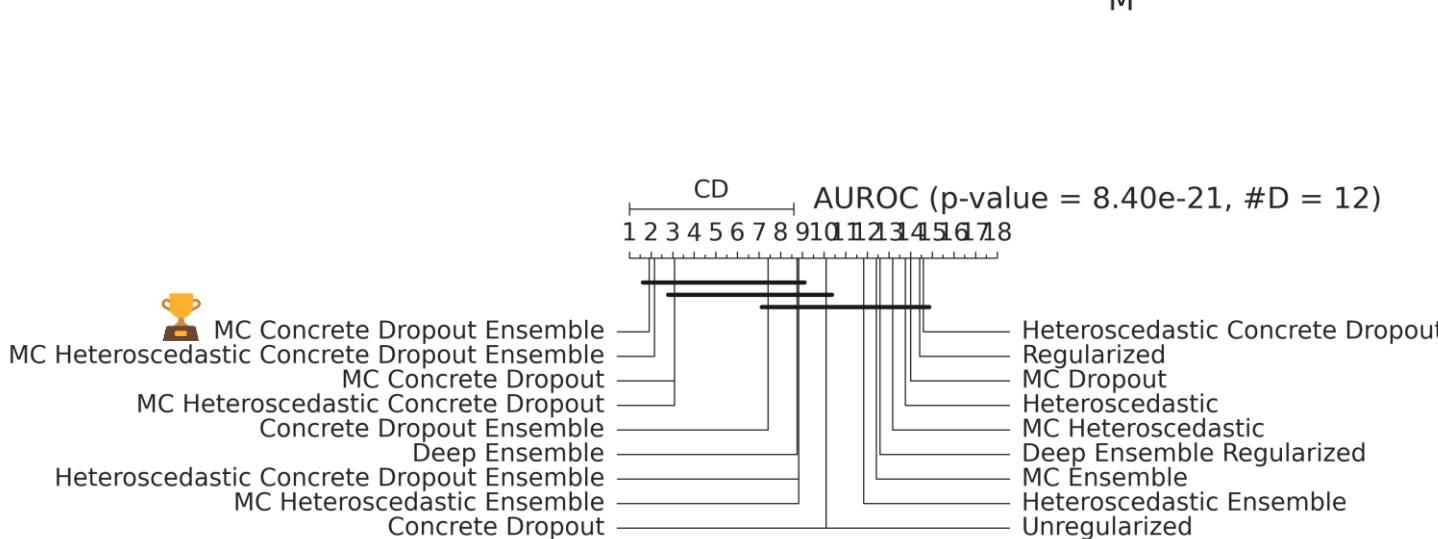
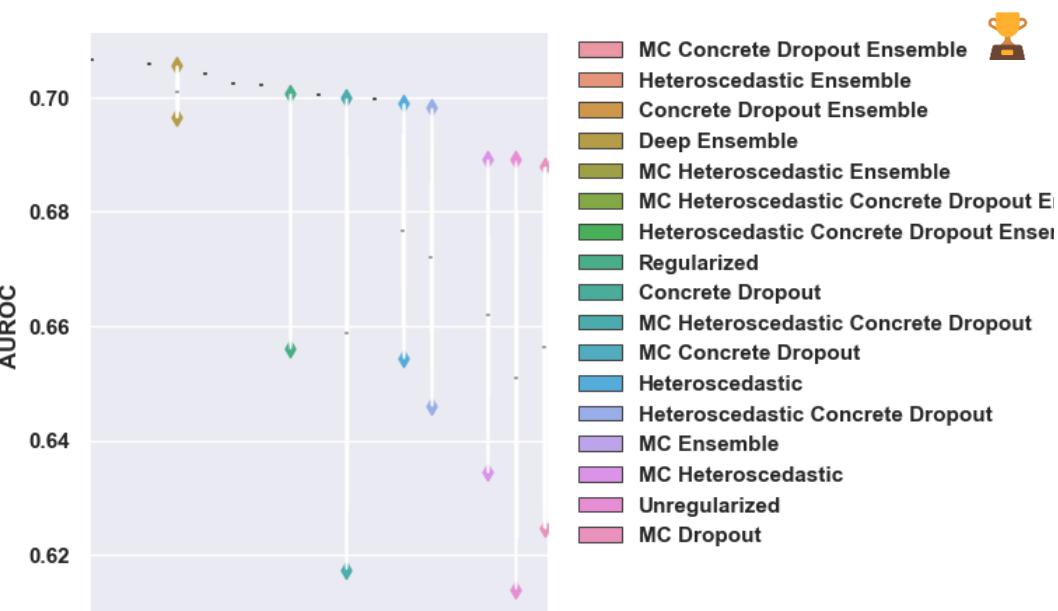




# Finding 1: Proposed hybrid method is superior, at higher efficiency

## MC Concrete Dropout Ensemble:

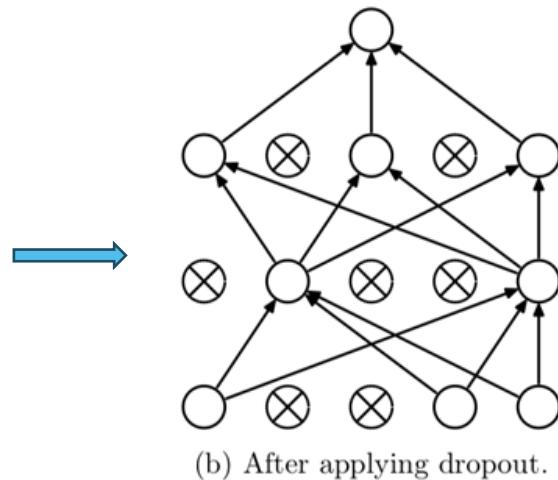
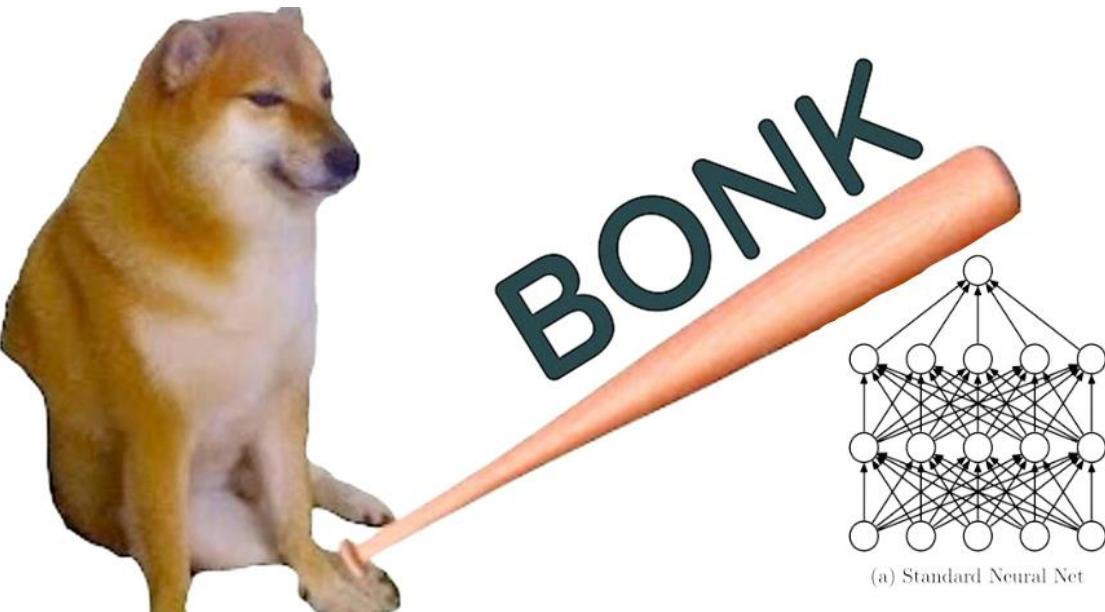
- Presented empirical evidence for theory of complementarity in function space
- Superior at novel class robustness and out-of-domain detection, even at a lower ensemble size





# Decomposing Monte Carlo Concrete Dropout Ensemble

Randomly set a %  $p$  of neurons/weights to 0



Algorithm 1: MCdropout

**Input:** data  $x^*$ , encoder  $g(\cdot)$ , prediction network  $h(\cdot)$ , dropout probability  $p$ , number of iterations  $B$

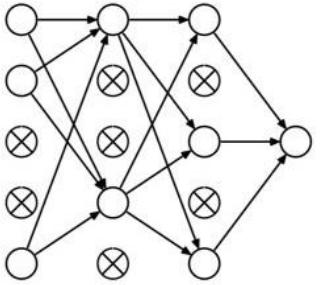
**Output:** prediction  $\hat{y}_{mc}^*$ , uncertainty  $\eta_1$

```
1: for  $b = 1$  to  $B$  do
2:    $e_{(b)}^* \leftarrow \text{VariationalDropout}(g(x^*), p)$ 
3:    $z_{(b)}^* \leftarrow \text{Concatenate}(e_{(b)}^*, \text{extFeatures})$ 
4:    $\hat{y}_{(b)}^* \leftarrow \text{Dropout}(h(z_{(b)}^*), p)$ 
5: end for
// prediction
6:  $\hat{y}_{mc}^* \leftarrow \frac{1}{B} \sum_{b=1}^B \hat{y}_{(b)}^*$ 
// model uncertainty and misspecification
7:  $\eta_1^2 \leftarrow \frac{1}{B} \sum_{b=1}^B (\hat{y}_{(b)}^* - \hat{y}^*)^2$ 
8: return  $\hat{y}_{mc}^*, \eta_1$ 
```

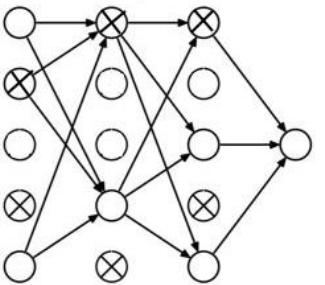




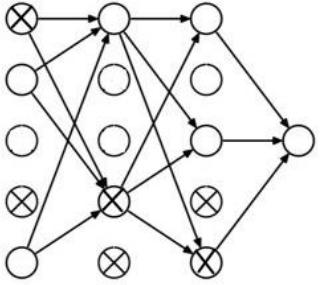
# MC dropout: low uncertainty



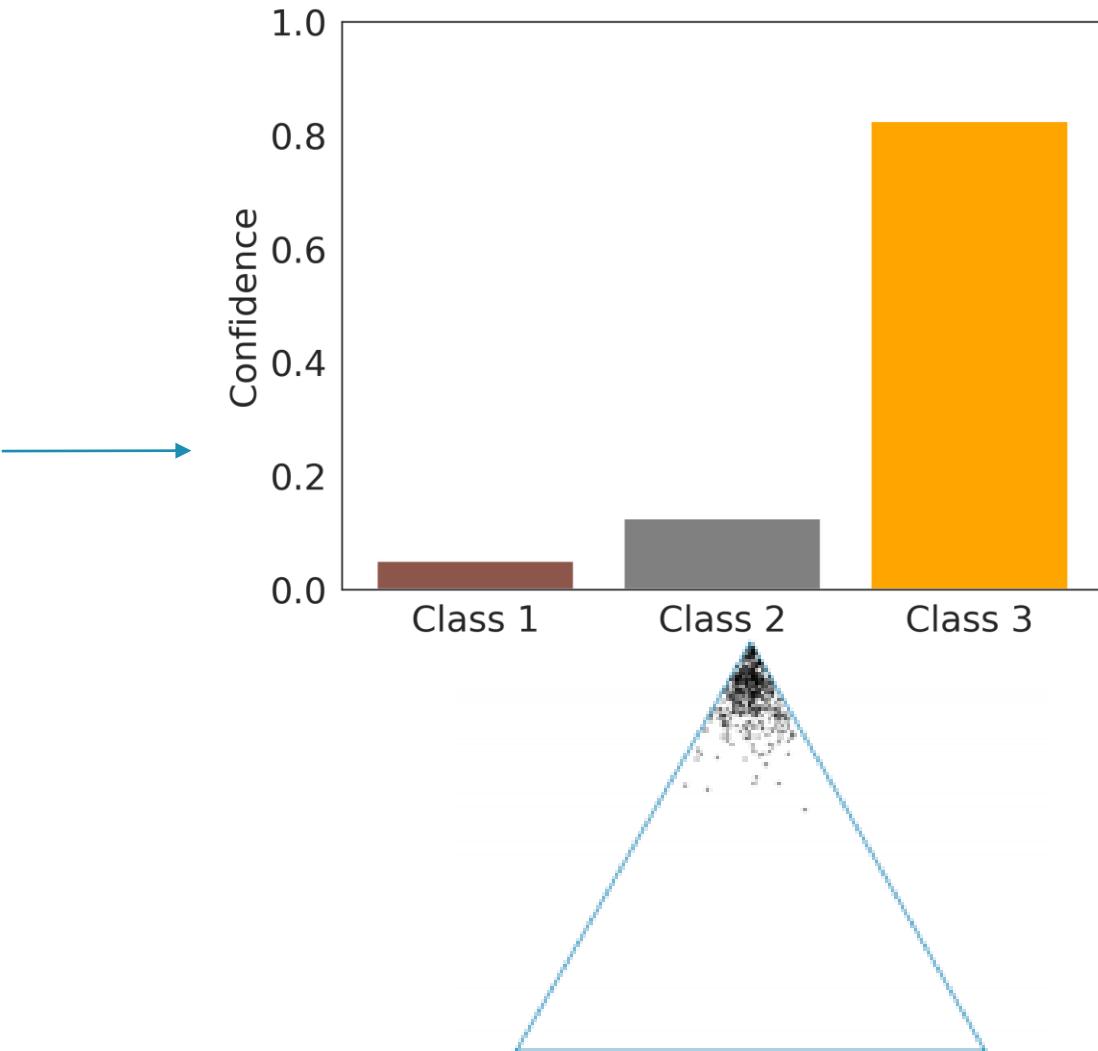
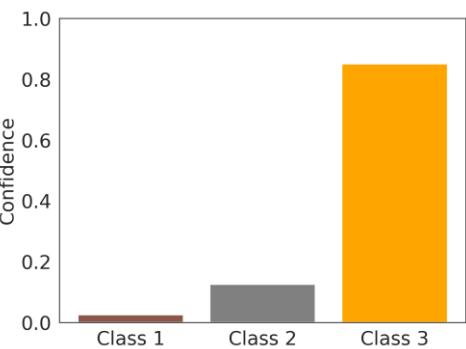
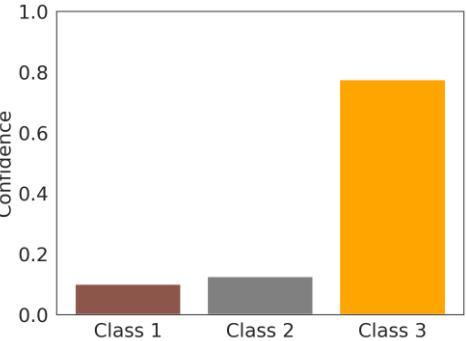
Sample  
1



Sample  
2

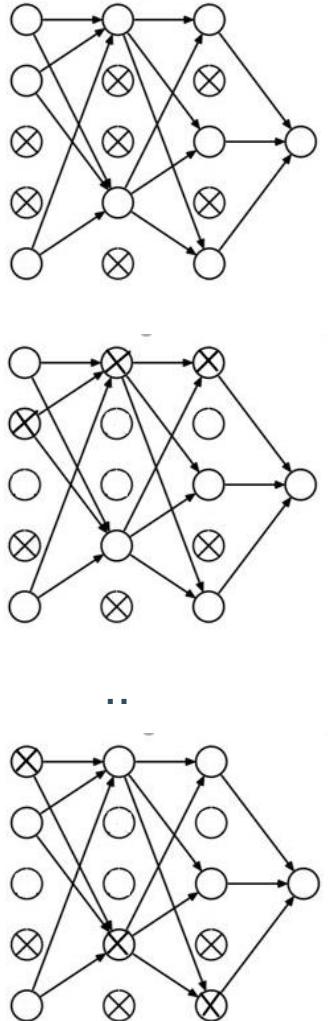


Sample  
T

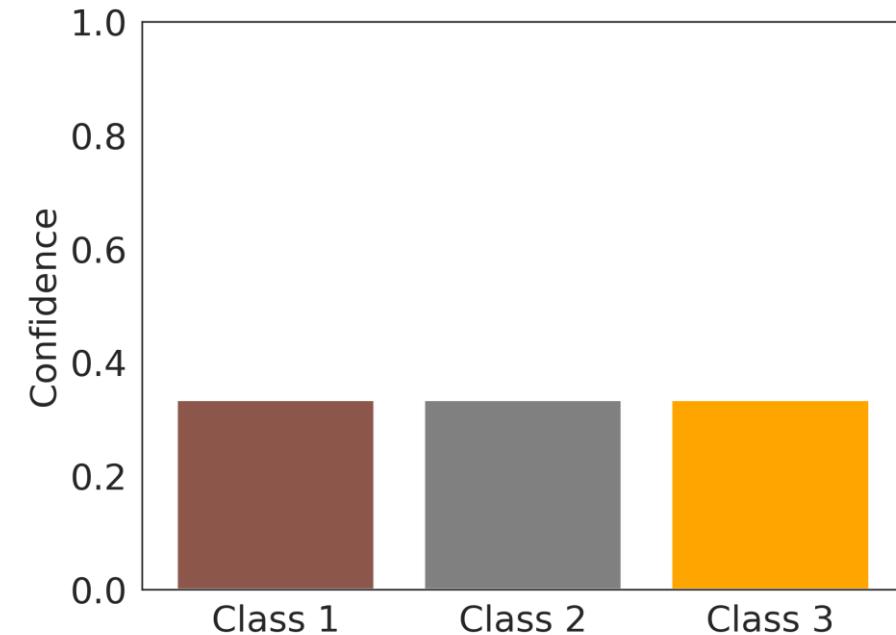
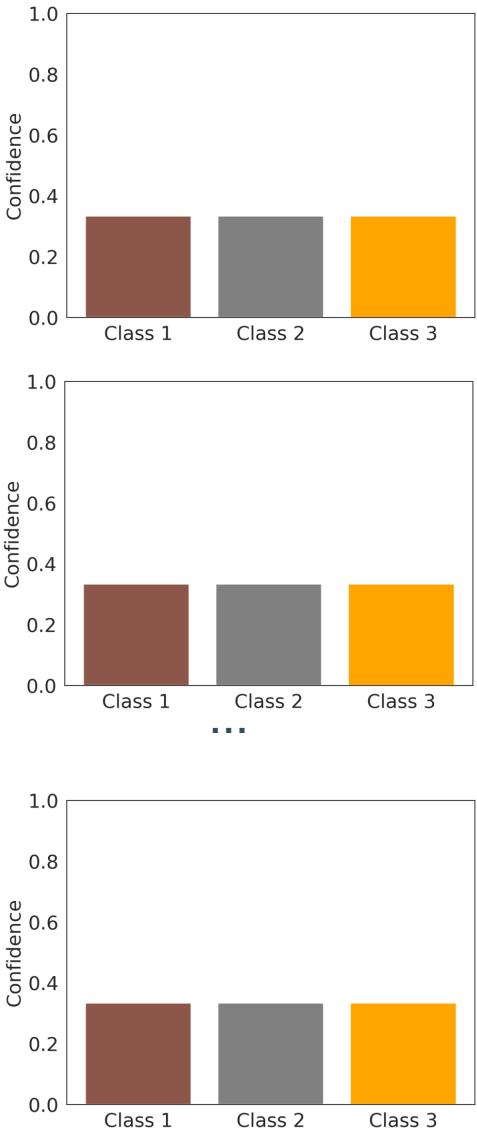




# MC dropout: max *data uncertainty*

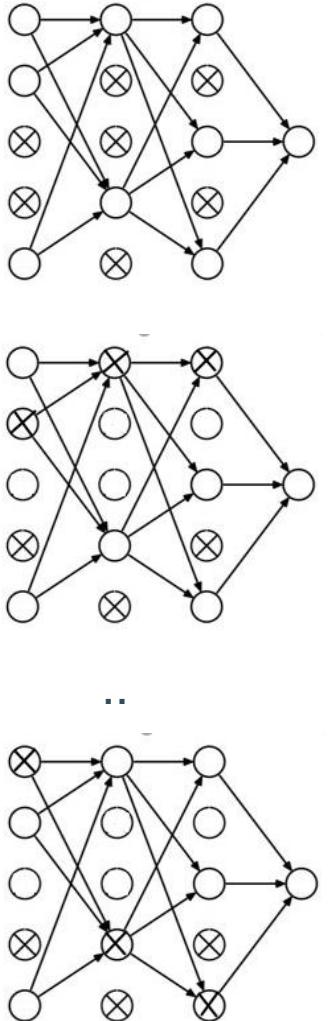


Sample  
1  
  
Sample  
2  
  
..  
  
Sample  
T





# MC dropout: max model uncertainty

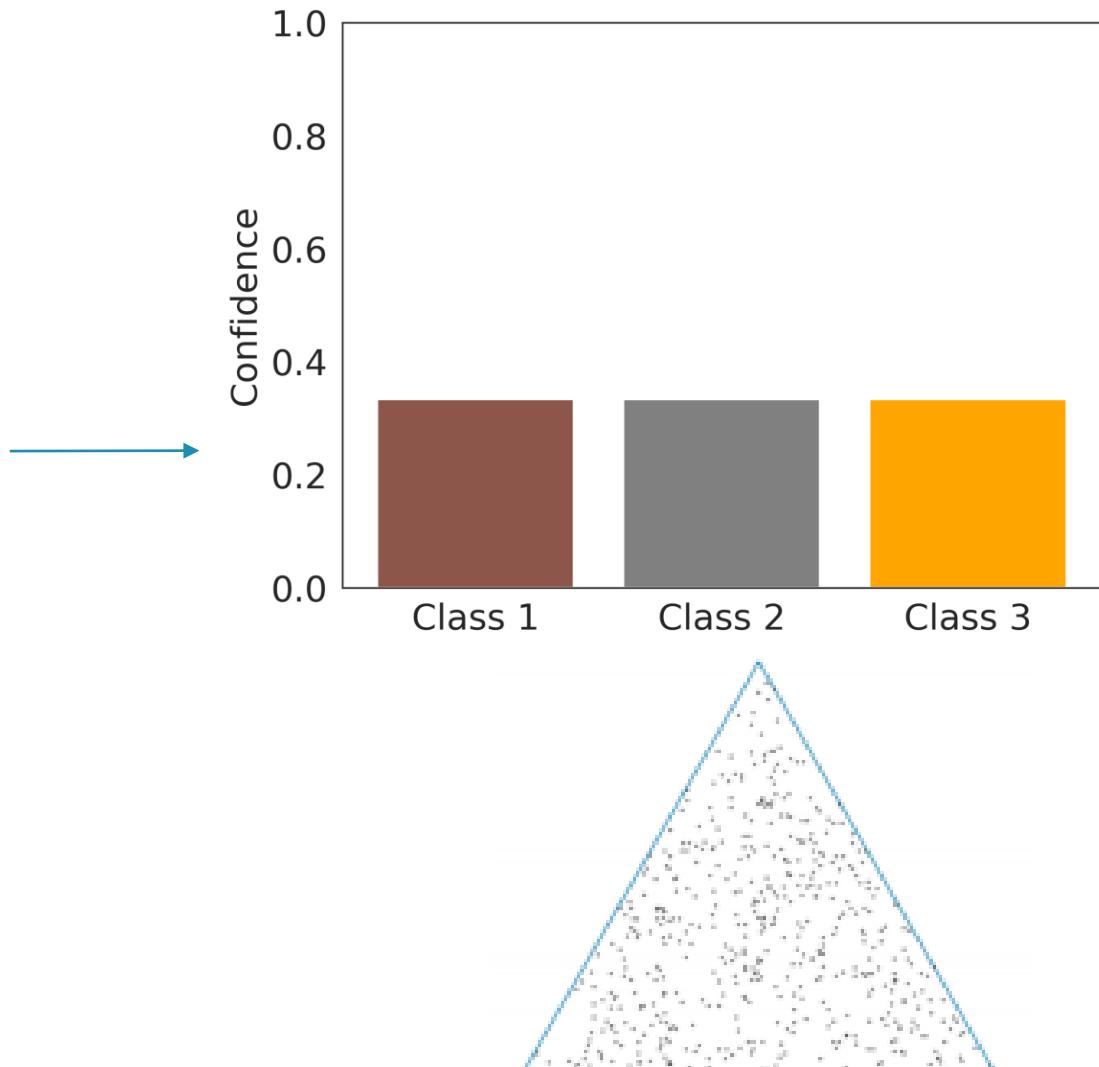
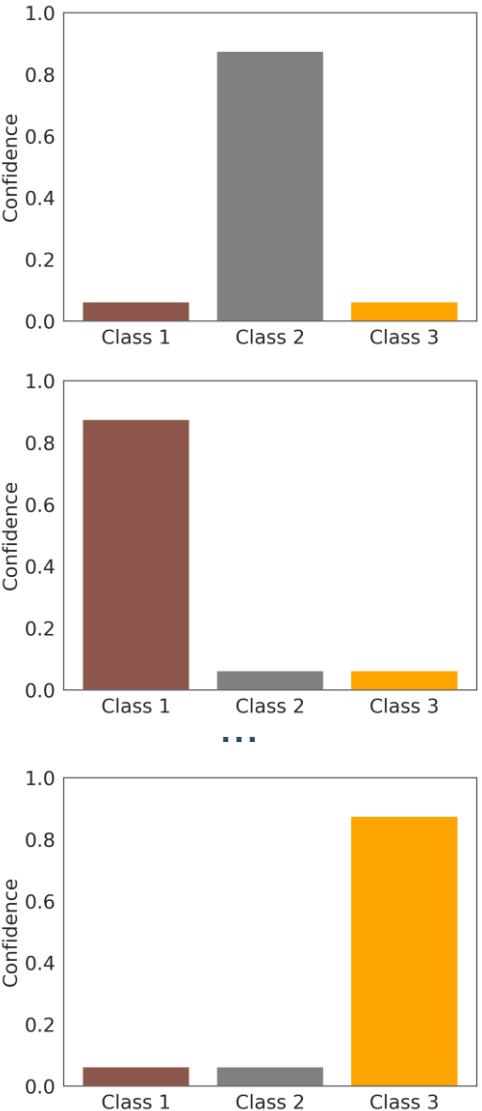


Sample  
1

Sample  
2

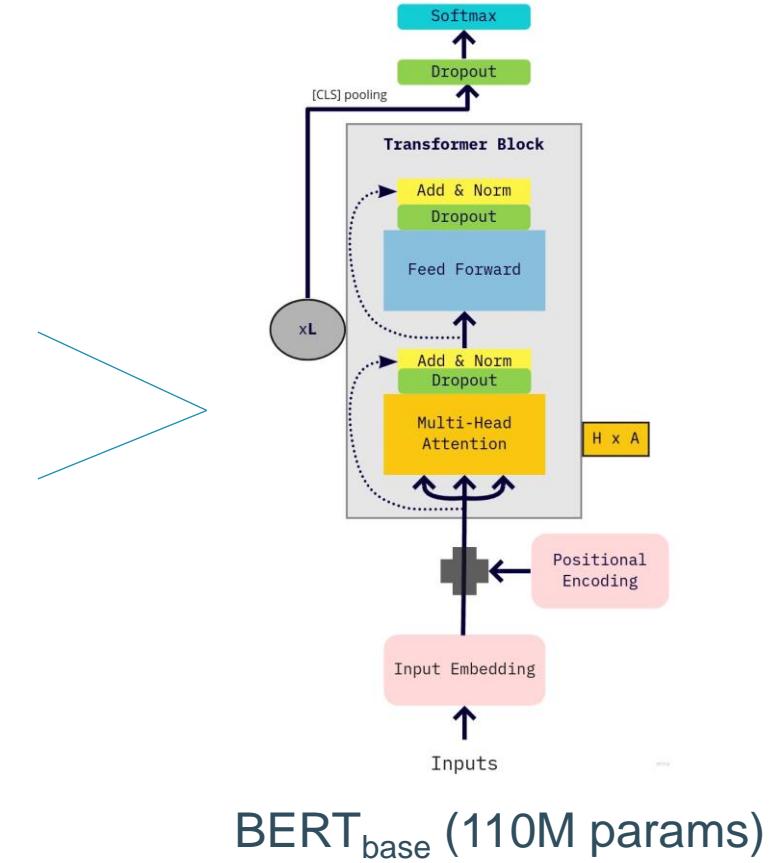
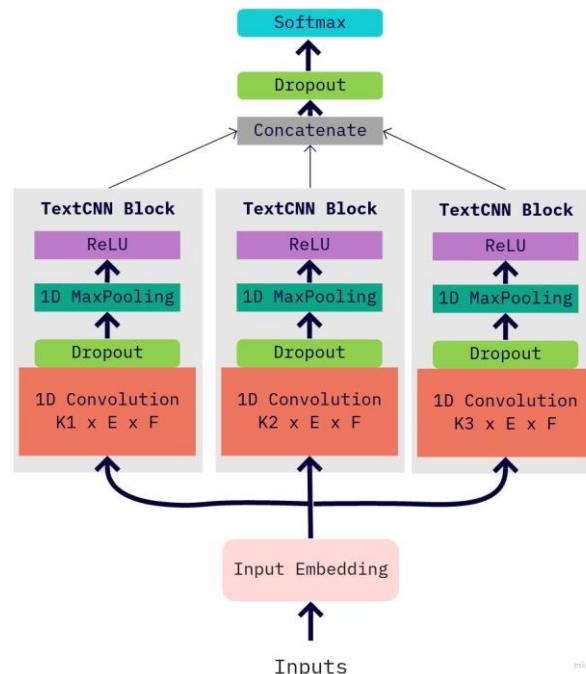
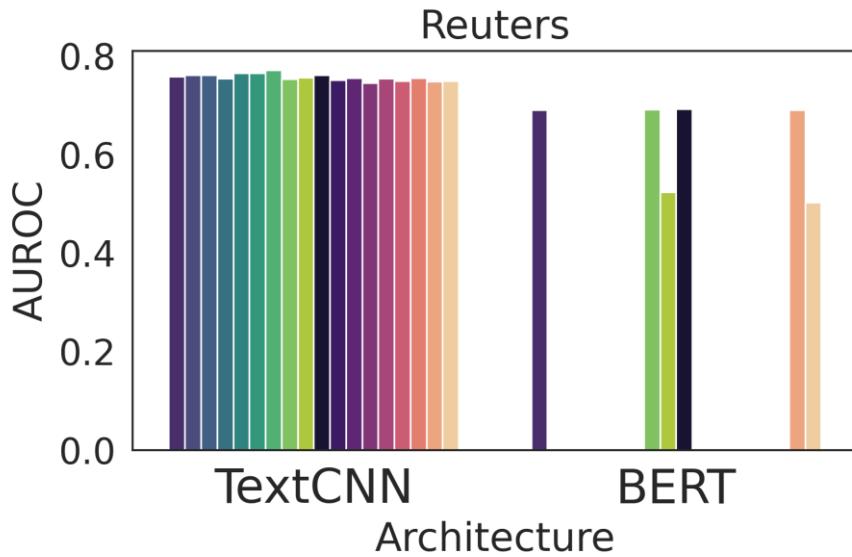
..

Sample  
T





# Finding 2: BERT underperforms in novel class detection

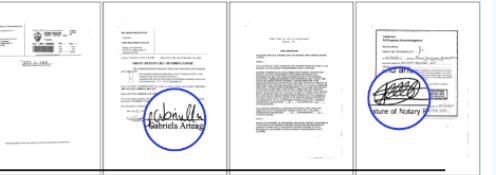


## II. Realistic and Efficient Document Understanding

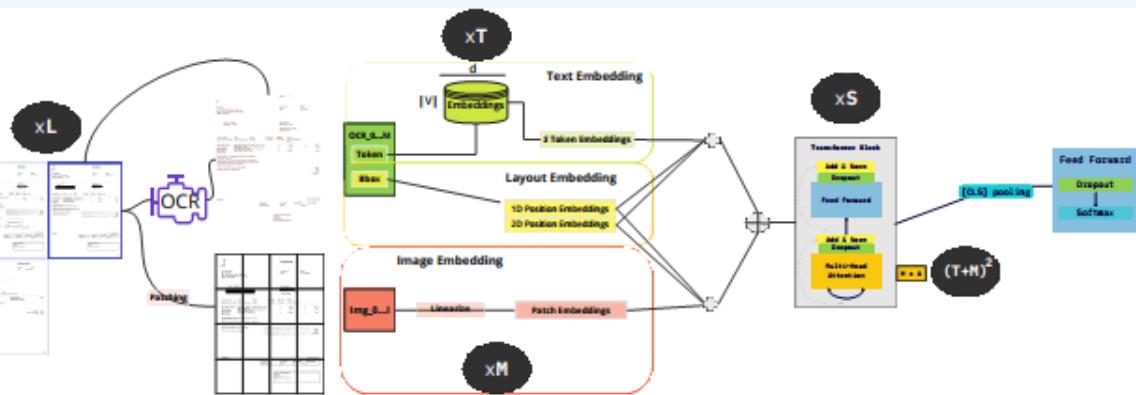


**Requires counting.** How many pages have a signature?

The question requires visual comprehension (recognition of signature), knowledge about layout, and counting.



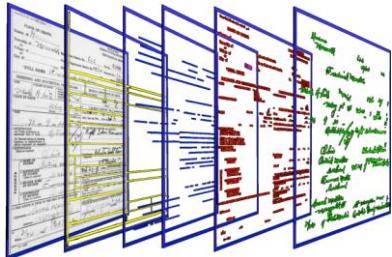
Source	Answer	ANLS	Conf.
Ground truth	2		
Human	2	1.0	—
T5	1	0.0	0.01
ChatGPT	4	0.0	—
GPT3	<i>[Not-answerable]</i>	0.0	—
T5-2D	4	0.0	0.69
HiVT5	4	0.0	0.41





# Shifting the focus to Document Understanding

Focus of the field



Optical Character Recognition



Document Understanding



## Document data unavailability

- Datasets lacking variety, scale and multipage documents
- Current benchmarks evaluation does not transfer downstream

## Pretrain-finetune | Foundation models

- Text-only LLMs for any document task?
- Foundation models more powerful, yet also more cumbersome

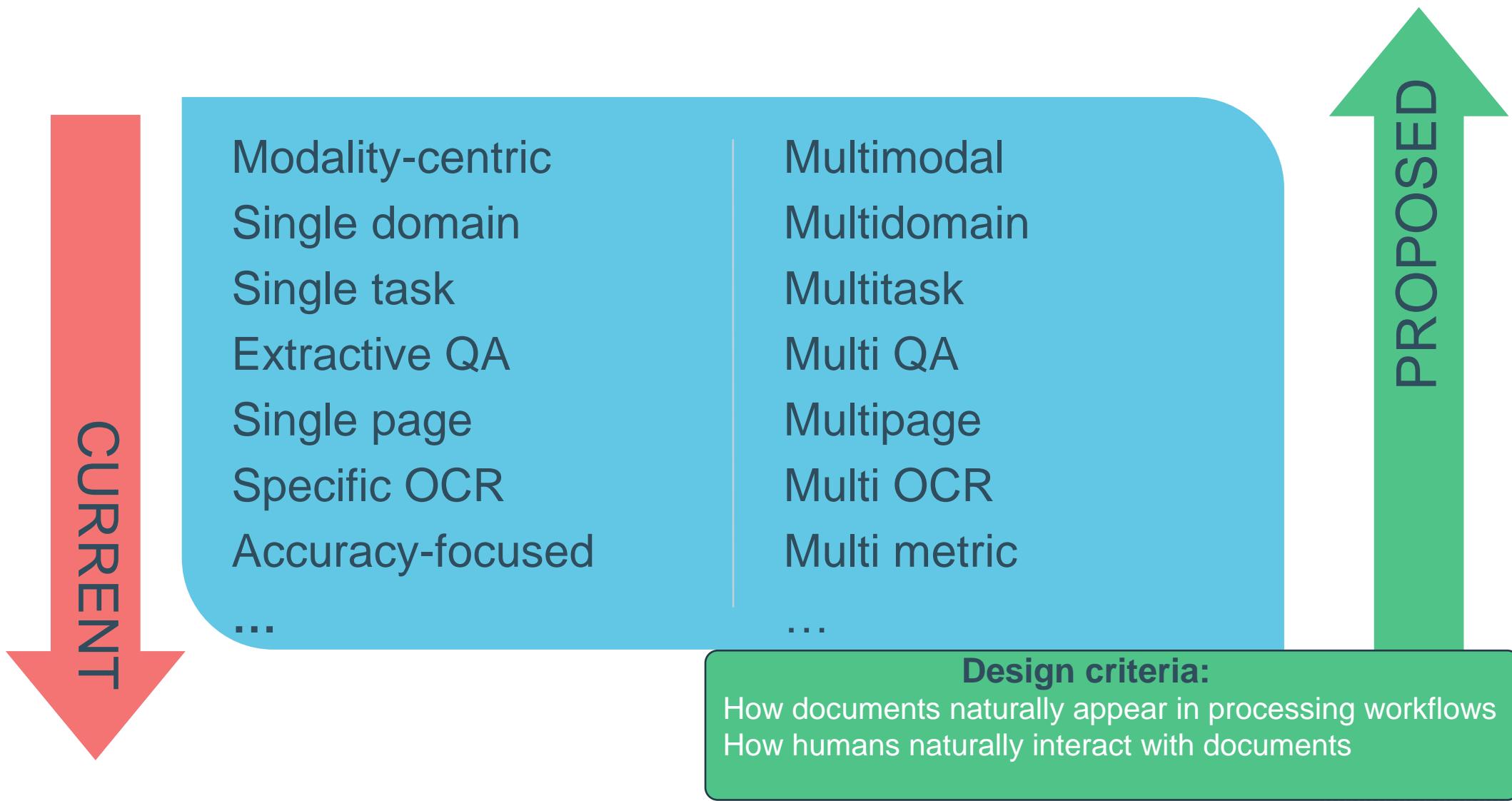
## Objectives

- More generally applicable, embrace real-world complexity
- More efficient at modeling the multimodality of documents
- Evaluation more in sync with downstream requirements





# What are DU benchmarks missing?



# Contributions: Realistic



Beyond Document Page Classification:  
Design, Datasets, and Challenges

WACV 2024 \*oral

- Formalization of multi-page DC
- Construction of two novel datasets
- Survey and recommendations:
  - Complete DC methodology
  - Dataset construction efforts

Document Understanding Dataset and  
Evaluation

ICCV 2023

Competition on Document  
UnderstanDing of Everything (DUDE)  
ICDAR 2023 \*oral



- Design of multi-faceted dataset
- Comprehensive evaluation of SOTA
- Baseline and competition results
- Calibrated, selective generation



# Document classification is more complex than reported

Covered in public research benchmarks					
INPUT TASK	Page $f_p$	Document $f_d$	Document bundle $f_b$	Page stream $f_s$	Page splits $f_m$
LABELS	collision form	purchase invoice	email; resume; application letter	wage slip, wage_slip; bank statement; id_back, id_front; wage_slip	ticket_1, ticket_2, ..., ticket_9
USE-CASE	Insurance claims	Robotic accounting	HR job screening	Loan application	Expenditure



# A multi-faceted benchmark for generic DU challenges the state-of-the-art



## Document UnderstanDing of Everything



### #non-answerable

Q: In which year does the Net Requirement exceed 25,000?

A: None

### #abstractive #counting

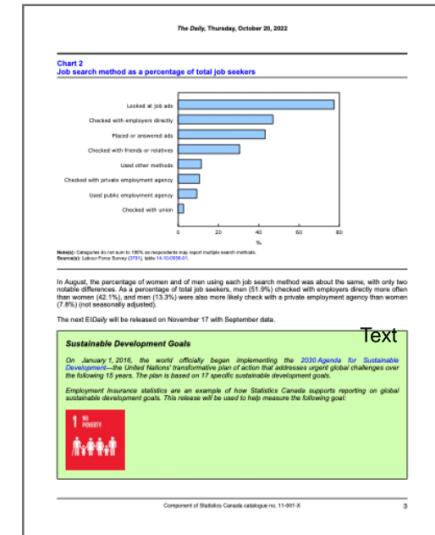
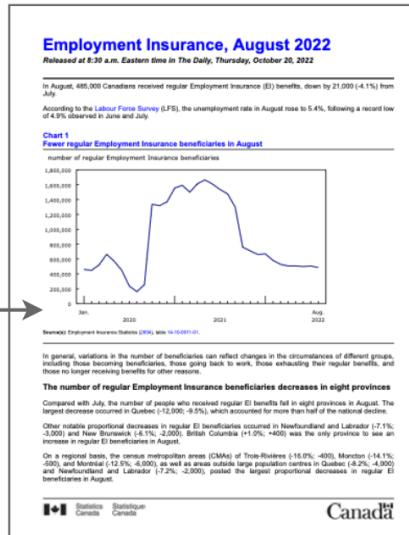
Q: How many attorneys are listed for the plaintiffs?

A: Two

### #layout-navigating #graphic-intensive

Q: Are the margins of the page uniform on all pages?

A: Yes



The figure shows a table titled "Beneficiaries receiving regular income benefits, by occupation, Canada - seasonally adjusted" from The Daily, Thursday, October 20, 2022. The table provides data on beneficiaries across different occupations.

### #extractive #list

Q: What are the Years mentioned in Chart 1?

A: [2020, 2021, 2022]

### #multi-hop #layout-navigating

Q: From the list of Top 10 Key Recovery Components, which is the last component listed on the second page?

A: Hope

### #abstractive #graphic-intensive

Q: Does this document contain any checkboxes?

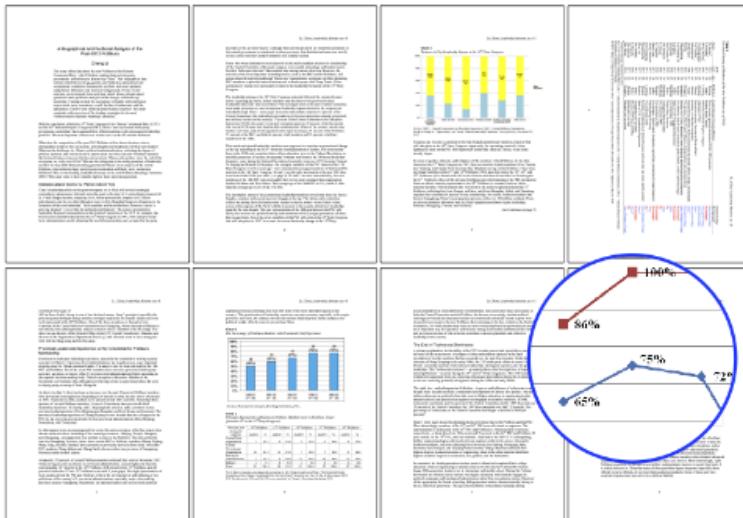
A: No





# -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.

Operator	Pay Rate (\$/hr)
Operator I	\$17/hr
Operator II	\$22/hr
Operator III	\$17/hr

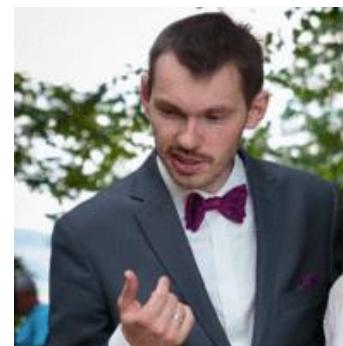
**Requires counting.** How many pages have a signature? The question requires visual comprehension (recognition of signature), knowledge about layout, and counting.

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





# Meet the DUDEs



contract.fit





# Our Baselines vs. Competition

Model	Init.	Params	Max Seq. Length	Test Setup	ANLS <sub>all</sub> ↑	ECE <sub>all</sub> ↓	AURC <sub>all</sub> ↓	ANLS <sub>do</sub>	ANLS <sub>do</sub> Abs	ANLS <sub>do</sub> Ex	ANLS <sub>do</sub> NA	ANLS <sub>do</sub> Li
<i>text-only</i> 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
<i>text-only</i> 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	<b>10.82</b>	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
<i>text-only</i> 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	<b>77.45</b>	17.74
GPT3	davinci3	175B	4000	Concat-0	-	-	-	43.95	18.16	54.44	73.53	36.32
				Concat-4	-	-	-	47.04	22.37	<b>57.09</b>	63.73	<b>40.01</b>
<i>text+layout</i> 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	<b>46.06</b>	14.40	<b>35.70</b>	<b>48.14</b>	<b>50.81</b>	55.65	68.62	5.43
<i>text+layout+vision</i> 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

	Answer	Calibration	OOD Detection	ANLS / answer type
Method	ANLS ↑ ECE ↓ AURC ↓		AUROC ↑	Ex Abs Li NA
UDOP+BLIP+GPT	<b>50.02</b>	<b>22.40</b>	<b>42.10</b>	<b>87.44</b>
MMT5	37.90	59.31	59.31	50.00
HiVT5+modules	35.59	28.03	46.03	51.24
				30.95 35.15 11.76 52.50

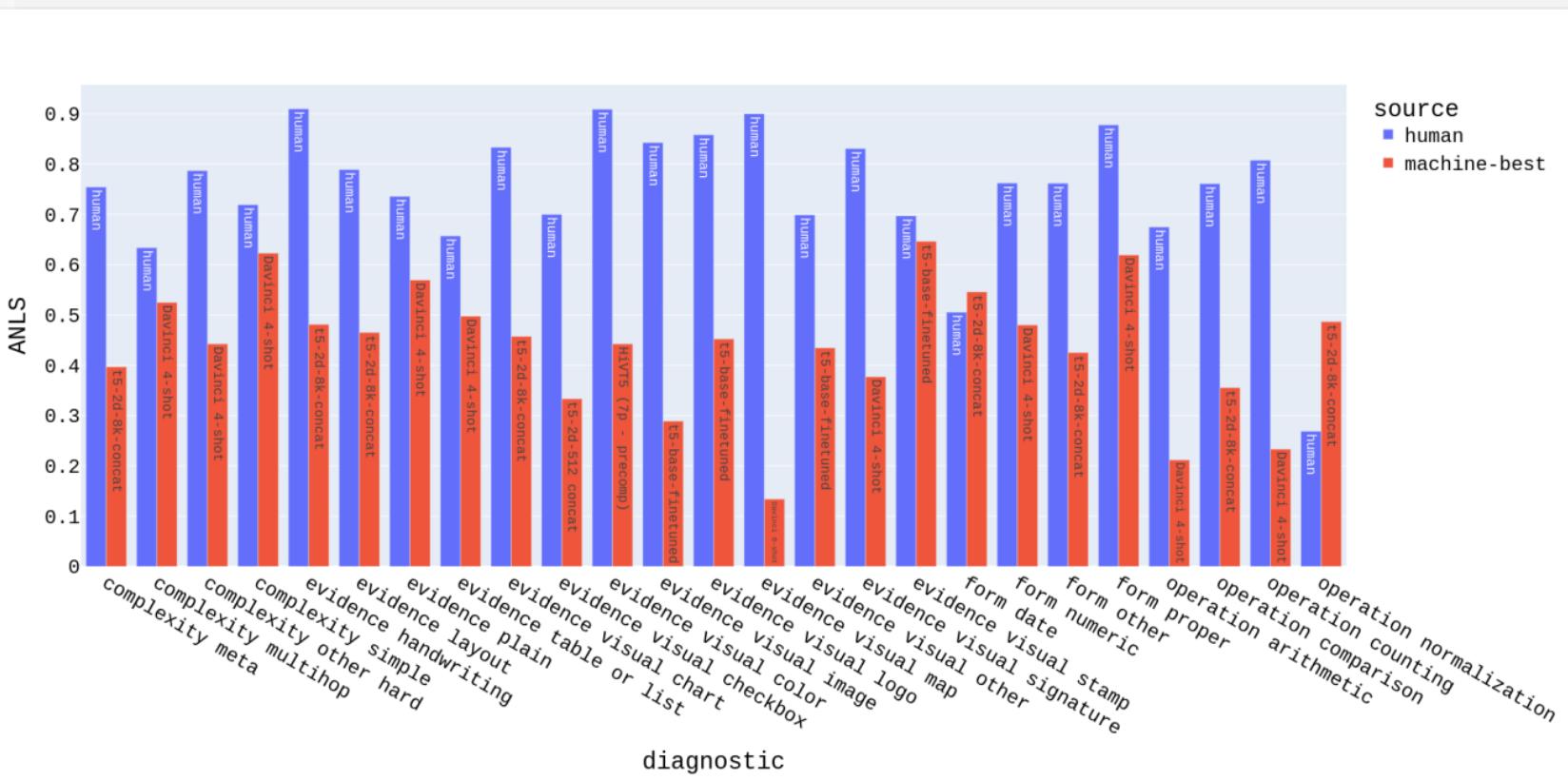
- i. Generative = must
- ii. LLMs are performant
- iii. Outperformed by models +layout understanding ++longer sequence length

SOTA ANLS <= 50%!





# Diagnostic categories shed more light on what models have most difficulty with



Diagnostic categories with

- visual evidence
- reasoning operations

**Baselines lagging far behind human baseline**

# Contributions: Efficient

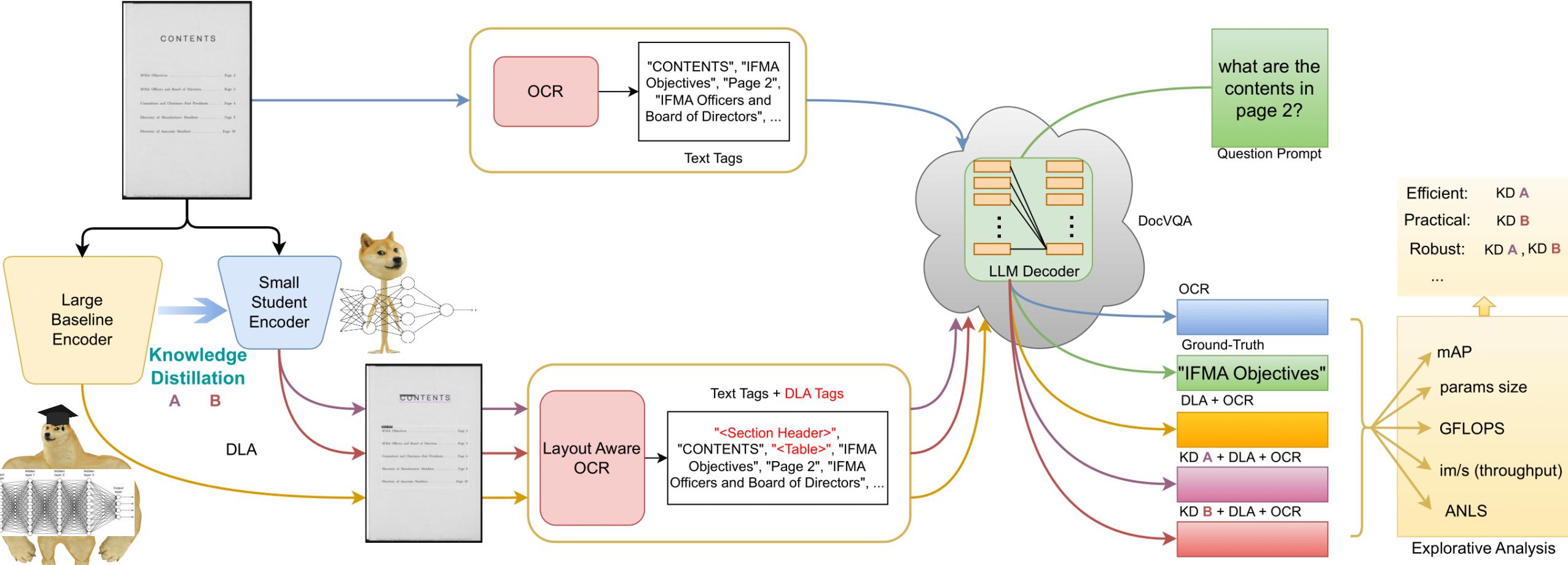


DistilDoc: Knowledge Distillation for  
Visually-Rich Document Applications →  
ICDAR 2024

- KD benchmark on VDU tasks
- Novel downstream evaluation
- Enrich LLMs with semantic layout



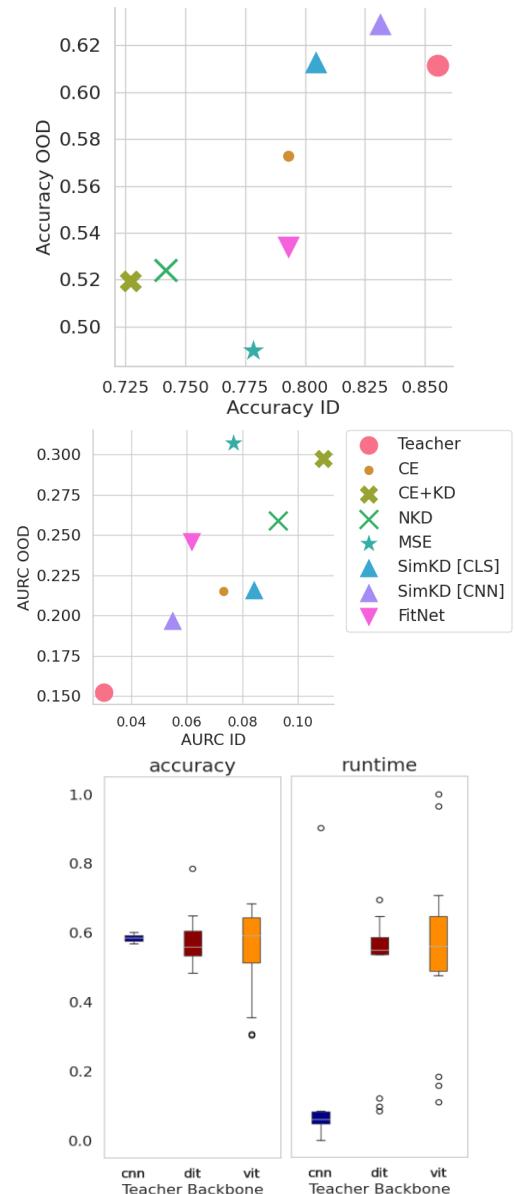
# Knowledge distillation facilitates small, specialized task modules that enrich downstream representations

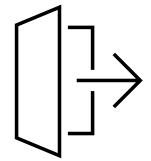




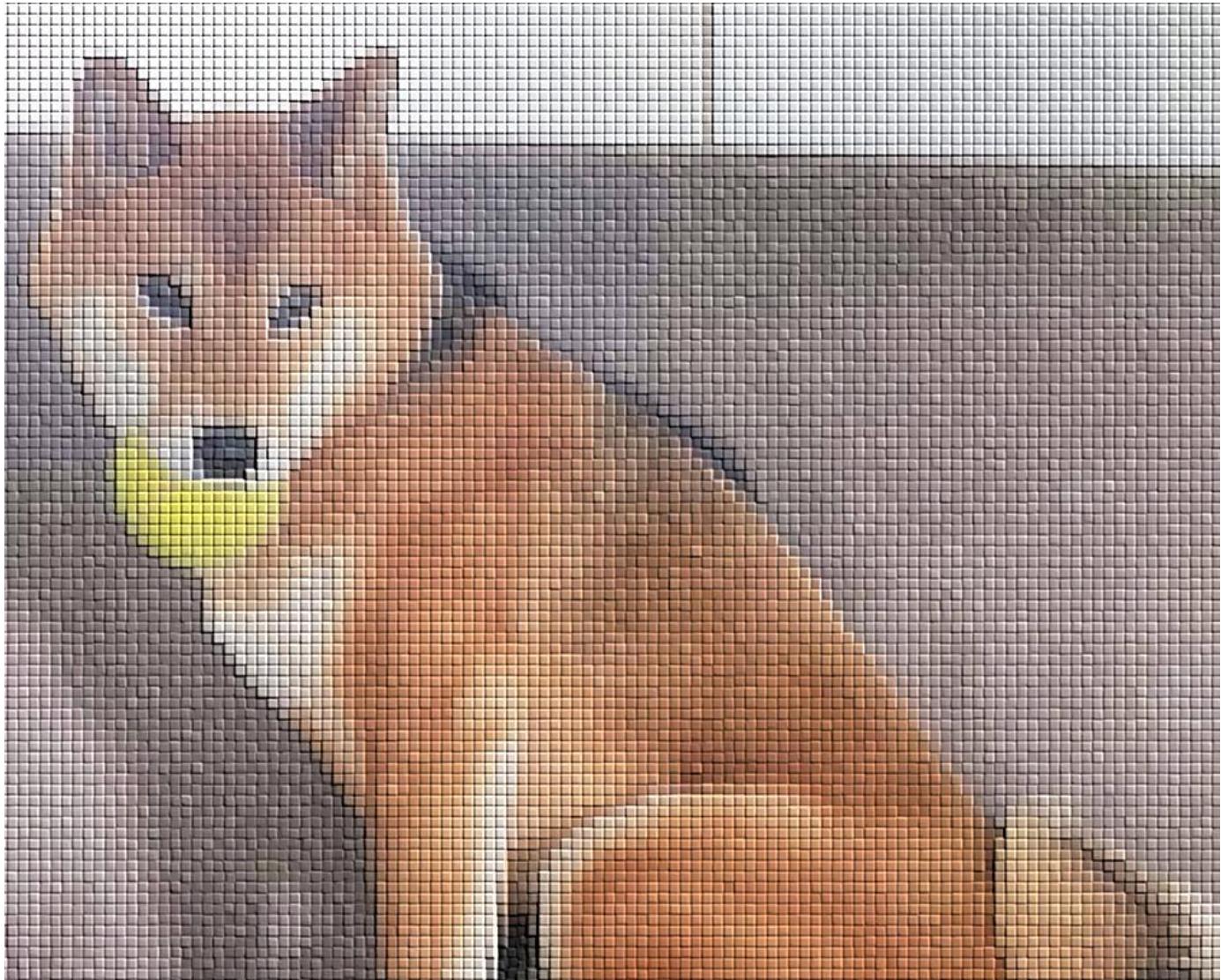
# DistilDoc streamlines research on compression tailored to VDU tasks

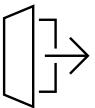
1. *Best KD method*
  - SimKD > vanilla KD, on par with teacher, + under covariate shift
2. *Teacher-Student capacity gap*
  - ViT-Tiny SimKD → 16x smaller model retains 90% rel. accuracy
3. *Impact of Pretraining on KD*
  - ViT2ViT > DiT2ViT, - under covariate shift
4. *Architecture influence*
  - Random initialization & DLA-KD: CNN > ViT
5. *Applicability for downstream tasks*
  - DLA-enriched spacing prompting contributes positively to DocVQA





## Conclusions

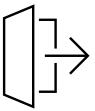




# Conclusions

My dissertation addresses gaps, proposes methodologies opening new opportunities:

1. Limited research on scalable uncertainty quantification in NLP
  -  Comprehensive survey and benchmark
  -  Design of hybrid PUQ methods, offering better robustness and scalability
2. Disconnect DU research and applications
  -  Complete redefinition of document classification and methodology
3. Unpredictable performance of SOTA for generic DU
  -  Multi-faceted benchmark and competition incorporating all document modalities
  -  Promote the layout modality and how to obtain it efficiently



# Takeaway messages

1. Evaluate AI capability, without forgetting about reliability and robustness
2. Need increasingly complex real-world benchmarks to track DU progress
3. Moving the goalpost to complete documents will drive efficiency research
4. A long way to understanding: *multimodality, compositionality and memory*



Thank You

A striking, ultra-realistic poster featuring a heartfelt "Thank You" message spelled out in a modern, bold font. The background is a visual representation of a million business documents, with different colors, patterns, and textures, creating a dynamic and visually rich atmosphere., poster @  Ideogram 50

# APP: ask my thesis



[huggingface.co/spaces/  
jordyvl/ask\\_my\\_thesis](https://huggingface.co/spaces/jordyvl/ask_my_thesis)

# Q&A



Ask my thesis: Intelligent Automation for AI-Driven Document Understanding

Chat with the thesis manuscript by asking questions and receive answers with reference to the page.

Click the visual above to be redirected to the PDF of the manuscript.

Technology used: [Llama-index](#), OS LLMs from HuggingFace

Spoiler: a quickly hacked together RAG application with a >1B LLM and online vector store can be quite slow on a 290 page document (10s+)

Question: What is ANLS?

Additional inputs:

Clear Submit

Answer

ANLS is a metric introduced for the evaluation of VQA (Visual Question Answering) that is adapted to support unanswerable questions. It is an average normalized Levenshtein similarity, which is a metric for measuring the similarity between two strings. The ANLS metric is invariant to the order of provided answers and is used to evaluate the performance of VQA methods.

Answer Page

24 FUNDAMENTALS

- I. Maximum softmax probability (MSP):  $g(x) = \max_{y' \in \mathcal{Y}} f_{y'}(x)$
- II. Maximum logit:  $g(x) = \max_{y' \in \mathcal{Y}} z_{y'}(x)$ , with logits  $\mathbf{z} \in \mathbb{R}^K$
- III. Negative entropy:  $g(x) = -\sum_{y' \in \mathcal{Y}} f_{y'}(x) \log f_{y'}(x)$
- IV. Margin:  $g(x) = \max_{y' \in \mathcal{Y}} f_{y'}(x) - \max_{y'' \in \mathcal{Y} \setminus \{y'\}} f_{y''}(x)$
- V. Distance-based measures
  - kNN distance: A 1D outlier score derived from the average distance of the feature representation of  $x$  to its  $k$  nearest neighbors in the