KU LEUVEN

Computer Vision Center

Meeting 𝔼xpectations:

Intelligent Automation for AI-driven Document Understanding

June 30, 2023

Jordy Van Landeghem
jordy@contract.fit

whoami

- Lead AI researcher @**Contract.fit** since 2017
- Ongoing Ph.D. project @**KU Leuven** on *Intelligent Automation* ***(IA) for*** *Artificial Intelligence* ***(AI)-Driven*** *Document Understanding* ***(DU)*** *[IA4AI-DU]*
- Research interests:

  calibration, predictive uncertainty, failure prediction

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

KU LEUVEN

Selected works

- Van Landeghem, J., Blaschko, M., Anckaert, B., & Moens, M. F. (2020). **Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification**. In *Proceedings ICML 2020 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., Blaschko, & Moens, M. F. (2021-2022). **Leaps-and-Bounds: Towards Stronger Calibration Measures for Structured Output Spaces.** [unpublished]
- 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., Tito, R., ..., Anckaert, B., Valveny, E., Blaschko, M, Moens, M. F, & Stanisławek, T. (2023). **Document Understanding Dataset and Evaluation (DUDE)**. *arXiv preprint arXiv:2305.08455*. (under review)
- Van Landeghem, J., Biswas, S., (2023). **Beyond Document Page Classification**. In ACIIDS 2023 (under review).

Ongoing explorations:

- Knowledge Distillation for Document Foundation Models
- A Multi-Modal Multi-Exit Architecture for Efficient Document Classification

Outline

- Intelligent Automation for AI-driven Document Understanding
  - How to enable, measure and improve IA?
- A primer on confidence estimation, calibration and failure prediction
- Linking back to collaborations with CVC
  1. Document UnderstanDing of Everything (DUDE)
  2. Beyond Document Page Classification
  3. Knowledge Distillation for Efficient Document Layout Analysis

KU LEUVEN

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 correctness) a lot of time and resources are spent on manually **categorizing documents** and **extracting crucial information**.

(E)mails

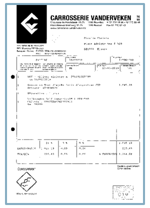
Attachments

Insurance policy

Car invoice

ID Card

Police report



Repair invoice

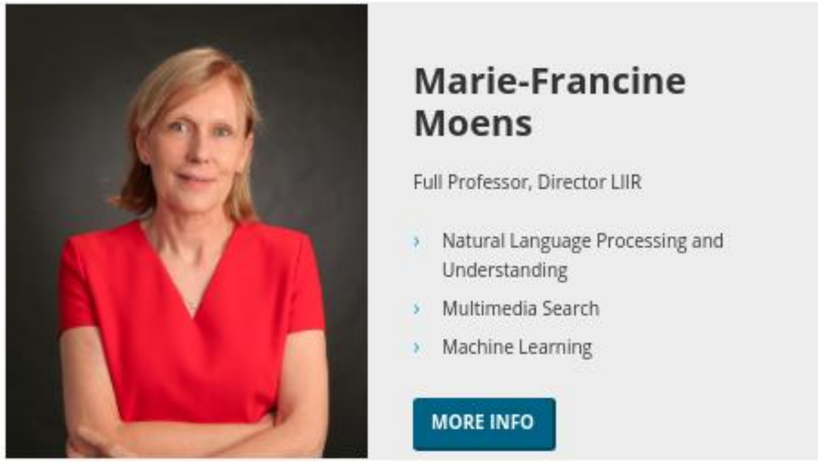
Accident form

IA4AI-DU

- Baekeland Ph.D. project: ⌛ 2020-2024
- Consortium involving University and Company

--> Strategic basic research with economic finality

--> Directed towards obtaining a doctorate diploma



Prof. dr. Matthew Blaschko

- Machine Learning theory
- Computer Vision
- Active Learning

What makes automation intelligent?

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

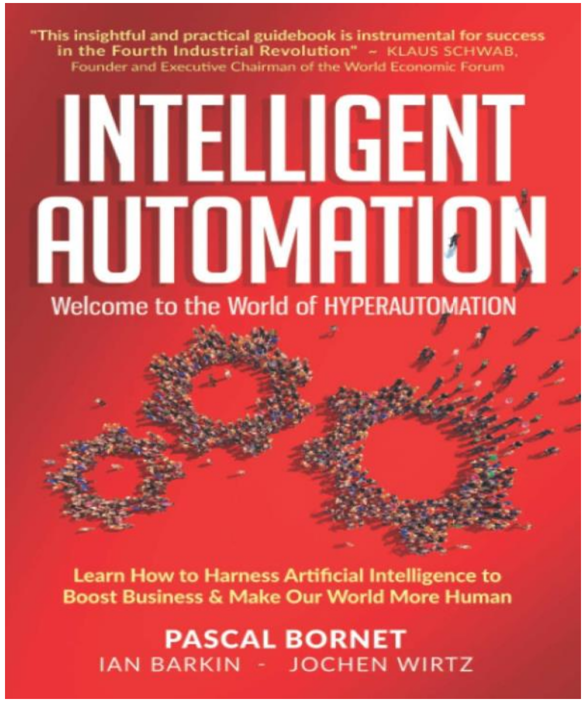
- A subset of Artificial Intelligence (AI) for automation of knowledge work
- Robotic Process Automatic (RPA): the macro on steroids
- Workflow & Business Process Management (BPM)

- jointly capable of solving major world problems
- when combined with people & organizations

- 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

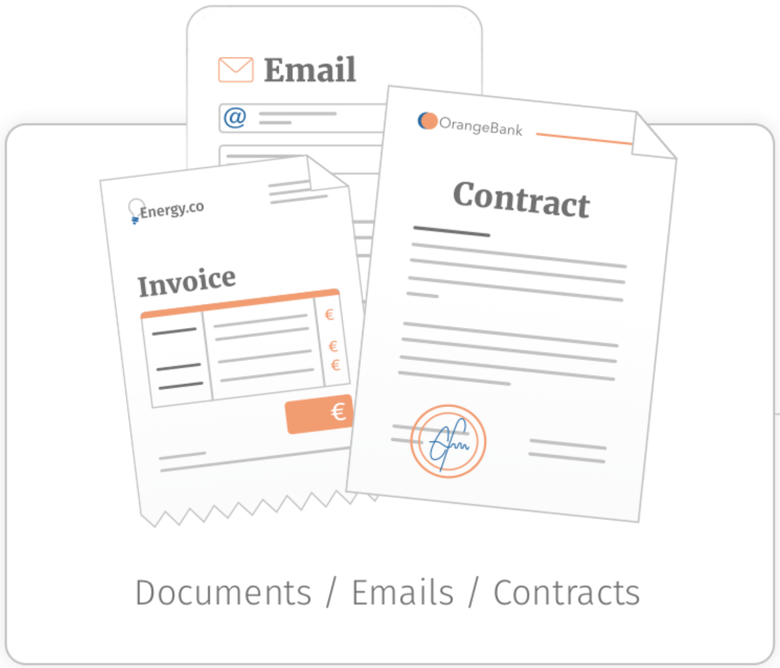
build **straight-through** business processes, which are more efficient (**productivity, processing speed, cost)** and often more effective (**quality and logic**).

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



Pascal Bornet, Ian Barkin and Jochen Wirtz (2020)

contract.fit : Intelligent Document Processing



1 — Classification
- [A][B] Document type
- Email routing
- Prioritize

2 — Capture
- [A,B,C…] Text
- Codes
- Image/Logo

3 — Validation
- [A+B] Calculation
- Signature
- Business rules

The core value proposition of our product involves IA for a variety of document understanding tasks

KU LEUVEN

Document Understanding

Document Understanding (DU) comprises a large set of skills, including the ability to holistically consume textual and visual elements structured according to rich semantic layouts.

The majority of efforts are directed toward the application-directed tasks of classification and key information extraction (KIE) in visually-rich documents (VRDs).

Popular document foundation models: *Document Image Transformer (DiT), LayoutLMv3, Donut, UDOP, Pix2Struct*, ...

Standard benchmark datasets: *RVL-CDIP, PubLayNet, DocBank*, DocLayNet, (DUDE).😎

Our solution brings bottom line impact for countless use-cases

Email routing

Accounts payable

Insurance claims handling

Credit application / disbursement

Expense management

Automated Bookkeeping

One-click onboarding

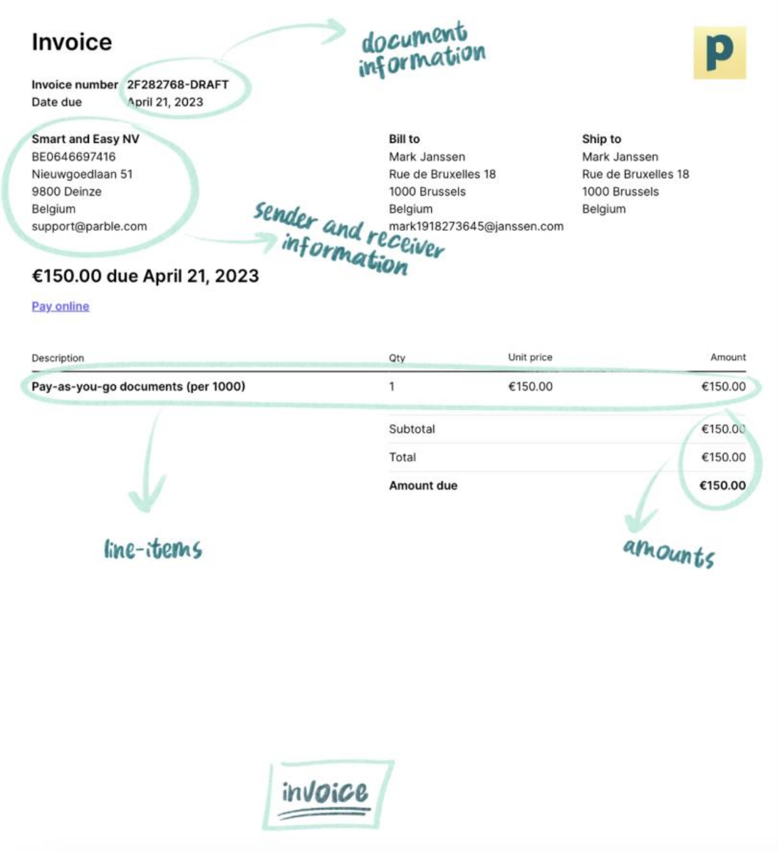
Personal Finance Management

KU LEUVEN

Parble

Process a mix of documents in seconds

Extracted data from an invoice


Invoice

Invoice number 2F282768-DRAFT
Date due April 21, 2023

Smart and Easy NV
BE0646697416
Nieuwgoedlaan 51
9800 Deinze
Belgium
support@parble.com

Bill to
Mark Janssen
Rue de Bruxelles 18
1000 Brussels
Belgium
mark1918273645@janssen.com

Ship to
Mark Janssen
Rue de Bruxelles 18
1000 Brussels
Belgium

€150.00 due April 21, 2023

Pay online

Description	Qty	Unit price	Amount
Pay-as-you-go documents (per 1000)	1	€150.00	€150.00
		Subtotal	€150.00
		Total	€150.00
		Amount due	€150.00

Supported document types include

- Invoices
- Receipts
- Purchase orders
- Delivery notes
- Emails
- ID cards
- Passports
- Photos

Simply integrate Parble using four lines of code

Python Node.js cURL

```python
# install the package using "pip install parble"
from parble import ParbleSDK

parble = ParbleSDK("https://api.parble.com/v1/{YOUR_TENANT}", "{YOUR_APIKEY}"
file = parble.files.post("{PATH_TO_YOUR_FILE}")
```

Try it now – just sign up via parble.com/signup (the first 300 documents are free 🎁)

Motivating example: what are the key ingredients for IA?

Decision-making under Predictive Uncertainty

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

Dear,

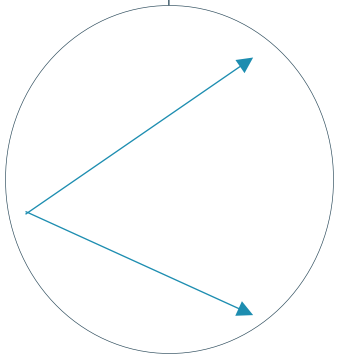
I would like to end the policy for my vehicle 1-CHA-123. The car has been sold and the license plate returned to the authorities (proof attached).

Kindly refund the unused portion of my premium.

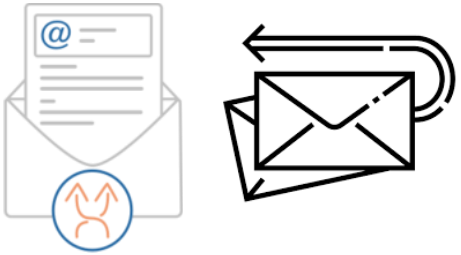
Best,
Jack

In-distribution

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



Automate action



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

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

e.g., novel class

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

Manual review

! **Catastrophically overconfident**

Bringing intelligent automation

Today's seminar 😉

- Enabling IA involves:
  - Confidence estimation
  - Operational thresholding for determining 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

Undervalued in DU studies

- As a proxy to the 'popularity' of IA-related topics, I did a comparative keyword search in the ICDAR 2021 proceedings.

document	3388
classification	242
key information	56
question answering	106
layout analysis	223

calibration/calibrate	33
temperature scaling	0
failure prediction misclassification detection	0
out-of-distribution OOD	25
predictive uncertainty	0

KU LEUVEN

Meeting $\mathbb{E}$xpectations

Jordy Van Landeghem

July 1, 2023

ToC

Notation I

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y}$ denote the output space as a finite set of discrete labels. Given a sample (x,y) drawn independently and identically distributed (*i.i.d*) from an unknown distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$:

Definition

Probabilistic predictor $f : \mathcal{X} \rightarrow \Delta^{\mathcal{Y}}$ that outputs a conditional probability distribution $P(y'|x)$ over outputs $y' \in \mathcal{Y}$.

Definition (Probability Simplex)

Let $\Delta^{\mathcal{Y}} := \{v \in \mathbb{R}_{\geq 0}^{|\mathcal{Y}|} : \|v\|_1 = 1\}$ be a probability simplex of size $|\mathcal{Y}| - 1$, where each vertex represents a mutually-exclusive label and each point has an associated probability vector v [Pistone and Sempi, 1995].

$\rightarrow$ Consider for simplicity, a multi-class classifier where $\mathcal{Y} = [K]$, for K=3 classes.

Basic setting I

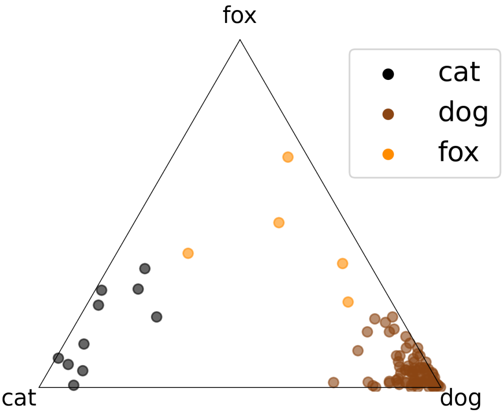


Figure 1: Scatter plot of ternary problem ($K = 3, N = 100$) in the probability simplex space.

Basic setting II

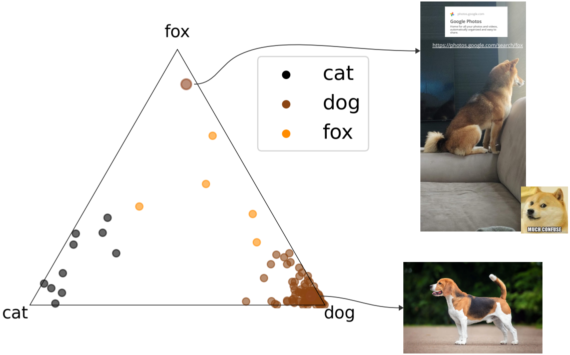


Figure 2: Example of overconfident misprediction (Pabu is a Shiba Inu dog) and correct sharp prediction (clear image of Beagle).

Notation I
continued

Considering standard Neural Networks (NNs), the last layer outputs a vector of real-valued *logits* $\mathrm{z} \in \mathbb{R}^K$, which in turn are normalized using a sigmoid/softmax activation function.

Sigmoid Function	Softmax Function
$\sigma(z) = \dfrac{1}{1 + \exp^{-z}}$	$\mathrm{softmax}(\mathrm{z}) = \dfrac{\exp(z)}{\sum_{k=1}^{K} \exp(z_k)}$

For convenience, $f_k(x)$ denotes the k-th element of the output vector.

- $\hat{y} = \mathrm{argmax}_{y' \in \mathcal{Y}} f_{y'}(X)$ is the top-1 class prediction
- $\hat{p} = \max_{y' \in \mathcal{Y}} f_{y'}(X)$ is the associated posterior probability

Notation II
continued

Some interesting distributions are defined:

$\mathcal{D}_{\text{in}}$ denotes the distribution over $\mathcal{X}$ of in-distribution (ID) data

$\mathcal{D}_{\text{out}}$ out-of-distribution (OOD) data

$\mathcal{D}_{\text{in}}^{\text{test},\checkmark}$ and $\mathcal{D}_{\text{in}}^{\text{test},\times}$ represent the distribution of correct and misclassified ID test samples

Section 2

Confidence Estimation

Confidence scoring function

From model outputs to probabilities

What is “confidence"?

🤔

a method in mathematical statistics for the construction of a set of approximate values of the unknown parameters of probability distributions. Math statistics

raw confidence score, or uncertainty, is a percentage (0-100%), that indicates whether the machine is not sure at all, somewhat sure or very sure about the correctness of a prediction. CF Blog

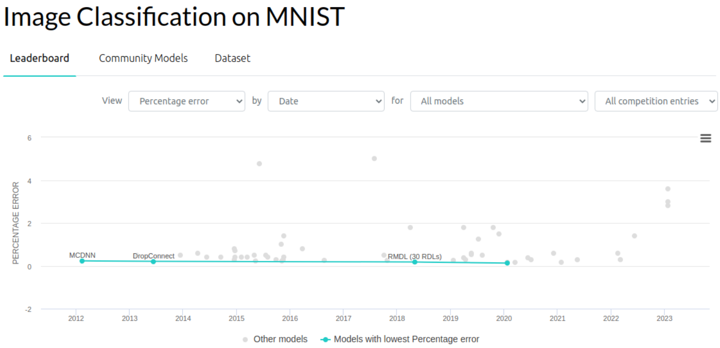
Definition (CSF)

Any function whose continuous output aims to separate a model's failures from correct predictions can be interpreted as a confidence scoring function (CSF). [Jaeger et al., 2023]

Why do we even need to estimate confidence?

🤔

→ ML models are continually improving, yet 0 test error is an illusion*
→ Once a model reaches production, expect deterioration due to *i.i.d* assumptions breaking
→ Generative models are prone to hallucinations, requiring some control mechanism to guide them



KU LEUVEN

CSFs in practice I

MSP and beyond

The most popular CSF is the *maximum softmax probability* (MSP) [Hendrycks and Gimpel, 2017], which is the probability of the top-1 prediction ($\hat{p}$), arising as the largest value from softmax normalization of logits from a final model layer (head).

- A *prediction* is translation of a model's output parameters (as a response to input) to which we apply a standard decision rule, e.g., to obtain the top-1/r predictions.
- For structured prediction models, inference involves decoding according to a function maximizing e.g., total likelihood, diversity, ...

For different tasks, architectures or *failure sources*, CSFs can be more complex.

🫣

CSFs in practice II

- predictive uncertainty quantification (PUQ) [Ghahramani, 2016, Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017, Wilson, 2020, Maddox et al., 2019, Van Amersfoort et al., 2020, Gawlikowski et al., 2021, Mukhoti et al., 2021, Van Landeghem et al., 2022, Mukhoti et al., 2023]
- learning explicit scoring functions (e.g., TrustScore [Jiang et al., 2018], Deep KNN [Papernot and McDaniel, 2018])
- assessing the similarity of inputs to the training distribution [Liang et al., 2018, Liu et al., 2020, Rabanser et al., 2019, Bulusu et al., 2020, Wei et al., 2022]
  - covariate shifts, concept drift, novelty detection, adversarial shifts, domain adaptation
- LLM confidence estimation
  - verbalized probability [Lin et al., 2022] for expressing uncertainty without access to logits
  - semantic entropy [Kuhn et al., 2023] for taking into account semantic equivalence
  - P(*I don't know*) [Kadavath et al., 2022]
  - prompt chaining
    Please give a confidence between 0 and 1 about how certain you are this is the correct answer.

Example outputs for DU task models

Focus: popular DU tasks such as document image *classification*, *KIE* (sequence labeling), *DocVQA* (discriminative span/generative)

Label	**Probability**
invoice	0.85
receipt	0.1
email	0.05

Hal	Jordan	was	the	best	Green	Lantern	ever
PER	PER	O	O	O	MISC	MISC	O
1	1	0	0	0	2	2	0
0.05	0.05	0.7	0.8	0.9	0.1	0.25	0.5
0.9	0.8	0.1	0	0.1	0.3	0.35	0.2
0.05	0.15	0.2	0.2	0	0.6	0.4	0.3

HuggingFace NER example
DUDE T5 DocVQA example

Confidence estimation in KIE

- **Input**: a sequence of tokens $x = \{x_1, x_2, ..., x_T\}$, where $x_t \in \mathcal{V}$ maps to (sub)words in a vocabulary $\mathcal{V}$.
- **Labels**: a sequence of labels $y = \{y_1, y_2, ..., y_T\}$, where $y_t \in \mathcal{Y}$ is a label from a *IOB,IOBES*-encoded labelset $\mathcal{Y}$ (B-Person, I-Person, ..., O).
- *Aggregation strategy* in (*first*, *average*, *max*) for combining subword logits into token logits.

```
# Standard forward pass
outputs = model(**inputs)

# mapping over subwords to token indices
prediction_masks = inputs.word_ids()

# get all unique token indices, skip special start token (None)
words = np.unique([mask if mask is not None else -100 for mask in prediction_masks])[1:]

# for each word spread over multiple subwords, obtain a word confidence
for word in words:
        word_idx = (prediction_masks == word).nonzero()[0]
if aggregation_strategy == "average":
word_logits.append(np.nanmean(logits[word_idx], 0))
        elif aggregation_strategy == "first":
word_logits.append(logits[word_idx[0]])
        elif aggregation_strategy == "max":
word_logits.append(np.nanmax(logits[word_idx], 0))

#TODO: for predicted span (start,end), obtain confidence by summing a range in word_logits
```

Confidence estimation in DocVQA

Comparing extractive (discriminative span prediction) vs. abstractive (generative) [1]

Encoder-based models will output logits for all possible start and end positions of the answer within the provided context. $\hat{y}$ is the predicted answer span, where $\hat{y} = (\hat{y}_{start}, \hat{y}_{end})$ and $\hat{y}_{start} \leq \hat{y}_{end}$. The logits at the final layer take the shape of $BS \times S \times S$, where BS is the batch size and S is the sequence length of the context.

Decoder-based models are not restricted to spans and can output an arbitrary, though often controllable, amount of text tokens, indicated as S'. The logits at the final layer take the shape of $BS \times S' \times V$, where V is the tokenizer's vocabulary size (32.1K for T5-base). Due to autoregressive decoding, the probability of the token at step t is dependent on steps $[0, t-1]$.

[1] Van Landeghem, Tito, Borchmann, Pietruszka, Józiak, Powalski, Jurkiewicz, Coustaty, Ackaert, Valveny, Blaschko, Moens, and Stanislawek [2023]

MSP for extractive QA

```
# Standard span prediction forward call
outputs = model(**inputs, start_positions=start_positions, end_positions=end_positions)

# Assumes masking all padding and special tokens after softmax with 0
start = outputs.start_logits.softmax(dim=1)
.unsqueeze(dim=0).unsqueeze(dim=-1) #1 x BS x S x 1
end = outputs.end_logits.softmax(dim=1)
.unsqueeze(dim=0).unsqueeze(dim=1) #1 x BS x 1 x S

# Compute the probability of each valid (end < start) start, end pair
candidate_matrix = torch.matmul(start, end).triu().detach().numpy() # 1 x BS x S x S

# Obtain highest scoring candidate span by multi-index argmax
flat_probs = candidate_matrix.reshape((1, -1)) # BS x S*S
batch_idx, start_idx, end_idx = np.unravel_index(np.argmax(flat_probs, 1),
↪  candidate_matrix.shape)[1:]
batch_answer_confs = candidate_matrix[0, batch_idx, start_idx, end_idx]
```

MSP for generative models

```python
# Standard decoder-based greedy forward pass (without labels)
outputs = model.generate(**input_ids, output_scores=True, return_dict_in_generate=True)

# BS x S' x V, dropping EOS token and applying softmax + argmax per token
batch_logits = torch.stack(outputs.scores, dim=1)[:, :-1, :]
decoder_outputs_confs = torch.amax(batch_logits.softmax(-1), 2)

# Remove padding from batching decoder output of variable sizes
decoder_outputs_confs_masked = torch.where(
outputs.sequences[:, 1:-1] > 0,
decoder_outputs_confs,
torch.ones_like(decoder_outputs_confs))

# Multiply probability over tokens
batch_answer_confs = decoder_outputs_confs_masked.prod(1)
```

Section 3

Probability Calibration

The history of calibration
Can we trust the weatherman?

Table 2. Precipitation Probability Forecasts for the Pecos Valley in New Mexico

Forcast Probability (%)	No. of Forecasts	No. of Precipitation Occurrences	Relative Frequency of Precipitation (%)
0–9	0	0	—
10–19	22	4	18.2
20–29	31	7	22.6
30–39	18	6	33.3
40–49	15	8	53.3
50–59	13	8	61.5
60–69	15	11	73.3
70–79	7	6	85.7
80–89	1	1	100.0
90–100	1	1	100.0
Total/average	123	52	42.3

NOTE: Data from Hallenbeck (1920).

Calibration error = | Forecast* Probability - Relative Frequency of Precipitation|

Source: https://twitter.com/PreetumNakkiran/status/1581841505647415297

Defining calibration

Definition (Perfect calibration)

[Dawid, 1982, DeGroot and Fienberg, 1983, Zadrozny and Elkan, 2002] Calibration is a property of an empirical predictor f, which states that on finite-sample data it converges to a solution where the confidence scoring function reflects the probability ρ of being correct. Perfect calibration, $\mathrm{CE}(f) = 0$, is satisfied iff:

$$\mathbb{P}(Y = \hat{Y} \mid f(X) = \rho) = \rho, \quad \forall \rho \in [0, 1] \tag{1}$$

(!) This definition will be worked out in detail later.

Proper Loss Functions I

The study of calibration originated in the meteorology and statistics literature, primarily in the context of **proper loss functions** [Murphy and Winkler, 1970] for evaluating probabilistic forecasts.

Proper Loss Functions II

Strictly proper loss functions like Brier score (BS) [Brier, 1950] and negative log likelihood (NLL) [Quinonero-Candela et al., 2005] calculate instance-level scores

- decompose into a sum of multiple metrics including both accuracy and calibration error [Hernández-Orallo et al., 2012].

$$\mathcal{L}_{\mathrm{BS}}(Y, f(X)) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\mathbb{I}\left(Y_i = k\right) - f_k(X_i)\right)^2 \tag{2}$$

$$\mathcal{L}_{\mathrm{NLL}}(Y, f(X)) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{I}\left(Y_i = k\right) \cdot \log\left(f_k(X_i)\right) \tag{3}$$

Proper Loss Functions III

Negative Log Likelihood (NLL) [Quinonero-Candela et al., 2005] is both a popular loss function (*cross-entropy*) and scoring rule which only penalizes (wrong) log probabilities q_i given to the true class, with $\mathbb{I}$ an indicator function defining the true class. This measure more heavily penalizes sharp probabilities, which are close to the wrong edge or class by over/under-confidence.

Brier Score [Brier, 1950] is a scoring rule that measures the accuracy of a probabilistic classifier and is related to the mean-squared error loss function (MSE). Brier score is more commonly used in industrial practice, since it is an ℓ^2 metric (score between 0 and 1), yet it penalizes tail probabilities less severely than NLL.

On Calibration of 'Modern' Neural Networks

- Research into calibration regained popularity after repeated empirical observations of overconfidence in Deep Neural Networks (DNNs) [Nguyen et al., 2015, Guo et al., 2017]

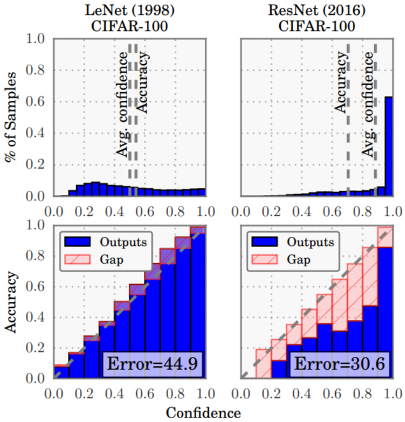


Figure 3: Confidence histograms and reliability diagrams from [Guo et al., 2017]

Broad characterization of calibration research

Characterizing calibration research

calibration metrics

CSF evaluation with both theoretical guarantees and practical estimation methodologies

- Estimators for calibration notions beyond top-1 [Vaicenavicius et al., 2019, Kull et al., 2019, Nixon et al., 2019, Kumar et al., 2019]
- Theoretical frameworks to *generalize* over (existing) metrics [Kumar et al., 2019, Widmann et al., 2019, 2021, Błasiok et al., 2023b]
- *specialize* towards a task such as multi-class classification [Vaicenavicius et al., 2019], regression [Kuleshov et al., 2018, Song et al., 2019], or structured prediction [Kuleshov and Liang, 2015]
- Alternative error estimation procedures, based on histogram regression [Nobel, 1996, Murphy and Winkler, 1977, Niculescu-Mizil and Caruana, 2005, Naeini et al., 2015, Guo et al., 2017], kernels [Kumar et al., 2018, Widmann et al., 2019, 2021, Popordanoska et al., 2022] or splines [Gupta et al., 2020]

(re)calibration methods

Improve calibration by adapting CSF or inducing calibration during training of f

- learn a post-hoc forecaster $F : f(X) \to [0,1]$ on top of f (overview: Ma and Blaschko [2021])
- modifying training procedure with regularization (overview: [Liu et al., 2021, Popordanoska et al., 2022])

$+$ PUQ methods (e.g., Deep Ensemble [Ovadia et al., 2019])

Notions of calibration

Three standard notions of calibration, differing in the subset of predictions considered over $\Delta^{\mathcal{Y}}$ [Vaicenavicius et al., 2019]

- `top-1` [Guo et al., 2017]
- `top-r` [Gupta et al., 2020]
- `canonical` calibration [Bröcker, 2009]

Expected Calibration Error

Definition (ℓ_p Calibration Error)

[Kumar et al., 2019, Vaicenavicius et al., 2019]
The ℓ_p calibration error of $f : \mathcal{X} \to \Delta^{\mathcal{Y}}$ over the joint distribution $(X \times Y)$ with the norm $p \in [1, \infty)$ is given by:

$$\mathrm{CE}_p(f)^p = \mathbb{E}_{(X,Y)} \left[\| \mathbb{E}[Y \mid f(X)] - f(X) \|_p^p \right] \tag{4}$$

- Popular `ECE` metric [Naeini et al., 2015] is a special case with $p = 1$
- $p = \infty$ defines the worst-case risk version known as `MaxCE`.

Formally, a classifier f is said to be *canonically* calibrated iff,

However, the most strict notion of calibration becomes infeasible to compute once the output space cardinality exceeds a certain size [Gupta and Ramdas, 2021].

For discrete target spaces with a large number of classes, there is plenty interest in knowing that a model is calibrated on less likely predictions as well.

Relaxations:

I. `top-label` [Gupta and Ramdas, 2022] (highly recommended)
II. `within-top-r` [Gupta et al., 2020]
III. `marginal` [Kull et al., 2019, Nixon et al., 2019, Kumar et al., 2019, Widmann et al., 2019]

Running example

	$f(x_i)$	$f(x_{ii})$	$f(x_{iii})$	$f(x_{iv})$	$f(x_v)$	$f(x_{vi})$
$f_1(\cdot)$	0.1	0.6	0.2	*0.0*	0.0	0.9
$f_2(\cdot)$	0.0	0.0	**0.7**	0.1	0.1	**0.1**
$f_3(\cdot)$	**0.6**	0.1	0.0	0.1	0.8	0.0
$f_4(\cdot)$	0.3	*0.3*	0.1	0.8	*0.1*	0.0
$\hat{p}$	0.6	0.6	0.7	0.8	0.8	0.9
$\hat{y}$	3	1	2	4	3	1
y	3	4	2	1	4	1

Table 1: Predictions of a fixed model $\mathrm{f} : \mathcal{X} \to \Delta^3$ ($K = 4$) on calibration/test data $\mathcal{D} = \{(i, 3), (ii, 4), \ldots, (vi, 1)\}$ ($N = 6$)

How to go from this to statistical estimates of calibration error? 🤔

🕹️ Let's do it together: Link

Dissecting CE for estimation I

Tractable and practical estimation of any ℓ_p calibration error requires measuring discrepancy between $\mathbb{E}(Y \mid f(X))$ and $f(X)$
$\rightarrow$ estimate conditional expectation for a discrete random variable Y conditioned on a continuous random variable $f(X)$
! $\rightarrow$ not trivially reduced to comparing distances between real-valued vectors

Definition

Binning scales down f to output the average value in each bin B_j:

$$f_{\mathcal{B}}(X) = \mathbb{E}\left[f(X) \mid f(X) \in B_j\right] \tag{6}$$

A binning scheme $\mathcal{B}$ discretizes a continuous random variable $\hat{P} \in [0,1]$ into a set of intervals B such as $B = \{[0, \frac{1}{|B|}], [\frac{1}{|B|}, \frac{2}{|B|}], \ldots, [\frac{|B|-1}{|B|}, 1]\}$ in the case of an *equal-range* binning scheme.

The choice of binning can drastically impact the shape of the reliability diagram and alter the estimated calibration error.

KU LEUVEN

Binning Estimator of ECE

Plenty of histogram-based ECE estimator implementations can be found online, yet many design parameters are not exposed:

I. Adapting number of bins (not the default $|B| = 15$)
II. Different binning scheme (equal-range, equal-mass)
III. Binning range to define operating zone
IV. Proxy used as bin accuracy (lower-edge, center, upper-edge)
V. ℓ_p norm

$\longrightarrow$ https://huggingface.co/spaces/jordyvl/ece

[Van Landeghem et al., 2023] introduced this generic implementation to the ICDAR 2023 Document UnderstanDing of Everything competition.

Understanding temperature scaling and its effect on the metrics of interest I

	Original	f_2: T=0.8	f_3: T=[0,0.8]	f_4: T=2	f_5: T=[1,2]
accuracy (↑)	0.500000	0.500000	0.500000	0.500000	0.500000
F1_macro (↑)	0.541667	0.541667	0.541667	0.541667	0.541667
BS(↓)	0.733333	**0.701493**	0.712969	0.868646	0.855173
NLL(↓)	6.503033	4.431991	**3.227383**	6.785734	6.789682
ECE($\|B\| = 10$, eqr) (↓)	0.300000	0.366667	**0.283333**	0.433333	0.433333
ECE($\|B\| = 10$, eqm) (↓)	0.400000	**0.374696**	0.408064	0.470121	0.516319
AURC* (↓)	0.483333	0.419444	**0.330556**	0.483333	0.586111

KU LEUVEN

Understanding temperature scaling and its effect on the metrics of interest II

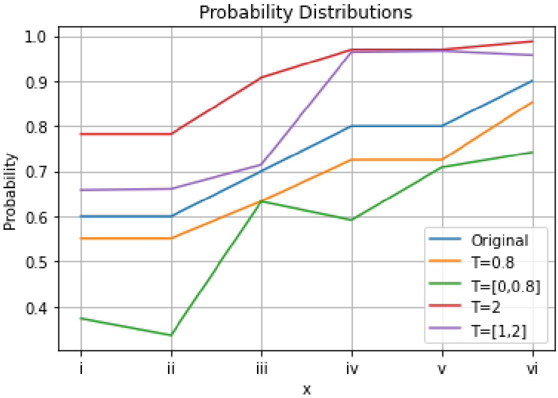


Figure 4: Tempering the probability of the original examples.

$\mathrm{ECE}_{\mathrm{ANLS}}$

Designing a calibration metric for continuous scores.

- As part of [Van Landeghem et al., 2023], we contributed a novel empirical estimator of top-1 calibration for the task of visual question answering, evaluated using average normalized Levenshtein distance (ANLS).

Thresholding for Continuous Scores

Prior work [Munir et al., 2022] introduced the strategy of thresholding continuous quality scores (in the case of IoU larger than τ) in order to be able to estimate ECE.
$\rightarrow$ In our setting, the exact accuracy condition $\mathbb{I}[Y = \hat{y}]$ is replaced by $\mathbb{I}[\mathrm{ANLS}(y, \hat{y}) > \tau]$.

Open problems in calibration

- Calibration metrics/methods adapted to specific tasks
  * Named entity recognition [Kong et al., 2020]
  * Object detection and segmentation [Pan et al., 2021, Dave et al., 2021, Küppers et al., 2022]
- Calibration metrics/methods adapted to specific output spaces
  * My attempt for sequence-structured output spaces (loss-weighted sampling / subgraph decomposition approximation)
- Efficient estimation of "stronger" calibration notions
  * A consistent and differentiable ℓ_p canonical calibration error estimator[Popordanoska et al., 2022]
- Understanding the link between non-convex optimization and calibration
  * Are flat minima required for calibration? [Zhu et al., 2022]
  * When does optimizing a Proper Loss Yield Calibration? [Błasiok et al., 2023a]

Section 4

Failure Prediction

Reflecting on ML evaluation practices

Both in academia and industry, benchmarks are pushing us to achieve higher predictive performance as measured by accuracy, BLEU, ROUGE, mAP, ANLS, ...

I. What if the opportunity resides in better modeling of CSFs, rather than chasing the next minimal goal post of ·% smaller test error?

II. Are we even (correctly) characterizing the errors? [Larson et al., 2023]

III. What (% of) errors can be tolerated in practice? [Flach, 2016]

Follow-up: are novel advances (pre-training, scaling, architectures) beneficial or hurting the detection of *iid* mispredictions? [Galil et al., 2023]

KU LEUVEN

Confidence Ranking

I. more sensible and practically useful notion to consider probabilistic predictions vs. calibration
II. Explicit assessment of i.i.d. failure detection performance is desired for safe deployment
III. Relation to intelligent automation - IDP and FTE savings (business metrics)

Evaluation Metrics:

AUROC (Area Under the ROC Curve)

AURC (Area Under the Risk Coverage Curve) [Geifman and El-Yaniv, 2017, Jaeger et al., 2023]

$\mathrm{E}-\mathrm{AURC}$ (discounting accuracy and normalization) [Geifman et al., 2018]

KU LEUVEN

AUROC

I. AUROC is a threshold-independent measure of the quality of a binary classifier.

II. plots the correct-reject (TN/N) vs. correct-accept (TP/P) ratio for all possible thresholds

III. Lies in the unit square with random choice corresponding to the diagonal, perfect discrimination corresponding to the edges.

AUROC_f [Hendrycks and Gimpel, 2017] for OOD-detection
$\sim$ the probability that a $+$ example (ID) is assigned a higher detection score than a $-$ example (OOD).

- AUROC is not sensitive to the magnitude of the scores, only to their ordering.
- AUROC is not sensitive to any class imbalance.
- AUROC is not a measure of the performance of the classifier.

AURC

Area-Under-Risk-Coverage-Curve (AURC) [Geifman and El-Yaniv, 2017, Jaeger et al., 2023] measures the possible trade-offs between coverage (proportion % of $\mathcal{D}^{\text{test}}$) and risk (error % under given coverage).

Assumptions:

- predictions come with a CSF estimate
- curve is obtained by sorting all CSF estimates and evaluating risk from high to low, together with their respective correctness

$\rightarrow$ AURC for evaluating highly-accurate settings (e.g., 95% accuracy) with risk control

Applied metrics on classic DU task

	layoutlmv3_rvlcdip
accuracy	0.927373
F1_macro	0.927225
BS	0.109506
NLL	1.147134
$\mathrm{ECE}(\lvert B\rvert = 100, \text{eqm})$	0.026559
AURC	0.009194

Table 2: LayoutLMv3-base [Huang et al., 2022] on test set of RVL-CDIP [Harley et al., 2015]

🕹 More examples from *Beyond Document Page Classification* (under review) Link

Choosing the right metric for the job

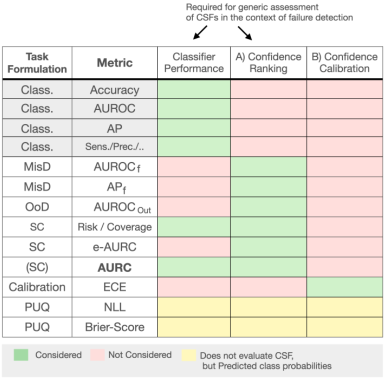


Figure 5: Evaluation metrics for failure prediction [Jaeger et al., 2023].

Why not just use (strictly) proper loss functions? → exclusively operate on the predicted class scores and are not compatible with arbitrary CSFs

Does calibration imply good failure prediction?

Not necessarily: [Zhu et al., 2022]

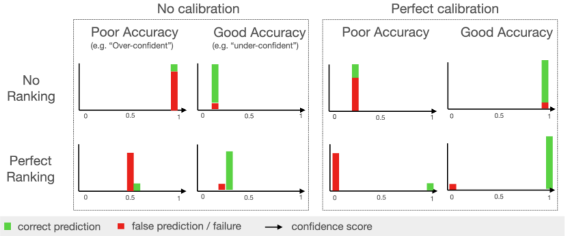


Figure 6: Following [Jaeger et al., 2023], this figure sketches the independent requirements of calibration and confidence ranking.

🕹️ Verified on CIFAR-10 - Resnet18 example, before and after temperature scaling

Open Problems

- ☐ Benchmarking beyond vision architectures (with the same methodological quality as [Galil et al., 2023])
- ☐ Extending failure prediction methodology to multi-task settings
- ☐ Understanding the link between feature space disentanglement and CSF ranking [Zhu et al., 2023]
- ☐ Investigating the relationship between stronger notions of calibration and failure prediction
- ☐ Sample-efficient failure prediction and exploring the connection to semi-supervised learning [Feng et al., 2023]

Additional resources

- Implementations at:

  https://github.com/Jordy-VL/calibration-primer
  https://github.com/Jordy-VL/DUDEeval
  https://huggingface.co/spaces/jordyvl/ece

- Great tutorial ECML 2020 classifier calibration and follow-up [Silva Filho et al., 2023]
- Literature overview:

  Awesome-Failure-Detection

- Slides available at https://jordy-vl.github.io/assets/230630_CVC-Seminar-JVL.pdf

Section 5

Intermediate Conclusions

Important takeaways

→ Lower test error is not all that matters

→ More fine-grained analysis of calibration and failure sources is important

→ The `top-1` (weak, yet popular and efficient to estimate) notion of calibration does not guarantee optimal failure prediction

→ While calibration literature is heavy-to-digest with a high barrier to entry, understanding of the basics already allows access to the low-hanging fruit

→ Collaborations can help bridge the gap between theory and practice

KU LEUVEN

Selected collabs

- 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., Tito, R., ..., Anckaert, B., Valveny, E., Blaschko, M, Moens, M. F, & Stanisławek, T. (2023). **Document Understanding Dataset and Evaluation (DUDE)**. *arXiv preprint arXiv:2305.08455*. (under review)
- Van Landeghem, J., Biswas, S., (2023). **Beyond Document Page Classification**. In ACIIDS 2023 (under review).

Ongoing explorations:

- Knowledge Distillation for Document Foundation Models
- A Multi-Modal Multi-Exit Architecture for Efficient Document Classification

DUDE

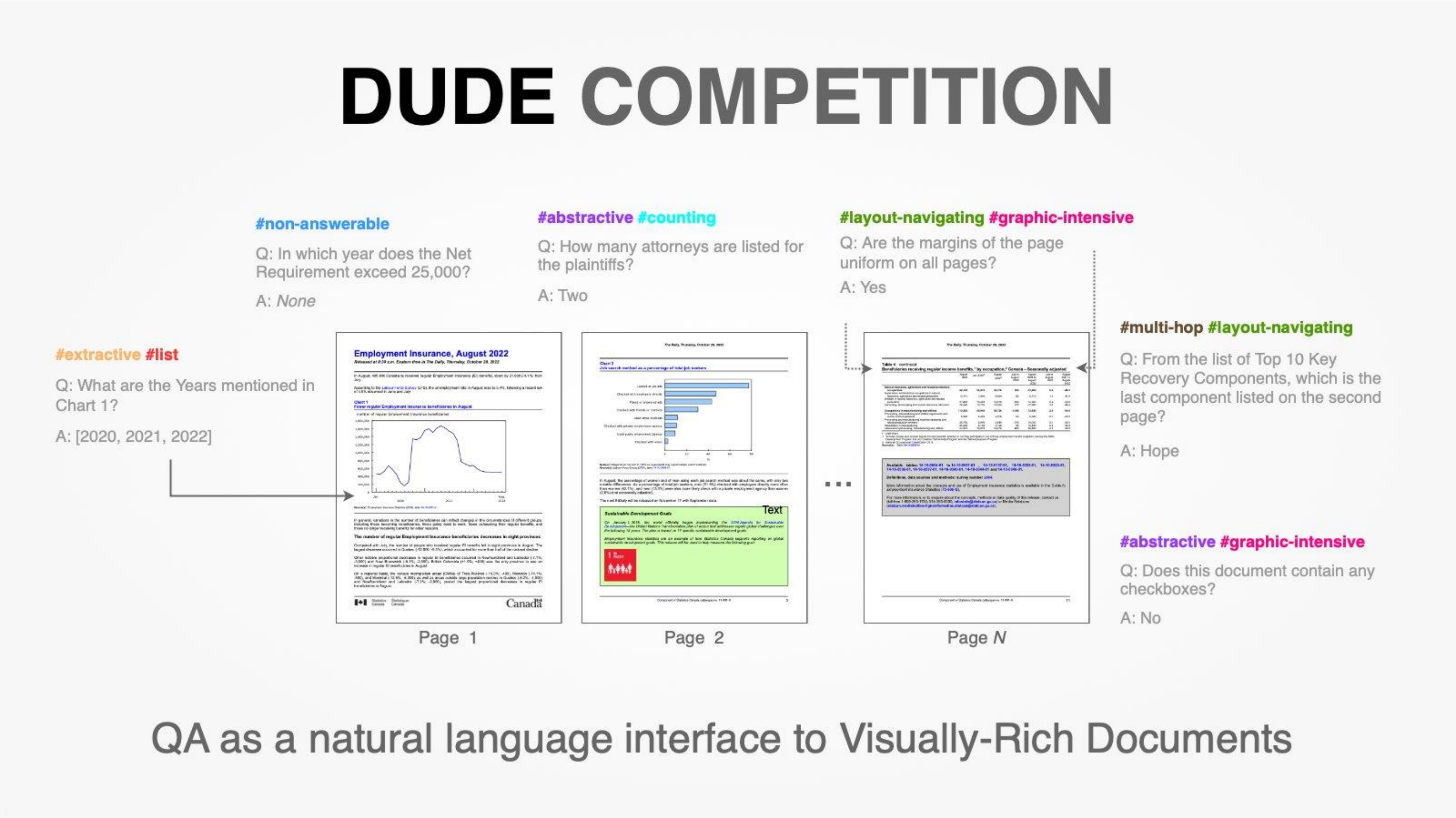
Building a long-standing document understanding benchmark

DUDE 😎

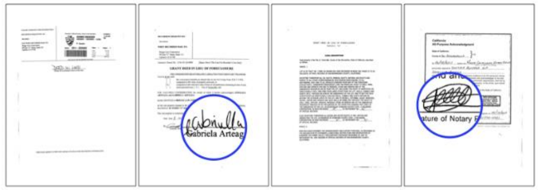
- Foster research on generic document understanding (DU)
- Sourced 5K opensource documents from *archive, wikimedia, documentcloud*
  - **multi-domain** (+15 industries)
  - **multi-type** (+- 200 document types)
  - **multi-page** (μ=5 pages)
  - **multi-QA** (extractive, abstractive, list, non-answerable)
- Bridge QA & DLA:
  - Particular layout semantics (stamp, signature, font style, checkbox)
  - Complex layout-navigating questions demanding multi-step reasoning

DUDE 😎 Competition Document UnderstanDing of Everything

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



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



@ICDAR 2023

DUDE Document Understanding Evaluation and Dataset

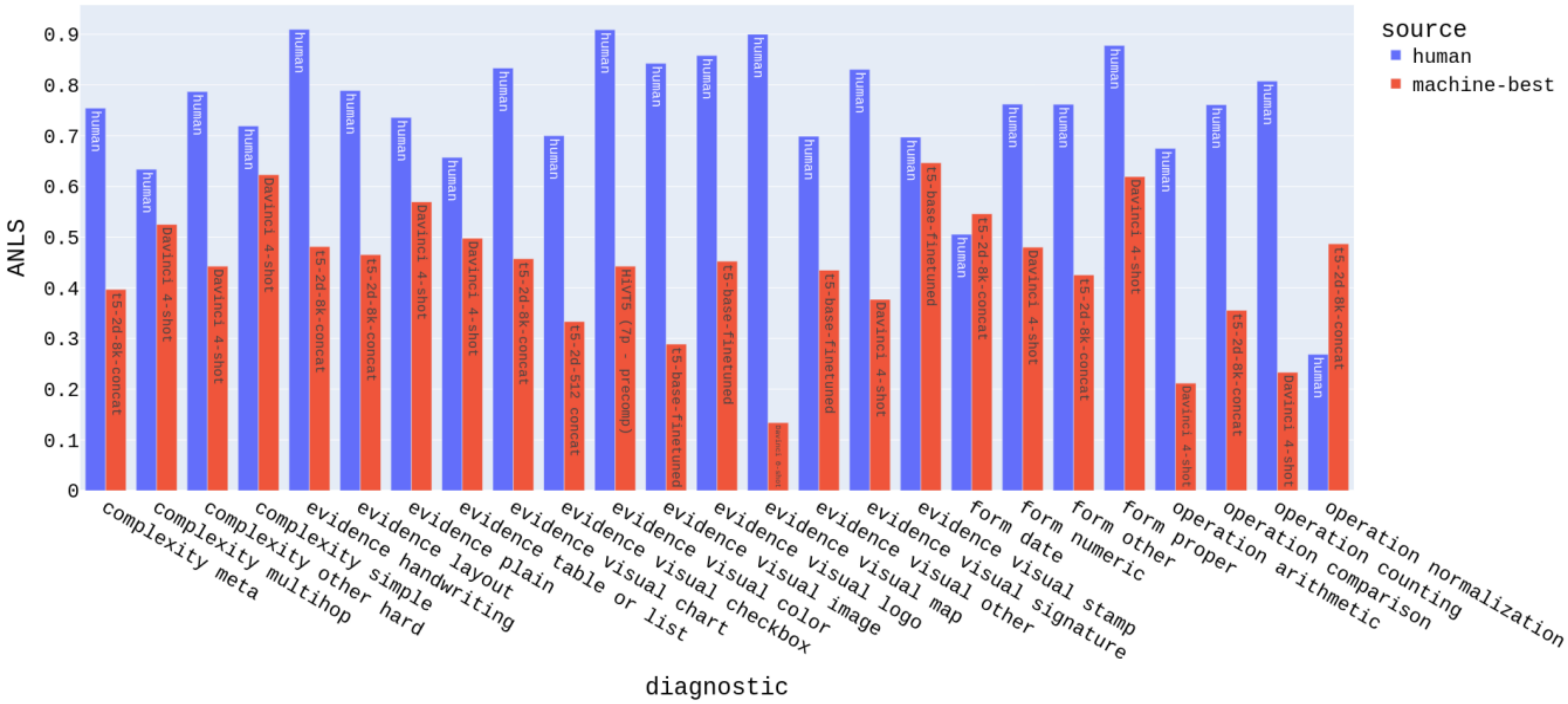
Model	Init.	Params	Max Seq. Length	Test Setup	$ANLS_{all}\uparrow$	$ECE_{all}\downarrow$	$AURC_{all}\downarrow$	$ANLS_{do}$	$ANLS_{do}$ Abs	$ANLS_{do}$ Ex	$ANLS_{do}$ NA	$ANLS_{do}$ Li
text-only Encoder-based models												
Big Bird	MPDocVQA	131M	4096	Concat*	26.27	30.14	44.22	30.67	7.11	40.26	12.75	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

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

I) strong performance of LLMs
II) Even stronger performance by models
+layout understanding
++longer sequence length

ICCV 2023, under review

DUDE Document Understanding Evaluation and Dataset



Diagnostic categories with

- visual evidence
- reasoning operations

Models lagging far behind human baseline

Beyond Document Page Classification

A reality check toward efficient multi-page document representations

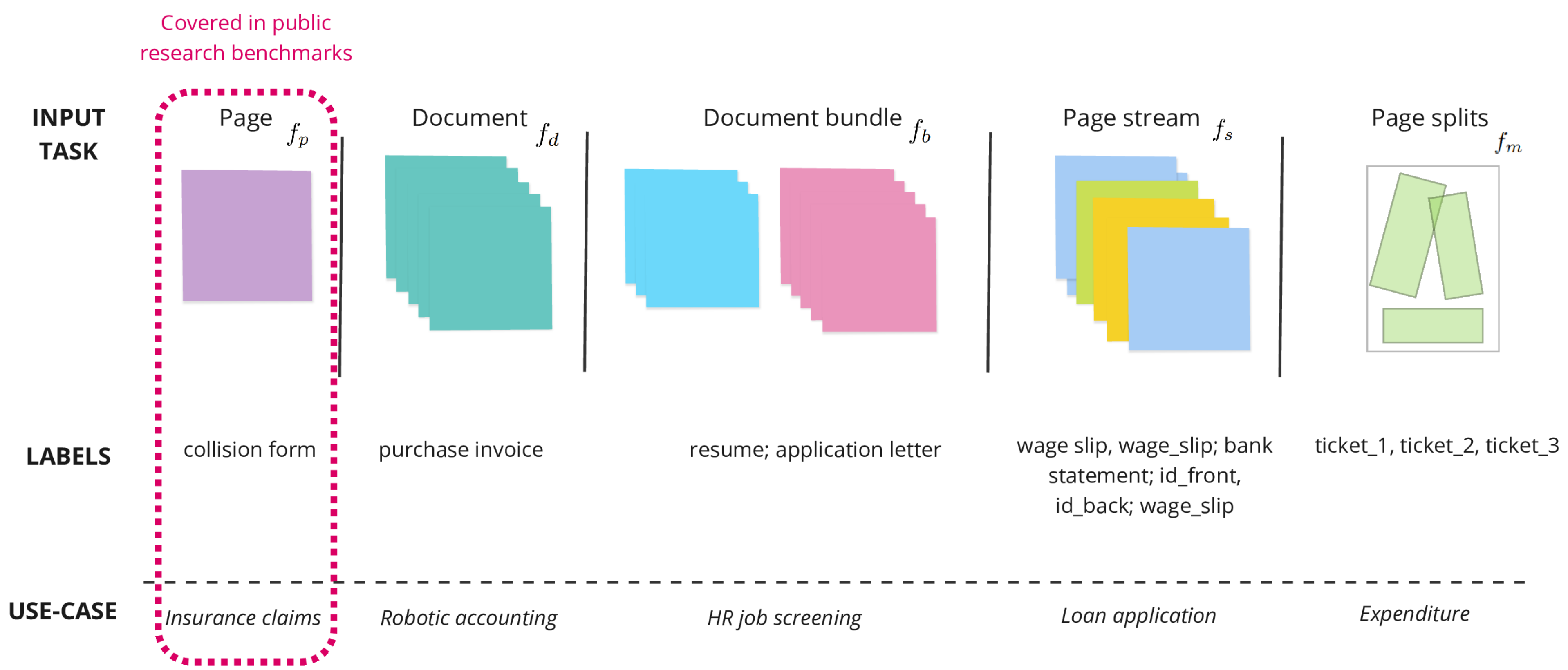
Beyond Document Page Classification (*DocPClf*)

- Position paper with following main points:
  I. Benchmark closer to applied document classification scenarios
  II. Experimental study on multi-page inference methods
  III. Reflect on evaluation practices & moving beyond *iid* test sets

Links to related calls from CVC collaborators (*deepdoc2022*, *scaledoc2023*)

ACIIDS 2023,
under review

I. Document classification scenarios



II. Multi-page inference experiments

Inference	Strategy	Scope
sample	first	page
	second	page
	last	page
sequence	max confidence	page
	soft voting	page
	hard voting	page
grid	grid	document
document	(not tested)	document

strategy	accuracy
single-page [23]	*92.11*
first	91.291
second	87.295
last	85.091
grid	72.642
hard voting	85.995
max confidence	**91.407**
soft voting	91.220
first+second$^{(*)}$	93.795
first+last$^{(*)}$	93.675
second+last$^{(*)}$	89.709
first+second/last$^{(*)}$	***94.454***

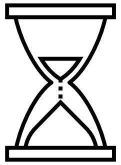
Table 4: Base classification accuracy of DiT-base [23] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_MP` per baseline f_d strategy. The lower table indicated with $^{(*)}$ refers to best-case accuracy when combining 'knowledge' over different pages.

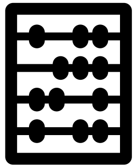
***Hypothesis**: Summary-detail document construction*

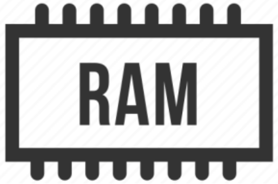
Inefficient (L pages) and dependent on calibration
cf. Table 4 in DUDE ICCV submission (Concat vs. Max Conf.)

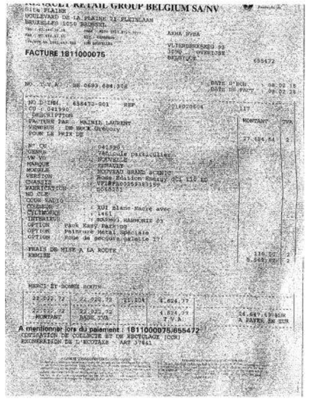
Multi-page document representations are promising for improving document classification

III. Evaluation beyond accuracy and *iid* settings








FACTURE 1811000075

Covariate shifts

Subclass shift


CALIFORNIA
MINI GUIDE

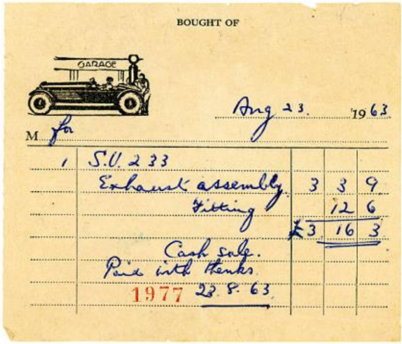
Your essential guide to travelling in California from The World Was Here First

WHAT TO KNOW BEFORE
VISITING CALIFORNIA

California is one of the most visited states in the US and it packs so much into it that it isn't hard to see why. With a population larger than those of Canada and Australia and with the world's fifth-largest economy, California can feel like its own country. With a landscape as diverse as its people and cuisine, the perfect California trip tops every traveller's wish list.

A trip to California is an experience that every traveller should have at least once in their lifetime. If you're planning on visiting the Golden State, follow this advice to get you started.

Near OOD


BOUGHT OF

GARAGE

Aug 23. 1963

M. for

S.U. 233

Exhaust assembly 3 3 9

Fitting 12 6

£3 16 3

Cash sale.

Paid with thanks

1977 23.8.63

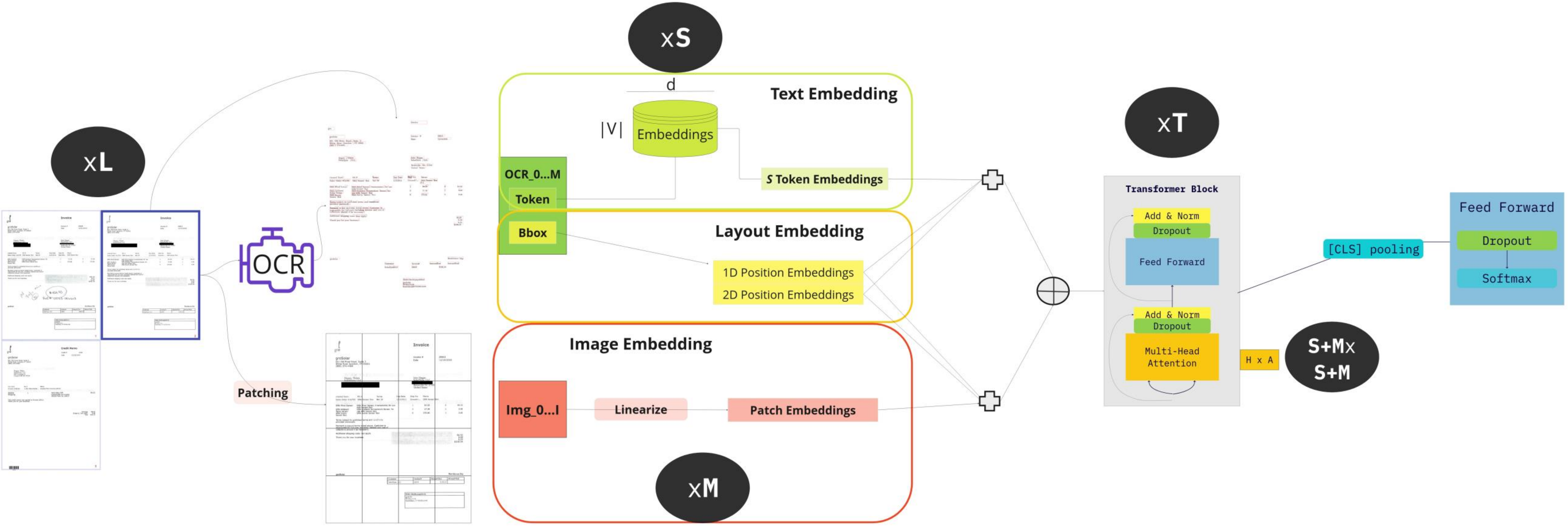
Concept drift

Interesting things to further explore

- All of the above on the introduced DUDE dataset
  - Improving CSF estimation for DocVQA
  - Benchmarking re-calibration methods, calibrated losses and failure prediction
- Multi-task calibration:
  - Hypothesis: joint training with multiple heads will improve joint calibration
  - No opensource dataset with KIE and classification annotations
- Efficient document understanding by e.g., adaptive inference, model compression
- The effect of knowledge distillation from/on calibration and failure prediction

Efficient Multi-Page Document Classification

- *How would we (naively) scale current architectures to multi-page documents and where are the current bottlenecks? (e.g., LayoutLMv3)*



Questions?

+ how to contact me:

jordy-vl.github.io/

Buy me a coffee

KU LEUVEN

References I

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? *arXiv preprint arXiv:2305.18764*, 2023a.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023b.

Glenn W. Brier. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3, 1950.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643): 1512–1519, 2009.

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8: 132330–132347, 2020.

Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating Large-Vocabulary Object Detectors: The Devil is in the Details. *arXiv preprint arXiv:2102.01066*, 2021.

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

References II

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

Li Dong, Chris Quirk, and Mirella Lapata. Confidence Modeling for Neural Semantic Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1069. URL https://aclanthology.org/P18-1069.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H. Abdi. Towards better selective classification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5gDz_yTcst.

Peter A. Flach. *Classifier Calibration*, pages 1–8. Springer US, Boston, MA, 2016. ISBN 978-1-4899-7502-7. doi: 10.1007/978-1-4899-7502-7_900-1. URL https://doi.org/10.1007/978-1-4899-7502-7_900-1.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p66AzKi6Xim.

References III

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

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

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers, 2018.

Zoubin Ghahramani. A history of Bayesian neural networks. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022.

Chirag Gupta and Aaditya K Ramdas. Top-label calibration and multiclass-to-binary reductions. *arXiv preprint arXiv:2107.08353*, 2021.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800*, 2020.

References IV

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995. IEEE, 2015.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *5th International Conference on Learning Representations*, 2017.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *The Journal of Machine Learning Research*, 13(1):2813–2869, 2012.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv preprint arXiv:2204.08387*, 2022.

Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YnkGMIh0gvX.

Abhyuday Jagannatha and Hong Yu. Calibrating Structured Output Predictors for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, 2020.

References V

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. *Advances in neural information processing systems*, 31, 2018.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective Question Answering under Domain Shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, 2020.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated Language Model Fine-Tuning for In- and Out-of-Distribution Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.102. URL https://aclanthology.org/2020.emnlp-main.102.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

References VI

Volodymyr Kuleshov and Percy S Liang. Calibrated structured prediction. *Advances in Neural Information Processing Systems*, 28:3474–3482, 2015.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12316–12326, 2019.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems*, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018.

Fabian Küppers, Anselm Haselhoff, Jan Kronenberger, and Jonas Schneider. Confidence calibration for object detection and segmentation. In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pages 225–250. Springer International Publishing Cham, 2022.

References VII

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30:6402–6413, 2017.

Stefan Larson, Gordon Lim, and Kevin Leach. On Evaluation of Document Classification with RVL-CDIP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *arXiv preprint arXiv:2205.14334*, 2022.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration. *arXiv preprint arXiv:2111.15430*, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems*, 33, 2020.

Xingchen Ma and Matthew B. Blaschko. Meta-cal: well-controlled post-hoc calibration by ranking. *Proceedings of machine learning research (PMLR)*, 2021.

References VIII

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. *arXiv:1902.02476 [cs, stat]*, December 2019. arXiv: 1902.02476 version: 2.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

Muhammad Akhtar Munir, Muhammad Haris Khan, M Saquib Sarfraz, and Mohsen Ali. Towards Improving Calibration in Object Detection Under Domain Shift. In *Advances in Neural Information Processing Systems*, 2022.

Allan H. Murphy and Robert L. Winkler. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34:273–286, 1970.

Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1):41–47, 1977.

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

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

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

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. In *CVPR Workshops*, volume 2, 2019.

Andrew Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you Trust your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019.

References X

Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On Model Calibration for Long-Tailed Object Detection and Instance Segmentation, 2021.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Giovanni Pistone and Carlo Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The annals of statistics*, pages 1543–1561, 1995.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35:7933–7946, 2022.

Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

References XI

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, pages 1–50, 2023.

Ron Slossberg, Oron Anschel, Amir Markovitz, Ron Litman, Aviad Aberdam, Shahar Tsiper, Shai Mazor, Jon Wu, and R Manmatha. On Calibration of Scene-Text Recognition Models. *arXiv preprint arXiv:2012.12643*, 2020.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

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

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 9690–9700. PMLR, 13–18 Jul 2020.

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Benchmarking Scalable Predictive Uncertainty in Text Classification. *IEEE Access*, 10:1–35, 2022. doi: 10.1109/ACCESS.2022.3168734.

KU LEUVEN

References XII

Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Marie-Francine Moens, and Tomasz Stanislawek. Document Understanding Dataset and Evaluation (DUDE), 2023.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Proceedings of the 32th International Conference on Neural Information Processing Systems*, pages 12236–12246. 2019.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021.

Andrew Gordon Wilson. The Case for Bayesian DeepLearning. *arXiv preprint arXiv:2001.10995*, 2020.

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

Shujian Zhang, Chengyue Gong, and Eunsol Choi. Knowing More About Questions Can Help: Improving Calibration in Question Answering. *arXiv preprint arXiv:2106.01494*, 2021.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking Confidence Calibration for Failure Prediction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 518–536. Springer, 2022.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12074–12083, 2023.