KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Intelligent Automation for AI-Driven Document Understanding

Jordy Van Landeghem

Supervisors:
Prof. Dr. Marie-Francine Moens
Prof. Dr. Matthew B. Blaschko

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

March 2024

Intelligent Automation for AI-Driven Document Understanding

Jordy VAN LANDEGHEM

Examination committee:
em. Prof. Dr. ir. Jean-Pierre Celis, chair
Prof. Dr. Marie-Francine Moens, supervisor
Prof. Dr. Matthew B. Blaschko, supervisor
Prof. Dr. ir. Johan Suykens
Prof. Dr. ir. Tinne Tuytelaars
Prof. Dr. Marcus Rohrbach
(TU Darmstadt)
Prof. Dr. Wenpeng Yin
(Penn State University)
Dr. Bertrand Anckaert
(Contract.fit)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

March 2024

Preface

This journey has been long and arduous, but I have finally reached an end. At this end, I have a thesis that I am proud of, and I have learned a lot. As I look back, I have been very fortunate to have had the support of many people, and I would like to take this opportunity to thank them.

First and foremost, I would like to thank my supervisors, **Sien and Matthew**, for their guidance and support throughout this journey. Sien has taught me the importance of being thorough and meticulous, striving for diligence and perfection from the get-go. I still remember how patiently she helped me with my first paper, holding a Sunday afternoon call from her attic/home-office, helping me hone the presentation and writing. Involving Matthew as the co-supervisor has been the best decision for my personal development, as he offered a different perspective on my work, always challenging me to look at problems from the lens of statistical theory and machine learning fundamentals. My knee-jerk reaction to start implementing things as soon as possible was often met with a "slow down, think about it first" from Matthew, which has been invaluable in my development as a researcher. I am grateful to both of them for their patience and understanding, and for giving me the freedom to explore my own ideas and interests.

Next, a sincere thanks to my jury members, for taking the time to read my thesis and for their valuable feedback. Furthermore, I would like to thank het Vlaams Agentschap Innoveren & Ondernemen (VLAIO) for awarding the Baekeland grant without which this PhD would not have been possible.

Pol & Bertrand, thanks for having me contribute to your dream to rid the world of boring administrative processes and paperwork. Technically my bosses, but in reality you are the embodiment of leadership by example, and I am grateful for the many lessons I have learned from you. I am grateful for the many opportunities you have given me to grow as a researcher and as a person. Many thanks to my past and present colleagues at *Contract.fit*, for always

preaching automation, inspiring me, and for having fun along the way. I am grateful to my *LIIR* colleagues at KU Leuven, particularly the folks from office 4.34 for the many interesting discussions and whiteboard sessions, whenever I occasionally popped into the office.

I was fortunate to travel to many places during my PhD (Lausanne, Lisbon, Barcelona, San Jose, Paris, Waikoloa), and I have met many people along the way. My DUDEs, you have been the trigger to complete my PhD, reinvigorating my passion for research and inspiring me for my future career. How crazy is it that we conceived the seeds of the **DUDE** 😎 project in a pirates bar, on a hotel rooftop, and from a hospital bed after my back surgery?

Finally, I would like to thank my family and friends for their support and encouragement throughout this journey. My parents, **Peter en Nadine**, you have showed me that hard work pays off, and *merci* for the many sacrifices you have made to give me the best possible education and life. **Marijke**, you are the love of my life, and although I am not religious, you are my goddess, *de mammiej*. **Feliz**, when you came into our lives, you added an extra dimension. I used to see in 2D, now I see in 3D. Forever your father, your *pappiej*. **Wes en Jen**, thanks for showing me to never give up, keep on pushing, even when you are at your lowest, there is a way out, and only hard work will get you there.

Cornbois -*Bryan, Emile, (even) Jan*, for our friendship, I fail to make an exhaustive definition. I wish for many more years of friendship from my like-minded brothers. *John, Teunen, Wannes*, if there is ever a zombie apocalypse, I know that I can count on you to have my window. **Kessel-city** - *Poohke, Vinny, Kweinch* etc., thanks for keeping on pushing the bar higher, and inspiring me with your ambition and drive. **Gustaf**, thanks for the many laughs (#velleke) and the much-needed distraction. *Elstipoes*, you are my oldest friend, and I am grateful for the many years of friendship. *Woutje*, thanks for your contagious optimism and the mancave during university. **Leuvenbende**, you were the ones that made university fun and enjoyable. Individually and together you are beautiful people, and I cherish our yearly reunions. *Lauren en Yannick*, thanks for letting me win at Mario Kart. I might be forgetting some people, but I would like to thank all my friends for bringing joy, for keeping me grounded, and for reminding me that there is more to life than work.

Having studied literature in my Bachelor's, it feels appropriate to finish with a quote wrongly attributed to Ernest Hemingway: "*Write drunk; edit sober.*"

Jordy Van Landeghem
Gurdo, Pogomeister, Jorre, De Van Laaandeghem
February, 2024
Kessel, Belgium

Abstract

Human communication is increasingly document-based, requiring machines to understand a wide variety of visually-rich documents to assist humans in their daily lives. Amid the digital evolution, documents continue to facilitate crucial human and organizational interactions but are tethered to manual processing, causing inefficiency. We examine why organizations lag in adopting automated document processing solutions and outline two primary challenges: the complexity of processing long, multimodal documents algorithmically and the necessity for reliability and control over associated risks. Automated decision-making is key to improving the efficiency of document processing, but the current state-of-the-art technology is not yet reliable and robust enough to be deployed in autonomous systems.

The practical objective set is to develop Intelligent Automation (IA) systems capable of estimating confidence in their actions, thereby increasing throughput without accruing additional costs due to errors. We analyze the key challenges and propose solutions to bridge the gap between research and practical applications, with a focus on realistic datasets and experimental methodologies. Building upon foundations of Document Understanding (DU), this dissertation introduces advanced methodologies combining Machine Learning, Natural Language Processing, and Computer Vision.

Addressing the evident gaps in research, this work presents novel methods for predictive uncertainty quantification (PUQ) alongside practical frameworks for evaluating the robustness and reliability of DU technologies. The contribution culminates in the introduction of two novel multipage document classification datasets and a multifaceted benchmark, **DUDE** 😎 , designed to rigorously challenge and assess the state-of-the-art in DU. Extensive experiments across these datasets reveal that while advancements have been made, significant room for improvement remains, particularly in long-context modeling for multipage document processing and calibrated, selective document visual question answering. Efficient DU is also explored, revealing the effectiveness

of knowledge distillation (KD) model compression in visually-rich document layout analysis (DLA) and classification.

Through empirical studies and methodological contributions, this dissertation has the following contributions and findings:
First, in a *benchmarking study of established PUQ methods on real-world text classification*, we find that our novel hybrid PUQ method 'Concrete Dropout Ensemble' performs best, enhancing in-domain calibration and novel class detection, even at a smaller ensemble size. Detailed ablation experiments reveal the impact of prior, neural architecture, and hyperparameter choices on PUQ estimation quality.
Second, on a prototypical DU task, we identify challenges in DU progress and propose a *formalization of multipage document classification scenarios*, constructed novel datasets, and conducted an experimental analysis showing the promise of multipage representation learning and inference.
Third, we *introduce* DUDE, *incorporating multifaceted challenges and principles for a comprehensive evaluation of generic DU*. Next to our own benchmarking, we organize a competition, revealing that while newer document foundation models show promise, they struggle with questions involving visual evidence or complex reasoning. Moreover, we find severe problems in the ability of Large Language Models (LLMs) to reason about documents in their entirety, highlighting issues with hallucination, long-context reasoning and control.
Fourth, we *propose the first methodology for enriching documents with semantic layout* structure using distilled DLA models. We apply KD to visual document tasks, unraveling the influence of various task and architecture components.

Finally, the dissertation concludes with a discussion of the findings and implications for future research, emphasizing the need for advancements in multipage document representation learning and the importance of realistic datasets and experimental methodologies to *measurably move forward to reliable and robust IA-DU technology.*

Beknopte samenvatting

Menselijke communicatie is in toenemende mate documentgebaseerd, waarbij machines een breed aanbod aan visueel-rijke documenten moeten begrijpen om mensen in hun dagelijks leven te assisteren. Te midden van de digitale evolutie blijven documenten cruciale menselijke en organisatorische interacties faciliteren, maar zijn ze gebonden aan handmatige verwerking, wat inefficiëntie veroorzaakt. We onderzoeken waarom organisaties achterblijven bij het adopteren van geautomatiseerde documentverwerkingsoplossingen en schetsen twee primaire uitdagingen: de complexiteit van het algoritmisch verwerken van lange, multimodale documenten en de noodzaak van betrouwbaarheid en controle over daarmee samenhangende risico's. Geautomatiseerde besluitvorming is essentieel voor het verbeteren van de efficiëntie van documentverwerking, maar de huidige stand van de technologie is nog niet betrouwbaar en robuust genoeg om ingezet te worden in autonome toepassingen.

Het praktische doel dat gesteld wordt, is het ontwikkelen van systemen voor Intelligente Automatisering (IA) die in staat zijn om vertrouwen in hun acties te schatten, daarmee de doorvoer verhogend zonder extra kosten vanwege fouten. We analyseren de belangrijkste uitdagingen en stellen oplossingen voor om de kloof tussen onderzoek en praktische toepassingen te overbruggen, met een focus op realistische datasets en experimentele methodologieën. Voortbouwend op de fundamenten van Documentinterpretatie (DI), introduceert dit proefschrift geavanceerde methodologieën die Machinaal Leren, Natuurlijke Taalverwerking en Computer Visie combineren.

Door de duidelijke hiaten in onderzoek aan te pakken, presenteert dit werk nieuwe methoden voor predictieve onzekerheidskwantificering (POK) naast praktische kaders voor het evalueren van de robuustheid en betrouwbaarheid van DI-technologieën. De bijdrage culmineert in de introductie van twee nieuwe datasets voor classificatie van multipagina documenten en een veelzijdige benchmark, **DUDE** 😎 , ontworpen om de state-of-the-art in DI rigoureus uit te dagen en te beoordelen. Uitgebreide experimenten met deze datasets

onthullen dat er weliswaar vooruitgang is geboekt, maar dat er nog significant veel ruimte is voor verbetering, met name in de lange-contextmodellering voor de verwerking van multipagina documenten en gekalibreerd, selectief visueel vraagbeantwoording van documenten. Meer schaalbaar DI wordt ook verkend, waarbij de effectiviteit van kennisdistillatie (KD) voor modelcompressie in visueel-rijke layoutanalyse (DLA) en classificatie van documenten aan het licht komt.

Door middel van empirische studies en methodologische bijdragen, heeft dit proefschrift de volgende bijdragen en bevindingen:
Ten eerste vinden we in een benchmarkstudie van gevestigde POK-methoden op tekstclassificatie in de echte wereld dat onze nieuwe hybride POK-methode 'Concrete Dropout Ensemble' het beste presteert, de kalibratie binnenshuis verbeterend en detectie van nieuwe klassen, zelfs met een kleiner ensemble. Gedetailleerde ablatie-experimenten onthullen de impact van voorafgaande kennis, neurale architectuur en keuzes van hyperparameters op de kwaliteit van POK-schatting.
Ten tweede identificeren we uitdagingen in de vooruitgang van DI en stellen een formalisatie voor van multipagina documentclassificatiescenario's, bouwen novel datasets, en voeren een experimentele analyse uit die de belofte van multipagina representatie-leren en inferentie toont.
Ten derde introduceren we DUDE, waarin veelzijdige uitdagingen en principes worden voorgesteld voor een uitgebreide evaluatie. Naast onze eigen benchmarking organiseren we een competitie, waaruit blijkt dat hoewel nieuwere modellen veelbelovend zijn, ze het moeilijk hebben met vragen die visueel bewijs of complex redeneren vereisen. Bovendien vinden we ernstige problemen in het vermogen van Grote Taalmodellen (LLMs) om over documenten in hun geheel te redeneren, wat problemen benadrukt met hallucinatie, redeneren met lange context en controle.
Ten vierde stellen we de eerste experimentele methodologie voor om documenten te verrijken met semantische layoutstructuur met behulp van gedestilleerde DLA-modellen. We passen KD toe op visuele documenttaken, waarbij we de invloed van verschillende architectuurcomponenten van taken ontrafelen.

Ten slotte sluit het proefschrift af met een bespreking van de bevindingen en implicaties voor toekomstig onderzoek, waarbij de noodzaak wordt benadrukt voor vooruitgang in multipagina documentrepresentatie-leren en het belang van realistische datasets en experimentele methodologieën om meetbaar vooruitgang te boeken naar betrouwbare en robuuste IA-DI technologie.

List of Abbreviations

AAPD Arxiv Academic Paper Dataset

Acc_ID Accuracy in-domain

Acc_OOD Accuracy out of domain

AI Artificial Intelligence

ANLS Average Normalized Levenshtein Similarity

AUPR Area Under the Precision-Recall Curve

AURC Area-Under-Risk-Coverage-Curve

AUROC Area Under the Receiver Operating Characteristic curve

BDL Bayesian Deep Learning

BNN Bayesian Neural Network

BPM Business Process Management

CE Cross-Entropy

CER Character Error Rate

COCO Common Objects in Context

CSF Confidence Scoring Function

CV Computer Vision

DC Document Classification

DG Document Generation

DL Deep Learning

DLA Document Layout Analysis

DNN Deep Neural Network

DocAI Document AI

DocVQA Document Visual Question Answering

DOD Document Object Detection

DU Document Understanding

DUDE Document UnderstanDing of Everything

ECE Expected Calibration Error

ELBO Evidence Lower Bound

ERM Empirical Risk Minimization

FasterRCNN Faster Region-based Convolutional Neural Network

FP False Positives

i.i.d. Independent and Identically Distributed

IA Intelligent Automation

ICDAR International Conference on Document Analysis and Recognition

IDP Intelligent Document Processing

IOB/IOBES Inside, Outside, Beginning / End, Single

KD Knowledge Distillation

KIE Key Information Extraction

LLM Large Language Model

MAP Maximum-a-Posteriori

mAP Mean Average Precision

MCD Monte Carlo Dropout

MCMC Markov Chain Monte-Carlo

MDLT Multi-Domain Long-Tailed Recognition

MECE Mutually Exclusive and Collectively Exhaustive

MI Mutual Information

ML Machine Learning

MSE Mean Squared Error

MSP Maximum Softmax Probability

MU Model Uncertainty

NLG Natural Language Generation

NLL Negative Log Likelihood

NLP Natural Language Processing

NN Neural Network

OCR Optical Character Recognition

OOD Out-of-Distribution

PCC Pearson Correlation Coefficient

PUQ Predictive Uncertainty Quantification

RERM Regularized Empirical Risk Minimization

ResNet Residual Network

RPA Robotic Process Automation

SaaS Software-as-a-service

SNGP Spectral-normalized Neural Gaussian Process

SOTA State-of-the-art

STP Straight-Through-Processing

TSR Table Structure Recognition

VDU Visual Document Understanding

VI Variational Inference

VLM Vision Language Model

VQA Visual Question Answering

VRD Visually-Rich Document

WER Word Error Rate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

“
Amid significant life events—*like buying a house or expecting your firstborn child*—lies a less cheerful reality that I experienced firsthand: the hassle of dealing with manual paperwork.

For the former case, this required a lot of back-and-forth with the bank, the notary, and the real estate agent, with each of them requiring a different set of documents (*e.g.*, monthly pay stubs, bank statements, copies of national registry, *etc.*) to be filled in, signed, and sent back for processing.
On the side of the document processors, each document needed to be classified, key information extracted, and the information validated against other documents to be able to prove my solvency in making an offer, applying for a loan, or being drafted as the future house owner. In between all parties and external organizations, even more documents were either created, adapted, or passed along such as the offer, the loan agreement, the deed of sale, a soil certificate, *etc.*

This juxtaposition of valuable moments in life with cumbersome administrative procedures involving **manual document processing** forms the backdrop against which I aim to explore and propose potential solutions in this thesis.

”

Documents are containers of information that are easily shareable. The concept of a document dates back to when humans started writing and has been a cornerstone of human communication ever since. In the age of digital technology, documents are still the primary means of communication between humans and organizations and form the backbone of many business processes. Human communication is increasingly happening through digital channels, and the COVID-19 pandemic has only accelerated this trend. We are increasingly living in a "document society" [53], dependent on documents in our daily lives or for recording second-hand knowledge. With instant gratification as the norm in the digital age, people expect similar seamless interactions with businesses and governments. While digitization has increased the speed and ease of document-based communication, document processing remains a largely human effort with organizations drowning under the sheer volume of documents they receive.

So why have organizations not switched en masse to automated document processing?

The answer lies for some part in (I) **the complexity of the task**, and for the other part in (II) **the need for reliability and risk control**.

(I) While it might be straightforward for a human (white-collar) worker to read a long, structured document, understand its contents, categorize it, and extract crucial information accordingly, this is not so easy for a machine. This could be perceived as an instance of Moravec's paradox [319], which states that tasks that are easy for humans are hard for machines, and vice versa. However, in recent times, significant strides forward have been made thanks to technological advances combining Natural Language Processing (NLP), Computer Vision (CV) and Machine Learning (ML). **Document Understanding** (DU) is the umbrella term for both the end-to-end solution and the research field studying to make machines interpret and understand documents (elaborated on in Section 2.3). It has seen a surge in interest in the past few years, with the rise of large-scale pretrained Language and Vision models (LLM, VLM) [52, 94, 101, 187, 380, 383, 502] capable of modeling document inputs.

What makes DU challenging is that it encompasses multiple subtasks, each of which is a research field in its own right, such as Optical Character Recognition (OCR), Document Layout Analysis (DLA), Document Classification (DC), Key Information Extraction (KIE), Visual Question Answering (VQA), *etc.* The complexity of the task is further increased by the fact that documents are multimodal, containing both text and images and that they are compositional, *i.e.*, the meaning of the document is not just the sum of its parts. Information can appear in a wide range of forms including text, images, tables or graphs, and be spread across multiple pages. Moreover, the meaning of a document

can change depending on the context in which it is used. As an artifact of the communication channel, not all documents are born digitally, and the quality of the document can vary greatly, with some documents being handwritten, scanned with low resolution, or even a picture of a document. Furthermore, documents are often not standardized templates and can be highly variable in terms of layout, structure, and content. Finally, the longer the document, the more computationally demanding it becomes to process, and the more likely it is to induce errors, which can be harder to detect.

Addressing the inherent challenges of document processing, and achieving high levels of accuracy, processing speed, reliability, robustness, and scalability in DU forms the applied scope of this thesis.

(II) Consider the example given of the birth certificate. While I might not appreciate as much the manual handling of this document, if they had registered my baby girl's name (*Feliz*, Spanish writing without an accent on the 'e') incorrectly, I would be pretty upset as this could have further repercussions. Whereas this error might be easily rectified, it is not so easy to do so in the case of a mortgage application, where the wrong information could lead to a rejection of the application, or even worse, a loan agreement with the wrong terms and conditions. This demonstrates that, even when full automation of document processing is in high demand, it is not always desirable if the risk of failure might be too large.
Nevertheless, a lot of the potential for automation remains untapped, and organizations are increasingly looking for solutions to fully automate their document processing workflows. However, full automation, implying perfect recognition of document categories and impeccable information extraction is an unattainable goal with the current state of technology [79].

The more realistic objective set is **Intelligent Automation** (IA) (elaborated on in Section 2.4), where the goal is to have the machine estimate confidence in its predictions, deriving business value with as high as possible volumes of perfect predictions (Straight-Through-Processing, STP) without incurring extra costs (False Positives, FP).
The leitmotif of this thesis will be the fundamental enablers of IA: confidence estimation and failure prediction.

Calibrated uncertainty estimation with efficient and effective DU technology will allow organizations to confidently automate their document processing workflow, while keeping a human in the loop only for predictions with a higher likelihood of being wrong. To date, however, little research has addressed the question of how to make DU technology more reliable, as is illustrated in a toy analysis (Table 1.1) reporting the absence of many IA-related keywords in the Proceedings of the 2021 International Conference on Document Analysis and

Recognition (ICDAR) [289].
The thesis aims to fill this gap by proposing novel methods for uncertainty estimation and failure prediction (Part I), and by providing a framework for benchmarking and evaluating the reliability and robustness of DU technology, as close as possible to real-world requirements (Part II).

Table 1.1. Comparative analysis of keywords in the ICDAR 2021 proceedings. While many DU subtasks are represented, there is a lack of keywords related to IA. Do note that calibration is used in the context of camera calibration, and not in the context of confidence estimation.

keyword	freq	keyword	freq
document	3388	calibration/calibrate	33
classification	242	temperature scaling	0
key information	56	failure prediction misclassification detection	0
question answering	106	out-of-distribution OOD	25
layout analysis	223	predictive uncertainty	0

In the remainder of the Introduction, I will sketch the surrounding research context, followed by the problem statement and research questions, and finally the outline of the thesis manuscript.

1.1 Research Context

All chapters of this dissertation have been executed as part of the Baekeland PhD mandate (HBC.2019.2604) with financial support of VLAIO (Flemish Innovation & Entrepreneurship) and Contract.fit. The latter is a Belgian-based software-as-a-service (SaaS) provider of Intelligent Document Processing (IDP) drawing on innovations in DU to power their product suite (email-routing, Parble), and my generous employer since 2017.

Some of the joint work (Chapter 5) has been partially funded by a PhD Scholarship from AGAUR (2023 FI-3-00223), and the Smart Growth Operational Programme under projects no. POIR.01.01.01-00-1624/20 (*Hiper-OCR - an innovative solution for information extraction from scanned documents*) and POIR.01.01.01-00-0605/19 (*Disruptive adoption of Neural Language Modelling for automation of text-intensive work*).

Moreover, given that the dissertation work has been performed over a large span of time, it warrants putting it in the larger context and dynamics of AI innovations, the state of DU as a field, how notions of 'reliability' have evolved over time, and finally the business context.

This thesis started almost concurrently with the rise of the global COVID-19 pandemic, making it hard to foster collaborations in the early stages. At the start of the PhD, DU methodology was fairly established, with OCR and Transformer-based pipelines such as BERT [94] and LayoutLM [502], which is why we first prioritized the more fundamental challenge of decision-making under uncertainty (Part I); which was followed by a step back, closer to applied DU research (Part II).

The research community's understanding of 'reliability' has also evolved over time. When starting the work of Chapter 3, the notion of reliability was mostly associated with uncertainty quantification and calibration. However, calibration is not a panacea, and only fairly recently, Jaeger et al. [193] proposed a more general framework encapsulating reliability and robustness. They promote the more concrete and useful notion of **failure prediction**, which still involves confidence/uncertainty estimation yet with an explicit definition of the failure source which one wants to detect or guard against, *e.g.*, in-domain test errors, changing input feature distributions, novel class shifts, *etc.* Since I share a similar view of the problem, I have focused following works on the more general notion of failure prediction, which is also more in line with the business context of IA.

Whereas we originally intended to work on multi-task learning of DU subtasks, the rise of general-purpose LLMs offering a natural language interface to documents rather than discriminative modeling (*e.g.*, ChatGPT [52, 344]), prompted us toward evaluating this promising technology in the context of DU. More importantly, we observed the lack of sufficiently complex datasets and benchmarks in DU that would allow us to tackle larger, more fundamental questions such as 'Do text-only LLMs suffice for most low-level DU subtasks?' (subsequently tackled in Chapter 5), which is why we shifted our focus to the more applied research questions of benchmarking and evaluation (Part II).

Finally, the business context has also evolved over time. Originally, IDP was practiced by legacy OCR companies; specialized vendors, offering a range of solutions for specific document types (*e.g.*, invoices, contracts, tax forms, *etc.*); or cloud service providers, offering IDP as part of a larger suite of services (*e.g.*, AWS Textract, Azure Form Recognizer, *etc.*). However, the rise of both open-source LLM development and powerful, though closed-source models has lowered the barrier to entry for any new entrants or incumbents. This has led to a commoditization of IDP, with the quality of the LLMs and the ease of integration with existing business processes becoming key differentiators.

1.2 Problem Statement and Questions

The general introduction sketches the context of the research, and motivates the research questions. In this Section, I will formulate the problem statement and research questions more formally and how they relate to the manuscript's contents.

1.2.1 Reliable and Robust Deep Learning

The dissertation opens with the more fundamental challenge of targeting *reliability and robustness in Deep Learning*, which covers fairly abstract concepts that have been used interchangeably and inconsistently in the literature. They will be defined more extensively in Section 2.2, but for now, consider reliability as the ability to avoid failure, robustness as the ability to resist failure, and resilience as the ability to recover from failure [373, 438, 455]. In Chapter 3, we focus on the more concrete objective of predictive uncertainty quantification (PUQ), which shows promise for improving reliability and robustness in Deep Learning (DL) [123, 140, 173, 455]. Concretely, PUQ methods are expected to elucidate sources of uncertainty such as a model's lack of in-domain knowledge due to either training data scarcity or model misspecification, or its ability to flag potentially noisy, shifted or unknown input data [136].

We observed that the majority of prior PUQ research focused on regression and CV tasks, while the applicability of PUQ methods had not been thoroughly explored in the context of NLP. As mentioned earlier, most DU pipelines (in 2020) were text-centric with a high dependency on the quality of OCR. Since OCR is often considered a solved problem [262], we hypothesized that the main source of error and uncertainty in DU would reside in the text representations learned by deep neural networks (DNN)s. This is why we focused on the more fundamental question of *how well do PUQ methods scale in NLP?* More specifically, we restricted the scope to the prototypical, well-studied task of text classification, for which we could leverage existing multi-domain datasets varying in complexity, size and label space (multi-class vs. multi-label).

This leads to the following research questions:

RQ 1. When tested in realistic language data distributions on various text classification tasks, how well do PUQ methods fare in NLP?

RQ 2. In which settings are PUQ methods most useful, *i.e.*, which failure sources / distribution shifts are they most sensitive to?

RQ 3. How can we obtain better PUQ estimates without overrelying on computationally prohibitive methods, *e.g.*, Deep Ensemble [238]?

RQ 4. How important are certain prior, neural architecture or hyperparameter influences on the quality of PUQ estimation?

In a later chapter (Chapter 5), we introduce a complex benchmark for generic DU that additionally tests for robustness to domain, visual and layout shifts, and explores the novel problem of hallucination and control in natural language generation (NLG) with LLMs from the perspective of *calibrated and selective* DocVQA. The general task formulation involves a natural language question (on content, aspect, form, visual/layout), an input document, and a set of reference answers. The model is expected to provide a natural language answer, an answer confidence and a (binary) abstention decision. Evaluation is done in terms of answer correctness, calibration and selective prediction. On the one hand, one expects a model to lower confidence when unsure about the correctness of a predicted answer. On the other hand, one expects a model to abstain from answering and refrain from hallucinations on unanswerable questions (which had been explicitly added in the dataset).

RQ 5. How severe is the problem of hallucination and control in LLMs when evaluated in a selective, free-form DocVQA task setting?

1.2.2 Realistic and Efficient Document Understanding

The second part of the dissertation focuses on the more applied research questions of *realistic and efficient* DU. The overall objective is to make DU technology more generically applicable (Chapter 5), evaluation more in sync with real-world requirements (Chapters 4 and 5), and more efficient at modeling the multimodal and compositional nature of documents (Chapters 5 and 6).

Due to the proximity to business applications and the risks of leaking personal information, DU research benchmarks have diverged substantially from the real-world distributions of document data. For instance, DU datasets are often limited to single-page document images, are from outdated sources (*e.g.*, IIT-

CDIP [252]), or are restricted to a single domain or a small set of document types.

We posit that larger, fundamental questions in DU remain unanswered due to a lack of sufficiently complex datasets and benchmarks with a rich methodology covering evaluation beyond the independent and identically distributed (i.i.d.) test set setting. While there exist performant models for DU subtasks such as OCR, DC, KIE, *etc.*, it is unclear how to move from these specific analysis and recognition tasks to models that can reason and understand documents. A truly end-to-end DU solution must handle the complexity and variety of real-world documents and subtasks, which could be expressed as natural language questions. Moreover, it should be able to generalize to any question on any document and reason over multiple pages and modalities.

The following research questions are addressed in Chapters 4 and 5:

RQ 6. How can we iteratively close the gap between research and practice in DU?

RQ 7. How can we design a resource that comprehensively challenges the state-of-the-art?

RQ 8. Which DU aspects are most challenging for current state-of-the-art LLMs? How can these be incorporated in a benchmark to allow proper measurements of future improvements?

However, moving the goalpost beyond a single-page context inevitably requires us to reconsider the research challenge of efficiency in DU. The rise of LLMs has enabled a new generation of DU pipelines, which are more flexible and easier to maintain than separate and specialized subtask modules, but also more computationally demanding. Importantly, most LLMs are not designed to handle the multimodality and long context windows of multipage documents, and are often unaware of the visual and layout semantics of documents.

The research questions for Chapter 6 address the *efficiency* challenge in DU:

RQ 9. How can we efficiently infuse LLMs with semantic layout awareness for more focused information extraction?

RQ 10. To what degree can model compression resolve the problem of efficiency in processing documents?

1.3 Outline

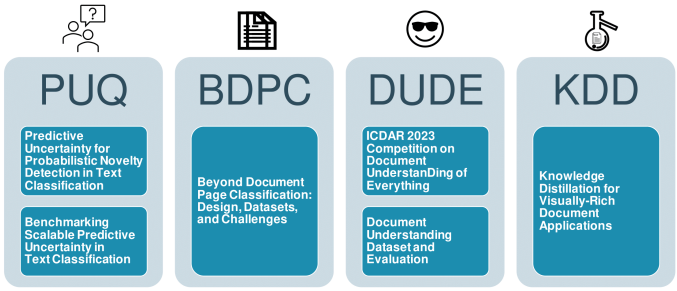


Figure 1.1. Overview of publications and how they relate to the chapters.

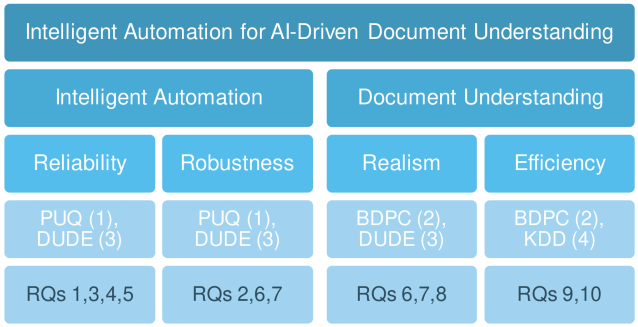


Figure 1.2. Visual Overview of the research questions and how they relate to the chapters.

After the introductory Chapters 1 and 2, we continue with the publication-based chapters that form the core of the thesis, which are structured in two parts.

Part I consists of a single chapter, Chapter 3, which presents a benchmarking study of PUQ methods applied on real-world text classification datasets with 1-D convolutional neural networks and pretrained transformers. It motivates a novel PUQ method, *Deep Ensemble* with *Concrete Dropout*, combining the benefits of both methods, and showing promise for improving reliability and robustness in NLP at a lower computational cost. The chapter concludes with a discussion of the results, including targeted ablation studies, and provides recommendations for future research.

Part II consists of three chapters, Chapters 4 to 6, which all focus on the more applied research questions of *realistic and efficient* DU.

Chapter 4 reflects on the current state of DU research, and proposes guidelines to foster document dataset construction efforts. It introduces two novel document classification datasets, RVL-CDIP_MP and RVL-CDIP-N_MP, as extensions of the RVL-CDIP dataset [165] with multipage documents. The datasets are accompanied by a comprehensive experimental analysis, which shows promise from advancing multipage document representations and inference.

Chapter 5 introduces the multi-faceted **DUDE** 😎 benchmark for assessing generic DU, that was also hosted as a competition to challenge the DU community. It describes the complete methodology and design of the dataset, targeting model innovations that can handle the complexity and variety of real-world documents and subtasks, and generalize to any documents and any questions. Next to a discussion of the competition results, it also presents our own comprehensive benchmarking study of SOTA LLMs with varying the context length and what modalities are represented.

Chapter 6 investigates how to efficiently obtain more semantic document layout awareness. We explore what affects the teacher-student knowledge gap in KD-based model compression methods, and design a downstream task setup to evaluate the robustness of distilled DLA models on zero-shot layout-aware DocVQA.

Finally, Chapter 7 concludes the thesis with a summary of the main contributions (Section 7.1), and a discussion of future research directions. As a logical follow-up to Chapter 5, we propose in Section 7.2.2.1 how the DUDE dataset could be extended to become the 'ultimate' DU benchmark. The thesis ends with a hypothetical, informed design of how the research presented would form part of an end-to-end, fully-fledged IA-DU solution (Section 7.2.2.2).

Chapter 2

Fundamentals

This chapter provides all the necessary background knowledge necessary to understand the contributions of this thesis.

The key questions covered here are:

i. *How to feed a document to an algorithm to perform arbitrary tasks on it?*
ii. *How to model language, vision, layout or structure?*
iii. *How does it learn and then operate at inference time?*
iv. *How does it estimate prediction uncertainty?*
v. *How to evaluate its performance?*
vi. *How to integrate it as a useful, end-to-end system in a document workflow?*

Section 2.1 explains the basic setting from the perspective of statistical learning theory [472], which is a mathematical framework for analyzing how algorithms learn from data with minimal error. Section 2.2 gives a primer on reliability and robustness, particularly calibration, failure detection and relevant evaluation metrics. Section 2.3 surveys the DU field, and discusses the state of the art in DU technology. Finally, Section 2.4 covers Intelligent Automation to illustrate how solving the challenges posed in this thesis will enable to augment human intelligence, creativity and productivity in straight-through business processes.

2.1 Statistical Learning

Two popular definitions of Machine Learning (ML) are given below.

> *Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.* [406]

> *A computer program is said to learn from experience* E *with respect to some class of tasks* T*, and performance measure* P*, if its performance at tasks in* T*, as measured by* P*, improves with experience* E. [317]

Following these, different types of learning problems [472] can be discerned, of which the most common (and the one used throughout our works) is **supervised learning**. It defines experience E as a set of input-output pairs for which the task T is to learn a mapping f from inputs $X \in \mathcal{X}$ to outputs $Y \in \mathcal{Y}$, and the performance measure P is the **risk** or expected loss (Equation (2.1)), given a (0-1) loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{P}}[\ell(Y, f(X))] \tag{2.1}$$

The mapping $f(\cdot;\theta) : \mathcal{X} \to \mathcal{Y}$ is typically parameterized by a set of parameters θ (omitted whenever it is fixed) and a hypothesis class $\mathcal{F}$, which is a set of possible functions. The objective is to find a function $f \in \mathcal{F}$ that minimizes the risk, or even better, the **Bayes risk**

$$f^* = \inf_{f\in\mathcal{F}} \mathcal{R}(f), \tag{2.2}$$

which is the minimum achievable risk over all functions in $\mathcal{F}$. The latter is only realizable with infinite data or having access to the data-generating distribution

$\mathcal{P}(\mathcal{X}, \mathcal{Y})$. In practice, Equation (2.2) is unknown, and the goal is to find a function $\hat{f}$ that minimizes the **empirical risk**

$$\hat{f} = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(x_i)), \tag{2.3}$$

where (x_i, y_i) are N independently and identically distributed (i.i.d.) samples drawn from an unknown distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$. This is known as **empirical risk minimization** (ERM), which is a popular approach to supervised learning, under which three important processes are defined.

Training or model fitting is the process of estimating the parameters θ of a model, which is done by minimizing a suitable *loss function* ℓ over a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of N i.i.d. samples.

Inference or prediction is the process of estimating the output of a model for a given input, which is typically done by computing the posterior probability $P(y|x)$ over the output space $\mathcal{Y}$. Classification output is a discrete label, while regression output is a continuous value.

Evaluation involves measuring the quality of a model's predictions, which is typically done by computing a suitable *evaluation metric* over a test set $\mathcal{D}_{\text{test}}$ of i.i.d. samples, which were not used for training.

However, ERM has its caveats concerning **generalization** to unseen data, requiring either additional assumptions on the hypothesis class $\mathcal{F}$, which are known as **inductive biases**, and/or **regularization** to penalize the complexity of the function class $\mathcal{F}$ [445]. In neural networks (discussed in detail Section 2.1.1), the former is controlled by the architecture of the network, while the latter involves specifying constraints to parameters or adding a regularization term to the loss function.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left(\hat{\mathcal{R}}(f) + \lambda \Psi(\theta) \right) \tag{2.4}$$

Equation (2.4) defines **regularized empirical risk minimization** (RERM), where $\Psi(\theta)$ is a regularization term and λ is a hyperparameter that controls the trade-off between the empirical risk (denoted with $\hat{R}$) and the regularization term.

All these concepts will be revisited in the context of neural networks in Section 2.1.1, where we will also discuss the optimization process of the model parameters θ, how inference differs in the case of probabilistic models to estimate

uncertainty (Section 2.2.5), and how regularization affects confidence estimation and calibration (Section 2.2.4).

2.1.1 Neural Networks

An artificial **neural network** (NN) is a mathematical approximation inspired by data processing in the human brain [396]. It can be represented by a network topology of interconnected **neurons** that are organized in **layers** that successively refine intermediately learned feature representations of the input [448] that are useful for the task at hand, *e.g.*, classifying an animal by means of its size, shape and fur, or detecting the sentiment of a review by focusing on adjectives.

A basic NN building block is a **linear layer**, which is a linear function of the input parameters: $f(x) = Wx + b$, where the bias term b is a constant vector shifting the decision boundary away from the origin and the weight matrix W holds most parameters that rotate the decision boundary in input space. **Activation functions** (*e.g.*, tanh, ReLu, sigmoid, softmax, GeLu) are used to introduce non-linearity in the model, which is required for learning complex functions.

The first **deep learning** (DL) network (stacking multiple linear layers) dates back to 1965 [191], yet the term 'Deep Learning' was coined in 1986 [398]. The first successful DL application was a demonstration of digit recognition in 1998 [244], followed by DL for CV [90, 223] and NLP [76]. The recent success of DL is attributed to the availability of large datasets, the increase in computational power, the development of new algorithms and architectures, and the commercial interest of large companies.

Consider a conventional DL **architecture** as a composition of parameterized functions. Each consists of a configuration of layers (*e.g.*, convolution, pooling, activation function, normalization, embeddings) determining the type of input transformation (*e.g.*, convolutional, recurrent, attention) with (trainable) parameters linear/non-linear w.r.t. the input x. Given the type of input, *e.g.*, language which is naturally discrete-sequential, or vision which presents a ready continuous-spatial signal, different DL architectures have been established, which will be discussed in Section 2.1.3.

A K-class classification function with an l-layer NN with d dimensional input $x \in \mathbb{R}^d$ is shorthand $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$, with $\theta = \{\theta_j\}_{j=1}^{l}$ assumed to be optimized, either partially or fully, using *backpropagation* and a *loss function*. More specifically, it presents a **non-convex optimization** problem, concerning multiple feasible regions with multiple locally optimal points within each. With **maximum-**

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

Table 2.1. Sigmoid and softmax activation functions for binary and multi-class classification, respectively.

likelihood estimation estimation, the goal is to find the optimal parameters or weights that minimize the loss function, effectively interpolating the training data. This process involves traversing the high-dimensional loss landscape. Upon convergence of model training, the optimized parameters form a *solution* in the weight-space, representing a unique *mode* (specific function $f_{\hat{\theta}}$). However, when regularization techniques such as weight decay, dropout, or early stopping are applied, the objective shifts towards **maximum-a-posteriori** (MAP), to take into account the prior probability of the parameters. The difference in parameter estimation forms the basis for several uncertainty estimation methods, covered in Section 2.2.5.

A *prediction* is a translation of a model's output to which a standard decision rule is applied, *e.g.*, to obtain the top-1/k prediction (Equation (2.5)), or decode structured output according to a function maximizing total likelihood with optionally additional diversity criteria.

$$\hat{y} = \text{argmax}\, f_{\hat{\theta}}(x) \tag{2.5}$$

Considering standard NNs, the last layer outputs a vector of real-valued **logits** $\mathbf{z} \in \mathbb{R}^K$, which in turn are normalized to a probability distribution over K classes using a **sigmoid** or **softmax** function (Table 2.1).

2.1.2 Probabilistic Evaluation

The majority of our works involves supervised learning with NNs, formulated generically as a probabilistic predictor in Definition 1.

Definition 1. *Probabilistic predictor $f : \mathcal{X} \to \Delta^{\mathcal{Y}}$ that outputs a conditional probability distribution $P(y'|x)$ over outputs $y' \in \mathcal{Y}$ for an i.i.d. drawn sample (x,y).*

Definition 2 (Probability Simplex)**.** *Let $\Delta^{\mathcal{Y}} := \{v \in \mathbb{R}^{|\mathcal{Y}|}_{\geq 0} : \|v\|_1 = 1\}$ be a probability simplex of size $|\mathcal{Y}| - 1$ as a geometric representation of a probability space, where each vertex represents a mutually exclusive label and each point has an associated probability vector v [368].*

Figure 2.1 illustrates a multi-class classifier, where $\mathcal{Y} = [K]$ for K=3 classes.

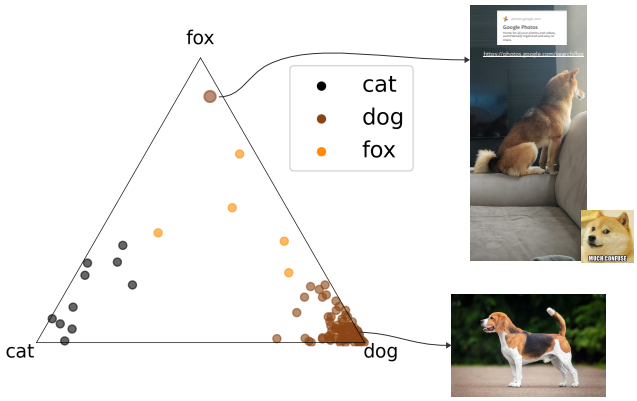


Figure 2.1. Scatter plot of a ternary problem ($K = 3, N = 100$) in the probability simplex space. Example of overconfident misprediction (above is a Shiba Inu dog) and correct sharp prediction (clear image of Beagle).

In practice, **loss functions** are proper scoring rules [330], $S : \Delta^{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}$, that measure the quality of a probabilistic prediction $P(\hat{y}|x)$ given the true label y. The **cross-entropy** (CE) loss is a popular loss function for classification, while the **mean-squared error** (MSE) loss is used for regression. In Section 2.2, we will discuss the evaluation of probabilistic predictors in more detail, including the calibration of confidence estimates and the detection of out-of-distribution samples.

2.1.3 Architectures

Throughout the chapters of the thesis, we have primarily used the following NN architectures: **Convolutional Neural Networks** (CNNs), **Transformer Networks** . We will briefly introduce the building blocks of these architectures, with a focus on how they are used in the context of document understanding.

2.1.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [244] are a class of DNNs designed primarily for visual and grid-spatial data such as images. They are inspired by the visual cortex of animals, which contains neurons that are sensitive to small subregions of the visual field, called a **receptive field**. The receptive fields of

different neurons partially overlap such that they cover the entire visual field, growing larger in deeper layers of the visual cortex.

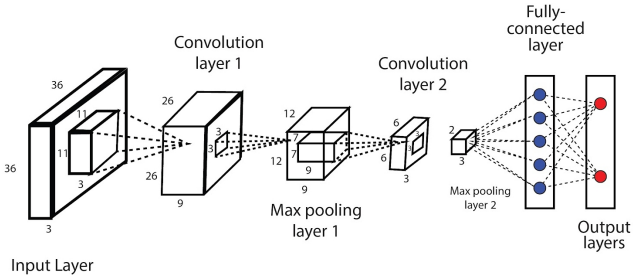


Figure 2.2. Sketch of a CNN architecture. The input is a 2D image, which is iteratively convolved with a set of learned **filters** detecting specific input features, *e.g.*, edges, corners, blobs, to produce **feature maps**. Feature maps are then downsampled using a **pooling** operation.

As illustrated in Figure 2.2, CNNs are composed of multiple convolutional layers, which hierarchically extract features from the input, followed by pooling and fully-connected layers to classify the input based on the downsampled features. A **filter** $\mathcal{K} \in \mathbb{R}^{d \times d}$ is a rectangular matrix of trainable weights with width and height d typically smaller than the input x. A convolutional layer applies filters sliding over the input, with each filter producing a feature map:

$$\mathcal{F} = \mathcal{K} * x, \tag{2.6}$$

where the convolution operation $*$ computes a dot product between filter entries and the covered portions of the input.

Thanks to the weight sharing property of the convolution operation, CNNs are able to learn **translation invariance**, *i.e.*, the ability to recognize an object regardless of its position in the image. This is particularly useful for object detection, where the position of the object in the image is unknown.

This architecture was used for document image classification and document layout analysis (Section 6.3.2). A special version is *1-D CNNs*, which we applied to one-hot encoded text data in text classification benchmarking (Section 3.4.3).

2.1.3.2 Language Neural Networks

The first step to represent language input into a format compatible with NNs is to convert units of language, words or characters or "tokens" as depending on

a **tokenizer**, into numerical vectors. This is done by means of **embeddings**, which are typically learned as part of the training process, and are used to represent the meaning of words in a continuous vector space. There have been multiple generations of word embeddings, starting with **one-hot** vectors that represent each word by a vector of zeros with a single one at its vocabulary index, which depends highly on the tokenizer used and does not capture semantic relationships between words. Alternatives are **frequency**-based embeddings, such as **TF-IDF** vectors, which represent each word by its frequency in the corpus, weighted by its inverse frequency in the corpus, capturing some lexical semantics, but not the context in which the word appears. The next generation are **Word2Vec** embeddings that are trained to predict the context of a word, *i.e.*, the words that appear before and after it in a sentence. **FastText** embeddings improve this by considering a character n-gram context, *i.e.*, a sequence of n characters. The current generation are **contextual word embeddings** that are trained to predict the context of a word, taking into account the surrounding context and learning the sense of a word based on its context, *e.g.*, 'bank' as a river bank vs. a financial institution in '*Feliz sits at the bank of the river Nete*'. Another important innovation is **subword tokenization** to deal with the **out-of-vocabulary** (OOV) problem, which is particularly relevant for morphologically rich languages, such as Dutch, where word meaning can be inferred from its subwords. A clever extension is **byte pair encoding** (BPE) [412], which is a data compression algorithm that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte, until a predefined vocabulary size is reached. This is particularly useful for **multilingual** models, where the vocabulary size would otherwise be too large to fit in memory.

The first embedding layer is typically a **lookup table**, which maps each word to a unique index in a vocabulary, and each index to a vector of real numbers. The embedding layer is typically followed by a **recurrent**, **convolutional** or **attention** layer, which is used to capture the sequential nature of language. **Recurrent Neural Networks** (RNNs) and recurrent architectures extended to model long-range dependencies such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were the dominant architectures for sequence modeling in NLP, yet they have been superseded by Transformers in recent years.

2.1.3.3 Transformer Network

A **Transformer** [473] is a *sequence-to-sequence* model that uses an attention mechanism to capture long-range dependencies in the input sequence, benefiting from increased parallelization. Traditionally, it consists of an encoder and a

decoder, each composed of multiple layers of self-attention and feed-forward layers.

Attention is a mechanism that allows for soft selection of relevant information from a set of candidates, *e.g.*, tokens in a document, based on a query, *e.g.*, a token in the document. The **scaled dot-product attention** is defined for a sequence of length n as follows: $\text{Att}(Q, K, V) = \sum_{i=1}^{n} \alpha_i V_i$. It utilizes three learnable weight matrices, each multiplied with all token embeddings in a sequence to build queries $Q \in \mathbb{R}^{n \times d_q}$, keys $K \in \mathbb{R}^{n \times d_q}$, and values $V \in \mathbb{R}^{n \times d_v}$. The output of the attention mechanism is a weighted sum of the unnormalized values, where each attention weight of the i-th key is computed by normalizing the dot product between the query and key vectors $\alpha_i = \frac{\exp(Q_i^T K_i)}{\sum_{j=1}^{n} \exp(Q_J^T K_j)}$. For training stability, the dot product is typically scaled by the square root of the dimensionality of the query and key vectors. This is followed by a feed-forward layer to capture non-linear relationships between the tokens in the sequence.

There exist different forms of attention, depending on the type of relationship that is captured. **Self-attention** computes the attention of each token w.r.t. all other tokens in the sequence, which changes the representation of each token based on the other tokens in the sequence. **Multi-head attention** is a set of h attention layers, which every Transformer uses to concurrently capture different types of relationships, concatenated together after the parallelized processing. **Cross-attention** computes the attention of each token in one sequence w.r.t. all tokens in *another sequence*, which is used in encoder-decoder Transformer architectures for *e.g.*, summarization and machine translation. Specific to decoder layers, **masked attention** is used to prevent the decoder from attending to future tokens in the sequence by masking the upper triangle of the attention matrix calculation.

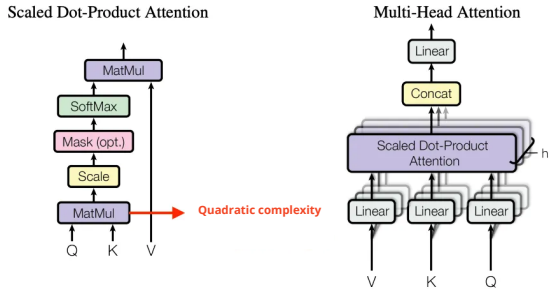


Figure 2.3. Illustration of the main attention mechanisms in a Transformer.

A major downside to Transformers is the quadratic complexity of the attention mechanism (Figure 2.3), which makes them computationally inefficient for long

sequences. This has been addressed by a wealth of techniques [120], such as sparsifing attention, targeting recurrence, downsampling, random or low-rank approximations.

Position Embeddings are indispensable for Transformers to be able to process sequences, as they do not have any notion of order or position of tokens in a sequence. The most common type of position embedding is a sinusoidal embedding with a fixed frequency and phase, $f(x) = \sin(\omega x + \phi)$, where ω is the frequency and ϕ is the phase which are learned as part of the training process, and they are typically shared across all tokens in the sequence. Integrating position information into Transformers can be achieved in different ways, which [105, Table 1] gives an overview for.

Transformers have gradually taken over as an end-to-end architecture for both NLP and CV tasks, albeit adoption in CV has been slower, due to the lack of spatial invariance in the original Transformer architecture. This has been addressed by recent works, such as Vision Transformer (ViT) [101], which uses a patch-based input representation with position embeddings.

A **large language model** (LLM) consists of a stack of Transformers that is pretrained on a large corpus of text, typically using a self-supervised learning objective, such as predicting the next token in a sequence. The goal of LLMs is to learn a general-purpose language representation that can be fine-tuned to perform well on a wide range of downstream tasks. LLMs have disrupted NLP in recent years, as they have achieved SOTA performance on a wide range of tasks thanks to pretraining on large amounts of data. The most popular LLMs are BERT [95], RoBERTa [287], ELECTRA [73], T5 [383], GPT-3 [52], Llama-2 [452], and Mistral [199]. Next to challenges specific to modeling document inputs, explained in Section 2.3.4, open challenges for LLMs include: (i) structured output generation, (ii) domain-specific knowledge injection (*e.g.*, does retrieval-augmented generation (RAG) suffice? [253, 347]), (iii) multimodality.

Vision-language models (VLM) are a recent development in multimodal learning, which combine the power of LLMs with vision encoders to perform tasks that require understanding both visual and textual information. The most popular VLMs are CLIP [381], UNITER [70], FLAVA [423] and GPT-4 [344].

In every chapter of this dissertation we have used Transformers, either as part of a *foundation model* for DU tasks (Chapters 4 to 6) or to contrast with 1-D CNNs in text classification (Chapter 3). Note that [265] share our concerns that NLP needs a new 'playground' with more realistic tasks and benchmarks, which extend beyond sentence-level contexts to more complex document-level tasks. Alternative sub-quadratic architectures have started addressing Transformer's

computational inefficiency on long sequences, *e.g.*, Mamba [152] and Longnet [99]. Time will tell if these will be able to compete with the Transformer's dominance in foundation models.

2.2 Reliability and Robustness

Chapter 3 contains a lot of relevant content on the basic relation between uncertainty quantification, calibration, and distributional generalization or detection tasks. Here, we will focus on the more general concepts of reliability and robustness, and how they relate to concepts used throughout the rest of the thesis. Next, we discuss the need for confidence estimation and appropriate evaluation metrics, followed by short summaries of the main research trends in calibration and uncertainty quantification.

Emerging guidance and regulations [2, 3, 475] place increasing importance on the reliability and robustness of ML systems, particularly once they are used in the public sphere or in safety-critical applications. In ML, reliability and robustness are often used interchangeably [78, 420, 455], yet they are distinct concepts, and it is important to understand the difference between them. This thesis uses the following definitions of reliability and robustness, adapted from systems engineering literature [395]:

Definition 3 [*Reliability*]. *Reliability* is the ability of a system to consistently perform its intended function in a specific, known environment for a specific period of time, with a specific level of expected accuracy [395]. Closer to the ML context, this entails all evaluation under the i.i.d. assumption, allowing for some benign shifts of the distribution, including predictive performance evaluation with task-dependent metrics (accuracy, F1, perplexity, *etc.*), *calibration*, *selective prediction*, *uncertainty estimation*, *etc.*

Reliability requires to clearly specify the role an ML component plays in a larger system, and to define the expected behavior of the system as a function of alignment with the training data distribution. This is particularly important in the context of *black-box* models, where the inner workings of the model are not transparent to the user. In this case, the user needs to be aware of the model's limitations, *e.g.*, model misspecification, lack of training data, and the model needs to be able to communicate its own uncertainty to the user. This is the focus of Chapter 3.

Definition 4 [*Robustness*]. *Robustness* is the ability of a system to maintain its intended function despite a wide range of disturbances, with a minimal

degradation of performance [395]. Such disturbances can take the form of adversarial attacks, distributional shifts, or other types of noise. In the ML context, this entails all evaluation violating the i.i.d. assumption, including adversarial and label noise robustness, out-of-distribution detection, domain generalization, extrapolation, *etc.*

Robustness is more involved with the application scope in which a model can perform well, assuming that the model can maintain some degree of its prediction capacity on non-i.i.d. data which might be unknown at training time. Detecting when the model is operating outside of its intended scope is an important part of robustness to prevent failure propagation to downstream systems.

Resilience is another component of the R^3**: reliability, robustness, resilience** concept in systems engineering, yet it is not a focus of this thesis, nor is it a relevant qualifier of the ML model in isolation, as it is more related to the system as a whole. Resilient systems are able to recover from disturbances, even those caused by model misspecification, *e.g.*, by adapting to new environments and unexpected inputs from unknown distributions or by self-healing.

2.2.1 Generalization and Adaptation

To complete the R^3 picture, we cannot overlook the **generalization-adaptation** spectrum, which has been less explored in our works, yet it is an important part of current practices in ML.

Definition 5 [*Generalization-adaptation*]. *Generalization* is the ability of a system to perform its intended function in a wide range of environments, including those not known at design time [395]. Each environment is defined by a data distribution over a domain and a task, and generalization is the ability of a model to perform well on new data drawn from the same distribution. *Adaptation* is the ability of a system to perform its intended function in a specific, known environment, despite changes in the system itself or its environment [395]. This entails the ability of a model to perform well on new data drawn from a different distribution, which is known at design time.
Different settings of generalization-adaptation are: *in-distribution* (same domain and task), *domain generalization* (same task, different domain), *task generalization* (same domain, different task), *out-of-distribution* (different domain or task). If the model has access to limited samples for training on the new distribution, it is referred to as *few-shot learning* or no samples at all, *zero-shot learning*; if it is able to adapt to new distributions over time, or accumulate knowledge over different tasks without retraining from scratch [87], it is referred to as *continual learning* or *incremental learning*.

Many of these settings are referred to in business as out-of-the-box, self-learning, yet without any formal definitions given. Domain and task generalization are major selling points of pretrained LLMs, which are able to perform well on a wide range of tasks and domains. In the case of very different distributions, *e.g.*, a different task/expected output or an additional domain/input modality, it is often necessary to fine-tune the model on a small amount of data from the new distribution, which is known as *transfer learning*. Specific to LLMs, *instruction tuning* is a form of transfer learning, where samples from a new distribution are appended with natural language instructions [69, 532]. This approach has been used in Chapter 5 to adapt pretrained LLMs to the task of DocVQA, in an effort to reduce the amount of annotated data required to generalize to unseen domains and questions.

2.2.2 Confidence Estimation

A quintessential component of reliability and robustness requires a model to estimate its own uncertainty, or inversely to translate model outputs into probabilities or 'confidence' (Definition 6).

Definition 6 [*Confidence Scoring Function*]. Any function $g : \mathcal{X} \to \mathbb{R}$ whose continuous output aims to separate a model's failures from correct predictions can be interpreted as a confidence scoring function (CSF) [193]. Note that while it is preferable to have the output domain of $g \in [0, 1]$ for easier thresholding, this is not a strict requirement.

Circling back on the question of why one needs a CSF, there are multiple reasons: i) ML models are continually improving, yet 0 test error is an illusion, even a toy dataset (MNIST) is not perfectly separable; ii) once a model is deployed, performance deterioration is expected due to i.i.d. assumptions breaking; iii) generative models are prone to hallucinations [198], requiring some control mechanisms and guardrails to guide them.

Below, we present some common CSFs used in practice [114, 172, 194, 539], where for convenience the subscript is reused to denote the k-th element of the output vector $g(x) = g_k(x)$.

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 training distribution
- Mahalanobis distance [390]: The minimum distance of the feature map (*e.g.*, penultimate layer activations) of a test input to class-conditional Gaussian distributions of the training data.

VI. Bayesian uncertainty estimation

Chapter 3 used MSP and negative entropy as CSFs, next to various PUQ methods for Bayesian uncertainty estimation. Other chapters used MSP as it is the most common CSF in practice, requiring only logits as input. From the use of CSFs also follows the need to evaluate their statistical quality next to task-specific predictive performance metrics, which is discussed next.

2.2.3 Evaluation Metrics

In an ideal world, the evaluation metric of interest would be the same as the loss function used for training, yet this is rarely the case in practice, as the gradient-based optimization process requires a continuously differentiable function, while the metric of interest is often non-differentiable, *e.g.*, accuracy vs. cross-entropy in classification.

Throughout our works, we have used (or extended) multiple predictive performance, calibration, and robustness metrics, of which the most interesting are respectively outlined.

Average Normalized Levenshtein Similarity (ANLS) is a metric introduced in [39] for the evaluation of VQA, which was then extended [449] to support *lists* and be invariant to the order of provided answers. We adapted the underlying Levenshtein Distance (LD) metric [251] to support *not-answerable* questions, $\text{NA}(G) = \mathbb{I}[\text{type}(G) = \text{ not-answerable }]$ (see Equation (2.7)).

Consider for simplicity, the evaluation of a single non-list ground truth answer G and prediction $\hat{P}$, each with string lengths $|G|$ and $|\hat{P}|$, respectively.

$$\mathrm{LD}(G,\hat{P}) = \begin{cases} 1 \quad \text{if } \mathrm{NA}(G) \wedge |\hat{P}| > 0, \\ 0 \quad \text{if } \mathrm{NA}(G) \wedge |\hat{P}| = 0, \\ |G| \quad \text{if } |\hat{P}| = 0, \\ \mathrm{LD}(\mathrm{tail}(G), \mathrm{tail}(\hat{P})) \quad \text{if } G[0] = \hat{P}[0], \\ 1 + \min \begin{cases} \mathrm{LD}(\mathrm{tail}(G), \hat{P}) & \text{if } G[0] \neq \hat{P}[0] \text{ (deletion)}, \\ \mathrm{LD}(G, \mathrm{tail}(\hat{P})) & \text{if } G[0] \neq \hat{P}[0] \text{ (insertion)}, \\ \mathrm{LD}(\mathrm{tail}(G), \mathrm{tail}(\hat{P})) & \text{if } G[0] \neq \hat{P}[0] \text{ (substitution)} \end{cases} \end{cases} \tag{2.7}$$

Each of the conditions is tested in turn, and the first one that is true is executed. The normalized similarity metric is then defined as

$$\mathrm{NLS}(G,\hat{P}) = \frac{1 - \mathrm{LD}(G,\hat{P})}{\max(1, |G|, |\hat{P}|)}.$$

Given multiple ground truth answer variants $G = \{a_1, a_2, ...\}$ and a predicted answer for $\hat{P}_{Q_i}$ for each question Q in the test set of size N, we define the complete metric as follows:

$$\mathrm{ANLS} = \frac{1}{N} \sum_{i=1}^{N} \left(\max_{a \in G_i} s\left(a, \hat{P}_{Q_i}\right) \right) \tag{2.8}$$

$$s\left(a, \hat{P}_{Q_i}\right) = \begin{cases} \mathrm{NLS}\left(a, \hat{P}_{Q_i}\right) & \text{if } \mathrm{NLS}\left(a, \hat{P}_{Q_i}\right) \geqslant \tau \\ 0 & \text{if } \mathrm{NLS}\left(a, \hat{P}_{Q_i}\right) < \tau \end{cases}, \tag{2.9}$$

where we follow prior literature [39, 449] in setting the threshold $\tau = 0.5$.

In the case of a *list*-type question, Hungarian matching is performed following [449] according to NLS between each ground truth answer part and each prediction answer part.

Proper scoring rules [330] are used for generic evaluation of predictive performance, which calculate scoring at the instance-level while measuring both the quality of the predictive function and predicted probability distribution (as they are not compatible with an arbitrary CSF):

- **Negative Log Likelihood (NLL)** [378] 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.

$$\ell_{\text{NLL}}(f) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}\left[y_i = k\right] \cdot \log\left(f_k(x_i)\right) \tag{2.10}$$

- **Brier Score** [50] is a scoring rule that measures the accuracy of a probabilistic classifier and is related to the *mean-squared error* (MSE) loss function. Brier score is more commonly used in industrial practice since it is an λ_2 metric (score between 0 and 1), yet it penalizes tail probabilities less severely than NLL.

$$\ell_{\text{BS}}(f) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \left(\mathbb{I}\left(y_i = k\right) - f_k(x_i)\right)^2 \tag{2.11}$$

All metrics following require a CSF $g(x)$ to be defined, and can pertain to specific evaluation settings [389] tested in Section 3.4.5.

Expected Calibration Error (ECE) [156, 332] is a default metric to evaluate top-1 prediction miscalibration. A calibration estimator (Definition 7) measures the $\mathcal{L}_p$ norm difference between a model's posterior and the true likelihood of being correct.

Definition 7 ($\mathcal{L}_p$ Calibration Error)**.** *[231, 463]*
The $\mathcal{L}_p$ calibration error of $f : \mathcal{X} \to \Delta^{\mathcal{Y}}$ over the joint distribution $(X \times Y)$ with the $\mathcal{L}_p$ norm $p \in [1, \infty)$ is given by:

$$\text{CE}_p(f)^p = \mathbb{E}_{(X,Y)}\left[\|\mathbb{E}[Y \mid f(X)] - f(X)\|_p^p\right] \tag{2.12}$$

The popular ECE metric [332] with condition $\mathbb{I}[Y = \hat{y}]$ is a special case of the above with $p = 1$, where the expectation is approximated using a histogram. `MaxCE` defines the worst-case risk version with $p = \infty$, effectively reporting on the bin with the highest error. As part of Chapter 5, we contributed a novel empirical estimator of top-1 calibration for the task of VQA, where the exact accuracy condition $\mathbb{I}[Y = \hat{y}]$ in ECEis replaced by $\mathbb{I}[\text{ANLS}(y, \hat{y}) > \tau]$. Prior work [329] used a similar strategy of thresholding continuous quality scores to be able to estimate ECE.

In practice, ECE is implemented as a histogram binning estimator that discretizes predicted probabilities into ranges of possible values for which conditional expectation can be estimated. Concretely, the probability space is partitioned into B bins b_i with $i \in \{1, ..., B\}$, where for each bin b_i the gap between observed accuracy and bin confidence $\bar{P}_b$ is measured, with a final

average weighted by the number of samples per bin $|b_i|$.

$$\mathrm{ECE} = \sum_{i=1}^{B} \frac{|b_i|}{N} \left| acc(b_i) - \bar{P}_b\left(b_i\right) \right| \tag{2.13}$$

To minimize the drawbacks inherited from histogram binning, as suggested by the literature [231, 342, 393, 463], we have applied an equal-mass binning scheme with 100 bins (close to $\sqrt{N}$). While plenty of histogram-based ECE estimator implementations exist, many design hyperparameters are not reported or exposed:

I. ℓ_p norm
II. The number of bins (beyond the unfounded default of $|B| = 15$)
III. Different binning schemes (equal-range, equal-mass)
IV. Binning range to define the operating zone
V. Proxy used as bin accuracy (lower-*e.g.*, center, upper-edge)

We upstreamed [1] a generic implementation of binning-based ECE as part of the ICDAR 2023 DUDE competition (Chapter 5).

Alternative formulations have been developed for multi-class [342, 370, 492] and multi-label calibration [493, 520]. Measurements of "strong" calibration, over the full predicted vector instead of the winning class, are reported less in practice. Possible reasons are that they render class-wise scorings, either based on adaptive thresholds or require estimation of kernel-based calibration error to derive hypothesis tests. While we are mindful of alternatives (revisited in Section 2.2.4), we have found that the simpler "weak" calibration measured by ECE meets the practical requirements for most of our benchmarking.

Area-Under-Risk-Coverage-Curve (AURC) [138, 193] measures the possible trade-offs between coverage (proportion of test set%) and risk (error % under given coverage). The metric explicitly assesses i.i.d. failure detection performance as desired for safe deployment. It has advantages as a primary evaluation metric given that it is effective both when underlying prediction models are the same or different (as opposed to AUROC or AUPR). Its most general form (without any curve approximation), with a task-specific evaluation metric ℓ and CSF g, is defined as:

$$\mathrm{AURC}(f, g) = \mathbb{E}_{x \sim \mathbb{P}(X)} \left[\frac{\mathbb{E}_{(\tilde{x},\tilde{y}) \sim \mathbb{P}_{XY}} [\ell([f(\tilde{x})], \tilde{y}) \mathbb{I}[g(\tilde{x}) > g(x)]]}{\mathbb{E}_{\tilde{x} \sim \mathbb{P}_X} [\mathbb{I}[g(\tilde{x}) > g(x)]]} \right] \tag{2.14}$$

This captures the intuition that the CSF g should be able to rank instances by their risk, and that the risk should be low for instances with high confidence.

[1] https://huggingface.co/spaces/jordyvl/ece

The standard curve metric can be obtained by sorting all CSF estimates and evaluating risk ($\frac{FP}{TP+FP}$) and coverage ($\frac{TP+FP}{TP+FP+FN+TN}$) for each threshold t (P if above threshold) from high to low, together with their respective correctness (T if correct). This is normally based on exact match, yet for generative evaluation in Section 5.3.5, we have applied ANLS thresholding instead. Formulated this way, the best possible AURC is constrained by the model's test error (1-ANLS) and the number of test instances. AURC might be more sensible for evaluating in a high-accuracy regime (*e.g.*, 95% accuracy), where risk can be better controlled and error tolerance is an apriori system-level decision [115]. This metric was used in every chapter of Part II.

For the evaluation under distribution shift in Chapter 3, we have used binary classification metrics following [172], **Area Under the Receiver Operating Characteristic Curve** (AUROC) and **Area Under the Precision-Recall Curve** (AUPR), which are threshold-independent measures that summarize detection statistics of positive (out-of-distribution) versus negative (in-distribution) instances. In this setting, AUROC corresponds to the probability that a randomly chosen out-of-distribution sample is assigned a higher confidence score than a randomly chosen in-distribution sample. AUPR is more informative under class imbalance.

2.2.4 Calibration

The study of calibration originated in the meteorology and statistics literature, primarily in the context of **proper loss functions** [330] for evaluating probabilistic forecasts. Calibration promises *i) interpretability, ii) system integration, iii) active learning, and iv) improved accuracy.* A calibrated model, as defined in Definition 8, can be interpreted as a *probabilistic* model, which can be integrated into a larger system, and can guide active learning with potentially fewer samples. Research into calibration regained popularity after repeated empirical observations of overconfidence in DNNs [156, 339].

Definition 8 (Perfect calibration)**.** *[86, 88, 520] 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:*

$$\mathbb{P}(Y = \hat{Y} \mid f(X) = \rho) = \rho, \quad \forall \rho \in [0, 1] \tag{2.15}$$

Below, we characterize calibration research in two directions: (A) CSF evaluation with both theoretical guarantees and practical estimation methodologies

- Estimators for calibration notions beyond top-1 [229, 231, 342, 463]

- Theoretical frameworks to *generalize* over existing metrics and design novel metrics [43, 231, 492, 493]
- *Specialize* towards a task such as multi-class classification [463], regression [228, 428], or structured prediction [227]
- Alternative error estimation procedures, based on histogram regression [156, 331, 332, 340, 343], kernels [230, 370, 492, 493] or splines [159]

(B) Calibration methods for improving the reliability of a model by adapting the CSF or inducing calibration during training of f:

- Learn a post-hoc forecaster $F : f(X) \to [0, 1]$ on top of f (overview: [298])
- Modify the training procedure with regularization (overview: [277, 370])

Due to its importance in practice, we will provide more detail on **train-time calibration** methods. It has been shown for a broad class of loss functions that risk minimization leads to Fisher consistent, Bayes optimal classifiers in the asymptotic limit [25, 495]. These can be shown to decompose into a sum of multiple metrics including both accuracy and calibration error [144, 177]. However, there is no –finite data, nor asymptotic– guarantee that classifiers trained with proper loss functions containing an explicit calibration term will eventually be well-calibrated. In practice, being entangled with other optimization terms often leads to sub-optimal calibration. For this reason, recent studies [12, 230, 492] have derived trainable estimators of calibration to have a better handle ($\gamma > 0$) on penalizing miscalibration, i.e., by jointly optimizing risk ($R(f) = \mathbb{E}_{X,Y}[\ell\,(Y, f(X))]$) and parameterized calibration error (CE) as in Equation (2.16).

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left(\mathrm{R}(f) + \gamma\,\mathrm{CE}(f)\right) \tag{2.16}$$

Many of these methods are implicitly or explicitly maximizing entropy of predictions or entropy relative to another probability distribution, *e.g.*, Entropy Regularization [361], Label Smoothing (LS) [327], Focal Loss [324], Margin-based LS [277], next to more direct (differentiable), kernel-based calibration error estimation [211, 230, 370, 492, 493, 526]. We had expected community contribution on the DUDE competition (Chapter 5) to take advantage of this wealth of calibration methods, yet the majority of submissions used uncalibrated models with MSP, requiring more education on the importance of calibration in practice.

For the sake of completeness, there exist different notions of calibration, differing in the subset of predictions considered over $\Delta^{\mathcal{Y}}$ [463]:

I. `top-1` [156]
II. `top-r` [159]
III. `canonical` calibration [51]

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

$$\mathbb{P}(Y = y_k \mid f(X) = \rho) = \rho_k \quad \forall k \in [K] \wedge \forall \rho \in [0,1]^K \text{ where } K = |\mathcal{Y}|. \quad (2.17)$$

However, the most strict notion of calibration becomes infeasible to compute once the output space cardinality exceeds a certain size [157].

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. Some relaxed notions of calibration have been proposed, which are more feasible to compute and can be used to compare models on a more equal footing. These include: `top-label` [157], `top-r` [159], `within-top-r` [159], `marginal` [229, 231, 342, 492].

2.2.5 Predictive Uncertainty Quantification

Bayes' theorem [26] is a fundamental result in probability theory, which provides a principled way to update beliefs about an event given new evidence. **Bayesian Deep Learning** (BDL) methods build on these solid mathematical foundations and promise reliable predictive uncertainty quantification (PUQ) [124, 136, 140, 238, 301, 325, 326, 464, 466, 496].

The Bayesian approach consists of casting learning and prediction as an **inference** task about *hypotheses* (uncertain quantities, with θ representing all BNN parameters: weights w, biases b, and model structure) from training *data* (measurable quantities, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N = (X, Y)$).

Bayesian Neural Networks (BNN) are in theory able to avoid the pitfalls of stochastic non-convex optimization on non-linear tunable functions with many high-dimensional parameters [300]. More specifically, BNNs can capture the uncertainty in the NN parameters by learning a distribution over them, rather than a single point estimate. This offers advantages in terms of data efficiency, avoiding overfitting thanks to regularization from parameter priors, model complexity control, and robustness to noise due to the probabilistic nature. However, they come with their own challenges such as the increased computational cost of learning and inference, the difficulty of specifying appropriate weight or function priors, and the need for specialized training algorithms or architectural extensions.

For a fixed model m, the analytically intractable Bayesian posterior distribution of the parameters θ is given by Bayes' rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta \mid m)}{P(\mathcal{D} \mid m)} \quad \begin{array}{ll} P(\mathcal{D} \mid \theta) & \text{likelihood of } \theta \text{ (in model } m) \\ P(\theta) & \text{prior probability of } \theta \\ P(\theta \mid \mathcal{D}) & \text{posterior of } \theta \text{ given data } \mathcal{D} \end{array} \tag{2.18}$$

The denominator $P(\mathcal{D}|m)$ is intractable, since it requires integrating over all possible parameter values weighted by their probabilities. This is known as the *inference problem*, which is the main challenge in BDL, as the posterior distribution is required to compute the *predictive distribution* for any new input (Equation (3.1) further explains this).

In practice, BNNs are often implemented as **Variational Inference** (VI) methods, which approximate the high-dimensional posterior distribution with a tractable distribution family, such as a Gaussian distribution [46]. Let $p(\theta \mid \mathcal{D})$ be the intractable posterior distribution of parameters θ given observed data $\mathcal{D}$, which will be approximated with a simpler, conjugate distribution $q(\theta|\mathcal{D}; \boldsymbol{\phi})$, parameterized by $\boldsymbol{\phi}$ (*e.g.*, mean and variance).

The key idea consists of finding the optimal variational parameters $\boldsymbol{\phi}^*$ that minimize the Kullback–Leibler (KL) divergence between the approximating distribution $q(\theta|\mathcal{D}; \boldsymbol{\phi})$ and the replaced true posterior $p(\theta \mid \mathcal{D})$. This is achieved by maximizing the *evidence lower bound* (ELBO), given by:

$$\text{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q(\theta|\mathcal{D};\boldsymbol{\phi})}[\log p(\mathcal{D}|\theta)] - \text{KL}[q(\theta|\mathcal{D}; \boldsymbol{\phi})||p(\theta)] \tag{2.19}$$

$$= \int q(\theta|\mathcal{D}; \boldsymbol{\phi}) \log \frac{p(\mathcal{D}|\theta)p(\theta)}{q(\theta|\mathcal{D}; \boldsymbol{\phi})} d\theta \tag{2.20}$$

$$= \int q(\theta|\mathcal{D}; \boldsymbol{\phi}) \log p(\mathcal{D}|\theta) d\theta - \int q(\theta|\mathcal{D}; \boldsymbol{\phi}) \log \frac{q(\theta|\mathcal{D}; \boldsymbol{\phi})}{p(\theta)} d\theta, \tag{2.21}$$

where the first term Equation (2.21) represents the expected likelihood of the data given the parameters, and the second term quantifies the dissimilarity between the variational distribution and the prior distribution over the parameters. Maximizing the ELBO with $\boldsymbol{\phi}$ is equivalent to minimizing the KL divergence between $q(\theta|\mathcal{D}; \boldsymbol{\phi})$ and $p(\theta|\mathcal{D})$, thereby providing a lower bound on the log marginal likelihood $\log p(\mathcal{D}) \geq \text{ELBO}(\boldsymbol{\phi})$, after the parameters θ have been integrated out. By optimizing the variational parameters $\boldsymbol{\phi}$, we simultaneously

fit the model to the data well and ensure that the approximate posterior is encouraged to be as close as possible to the true posterior distribution.

Even a non-Bayesian, classic NN can be interpreted in this framework as an approximate, degenerate posterior distribution, *i.e.*, a Dirac delta function centered on the MAP estimate of the parameters, $q(\theta|\mathcal{D};\boldsymbol{\phi}) = \delta(\theta - \hat{\theta}_{\text{MAP}})$. More PUQ methods based on different posterior approximations are discussed in detail in Chapter 3, with additional updates on the state-of-the-art.

2.2.6 Failure Prediction

Based on the principle of selective prediction [138, 139], **failure prediction** is the task of predicting whether a model will fail on a given input. In every chapter following Chapter 3, this topic is addressed in the context of the respective task. Since it is an important topic in the context of IA-DU that is generating increasing interest [81, 114, 127, 193, 391], it warrants a brief overview of how it provides a unified perspective. We refer the reader to [171, 536] for a comprehensive survey.

Failure prediction subsumes many related tasks in the sense that it requires a failure source to be defined to form a binary classification task. The failure source can be i.i.d. mispredictions, covariate shifts (*e.g.*, input corruptions, concept drift, domain shift), a new class, domain, modality, task, or concept. The goal of failure prediction is to predict these failures before they occur, allowing for more reliable and robust ML systems.

First, note that calibration does not imply failure prediction, as a calibrated model w.r.t. i.i.d. data can still be overconfident on OOD inputs [549]. The example in Example 2.2.1 sketches the independent requirements of calibration and confidence ranking.

Example 2.2.1. Classifier A scores 90% accuracy on the test set, with a CSF using the entire range $[0, 1]$. Classifier B scores 92% accuracy on the test set, but the CSF always reports 0.92 for any input. Which classifier is preferred in a real-world setting?

- Classifier A is calibrated, but it is not possible to know whether it will fail on a given input.
- Classifier B might be less calibrated, but the CSF allows separability to predict failure on a given input.

Specific to OOD failure prediction, [527] provides a comprehensive categorization of failure tasks and methods.

2.3 Document Understanding

This Section focuses on the history and definition of DU as a field of AI.

Like all subfields of AI, DU has been evolving rapidly, and the definition of a document has been changing accordingly. We identify three main stages in the evolution of the field, dependent on a) the type of learning, b) the unit of study, and c) the modality of the input.

Regarding a), it has followed the natural evolution of rule-based systems, to learning-based systems, to deep learning systems to build representations of documents. Regarding b), the field has evolved from region-based analysis, to page-level analysis, and now moving to document-level analysis, as we have advocated in our research (Chapters 4 and 5). Regarding c), the field was originally dominated by OCR, particularly CV, then by KIE, emphasizing NLP, and now by both CV and NLP, with more attention given to multimodality and generative models by which new tasks can be approached, *e.g.*, DocEdit [311].

Below, we expound on the evolution of the field through the lens of each modality, and the tasks that are typically associated with it. We also provide an overview of the most popular datasets and models in each task/modality.

The term **Document Understanding** (DU) is used in a variety of contexts (historical, research, commercial), and its definition deserves some attention. A seminal reference [430] dates back to 1992, which defines DU as 'the study of all processes involved in taking a document through various representations': from a physical object to a digital image, from an image to a symbolic description, and from a symbolic description to a high-level semantic representation. At the time, the field was dominated by **Optical Character Recognition** (OCR), particularly CV, and the definition was focused on the physical-to-digital conversion of documents, excluding born-digital documents.

Furthermore, the subterm **document** is used in the context of NLP (in particular in summarization) to denote a textually-rich document: a sequence of words exceeding a sentence or paragraph or a single unit in a corpus. However, in DU it denotes a **visually-rich document** (VRD), which can be a combination of text, images, tables, and other elements. There is no universally established definition of a document [53], and it is used interchangeably with the term page, which is a physical, symbolic unit. In Chapter 4, we come back to this definition, addressing the misalignment of research with how documents occur in practice.

Over time, the quality of OCR has improved, and the focus of the field has shifted from OCR to **document image classification** (DIC) and **key information**

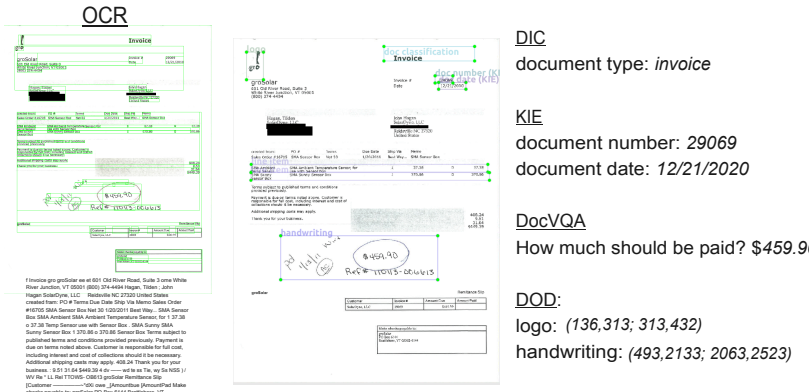


Figure 2.4. A simple illustration of common DU tasks on an example document.

extraction (KIE), which are more application-directed recognition tasks. Arguably, most businesses are interested in the unstructured information contained in documents, rather than the documents themselves. On the commercial side, the combination of these tasks is often referred to as **Intelligent Document Processing** (IDP), albeit 'understanding' has been similarly marketed by *e.g.*, UIPath (originally an RPA company, now looking at AI as the next frontier of automation). The scientific community has been more careful in using the denomination 'understanding' [29], with the DUE benchmark [47] defining it, on the one hand, as an end-to-end process involving a subset of human cognitive skills, and on the other hand, enumeratively with several well-defined problems (OCR, KIE, VQA as defined in Section 2.3.1).

In our research, we have extended DU to denote '*the ability to holistically consume textual and visual elements structured according to rich semantic layouts, and reason over compositional information extracted from a VRD to generate meaningful insights or actions.*'. There is no specific notion of tasks, but rather an emphasis on the end-to-end process leveraging all modalities intrinsic to documents, where a generic DU model is expected to generalize to any task on any document from any domain. This stands in shrill contrast to only DIC and KIE, where local context generalization (key-value pairs) is rewarded, whereas DU as defined here aims to generalize beyond the local context of a document.

2.3.1 Task Definitions

For thorough understanding, each task will be defined in terms of the following components: input, output, model, and evaluation. Most tasks use a single document page as input (for both legacy and computational reasons), and the output depends on the task.

Formally, a **page** p consists of an image $\boldsymbol{v} \in \mathbb{R}^{C \times H \times W}$ (number of channels, height, and width, respectively) with T word tokens $u = \{w_t\}_{t=1}^{T}$, where w_t maps to (sub)words in a vocabulary $\mathcal{V}$, organized according to a layout structure $s = \left\{\left(x_t^1, y_t^1, x_t^2, y_t^2\right)\right\}_{t=1}^{T}$, typically referred to as token bounding boxes (top-left to bottom-right corner), coming from OCR or available from a born-digital document. Standardized notation for document inputs beyond a single page has been established in Chapter 4 [470].

Optical Character Recognition (OCR) is the task of converting a document image to a sequence of characters. The input is a document image, and the output is a sequence of characters. The output space $\mathcal{Y}$ is the set of all possible characters (*e.g.*, *a, b, c, ..., A, B, C, ...*), typically restricted to a subset of characters based on the document language and orthography. The quality is evaluated with a metric such as the word error rate (WER) or the character error rate (CER).

Document Classification (DC) is the task of assigning a document to a predefined class. The input is a document image, and the output is a class label. The output space $\mathcal{Y}$ is the set of all document classes (*e.g.*, *invoice, email, form, advertisement*). Standard metrics are accuracy and F1 score (if class imbalance).

Key Information Extraction (KIE) is the task of extracting key information from a document. The input is a document image, and the output is a set of key-value pairs. The output space $\mathcal{Y}$ is the set of all key-value pairs (*e.g.*, *date: 2024-01-01, total: 1000.00, ...*), where keys are pre-defined as part of a format relevant to the document class in scope. In practice, it is implemented as sequence labeling with $y = \{y_1, y_2, ..., y_T\}$, where $y_t \in \mathcal{Y}$ is a label from a *IOB,IOBES*-encoded labelset $\mathcal{Y}$ *(*B-DATE, I-DATE, ..., O). Extraction quality is evaluated with the sequence F1 score to account for the imbalance with the 'O' token.

Document Visual Question Answering (DocVQA) is the task of answering a question about a document. The input is a document image and a question, and the output is an answer. Depending on the type of question, the output space changes. Extractive questions (ExQA) require a subspan of the document's text as answer, $y = (y_{start}, y_{end})$ with $y_{start} \leq y_{end}$ and $y_{start}, y_{end} \in \{1, ..., T\}$.

Abstractive questions (AbsQA) require a sequence of tokens as answer, $y = \{y_1, y_2, ..., y_{T'}\}$ with $y_t \in \mathcal{V}$. The latter is more complex to evaluate, yet more interesting to test 'understanding' than restricting evaluation to answer spans, which is why we introduced AbsQA as part of Chapter 5. Orthogonal to the previous two types, DUDE introduces *list* questions with multiple or multi-span (ExQA) answers. Predicted answers are evaluated using ANLS, with multiple extensions defined in Section 2.2.3.

Document Layout Analysis (DLA) is the task of analyzing the layout of a document in terms of logical layout elements (*e.g.*, *text blocks, headers, figures, figure, plots, tables, text*). The input is a document image, and the output is a set of bounding boxes and their respective labels. The output space $\mathcal{Y}$ is the set of all possible bounding boxes and labels. More formally, it outputs a set of tuples, where each tuple (b_j, c_j) represents one of J detected logical layout elements. For each, b_j denotes the bounding box for the j-th detected element, defined as (x_j, y_j, w_j, h_j) (in the popular COCO format). c_j is the class label for the j-th element, indicating its object category. Evaluation is done with the standard COCO metrics, *i.e.*, average precision (AP) over different intersection-over-union (IoU) thresholds, and mean AP (mAP).

Document Generation (DG) is the task of generating a document from a set of key-value pairs and potential metadata attributes, *e.g.*, visual appearance, color scheme. The output space $\mathcal{Y}$ is the set of all possible document images, which makes it hard to evaluate in a quantitative manner. Some efforts have been made to define metrics for document generation, *e.g.*, Document Earth Mover's Distance [169], but they are not yet widely adopted.

Other lesser known tasks include document object detection (DOD), table structure recognition (TSR), document retrieval, document editing, document translation, document summarization, document authenticity verification With the rise of multimodal models, more data types are being considered jointly with documents under the umbrella term *visually-situated language*, such as charts, tables, handwriting, text-heavy scenes or illustrations, webpage and user interface screenshots *etc.*

2.3.2 Datasets

With the variety of tasks, there is a large number of datasets available for each DU task. Instead of exhaustively enumerating datasets for each task defined above, we will link to the tables in the respective chapters treating these tasks. We will only highlight some more recent datasets, which are not yet included in the tables.

An overview of document source datasets for pretraining or dataset construction is presented in Table 4.1 as part of Chapter 4.

For an overview of *DC* datasets, see Table 4.2 in the same chapter. For an overview of *KIE* datasets, we refer to [47], with some newer datasets [422, 485] linked here. An overview of *DocVQA* datasets is presented in Table 5.1, with the introduction of the DUDE dataset (Chapter 5). An interesting new addition is PDFTriage [400] which focuses more on retrieval than on QA. Finally, some datasets for *DLA* are presented in Table 6.1 as part of Chapter 6. Other essential datasets are PubLayNet [544] and DocBank [261]; and the novel multidomain M^6 dataset [71].

2.3.3 Models

A model taxonomy is presented in [407] that differentiates models based on the input modalities they use, the geometric approach, dependence on OCR, or the type of output they produce. However, it is far from comprehensive due to missing out on various DU tasks and more recent models. Table 2.2 presents an overview of models that we have applied to various DU tasks, extending the taxonomy with our observations.

Depending on the modalities considered and the requirements of the task, different pretrained models have been used in practice, instead of the document foundation models presented above.

For *document text*, the most popular models are BERT [95], RoBERTa [287], and T5 [383]. Additionally, text-only LLMs such as GPT-3 [52], Llama [452], and Mistral [199] are increasingly applied to document text.

For *document images*, the most popular models are ResNet [167], EfficientNet [439], and DiT [259].

For all modalities combined, the most popular models are the LayoutLM series [187, 502, 503], DocFormer(v2) [15, 16], and UDOP [443]. The former are OCR-based pipelines, with pixel-only models such as Donut [216] and Pix2Struct [247] gaining popularity for increased efficiency, albeit they are still catching up on performance. Alternative approaches include the use of graph neural networks [286, 341, 517] and grid-based models [212, 275], yet their performance lags behind the aforementioned sequence models.

Most of the above-mentioned models have been applied during the Chapter 5 benchmark experiments, with only results missing for multimodal LLMs, which were introduced after the publications of the chapter. An up-to-date overview of newer multimodal LLMs, *e.g.*, GIT2, PaLi, Flamingo, Kosmos-2, GPT-4, Fuyu,

Model	Year	Conf.	Arch.	Input Mod.	Vision Branch
LayoutLMv1 [502]	2020	KDD	E	T + S	-
DocStruct [484]	2020	EMNLP	E	T + V + S	Resnet50
StrucText [266]	2021	ACM	E	T + V + S	Resnet50 + FPN
StructuralLM [254]	2021	ACL	E	T+S	-
LayoutLMv2 [503]	2021	ACL	E	T + V + S	ResNeXt 101
SelfDoc [263]	2021	CVPR	E	-	-
LamBERT [134]	2021	ICDAR	E	T + S	-
TILT [371]	2021	ICDAR	E + D	T + V + S	U-Net
DocFormerv1 [15]	2021	ICCV	E	T + V + S	Resnet50
UniDoc [153]	2021	NeurIPS	E	T+V+S	Resnet50
DiT [259]	2022	ACM	E	V	ViT
LayoutLMv3 [187]	2022	ACM	E	T + V + S	Linear
BROS [181]	2022	AAAI	E	T + S	-
XYLayoutLM [154]	2022	CVPR	E	T + V + S	ResNeXt 101
FormNet [245]	2022	ACL	E	-	-
ERNIE-Layout [264]	2022	EMNLP	E	T + V + S	F-RCNN
LiLT [481]	2022	ACL	E	T + S	-
XDoc [66]	2022	EMNLP	E	T	-
GeoLayoutLM [296]	2023	CVPR	E	T + V + S	F-RCNN+ConvNeXt
Vision Grid Transformer [80]	2023	ICCV	E	T + V + S	ViT
DocFormerv2 [16]	2023	-	E + D	T + V + S	Linear
Donut [216]	2022	ECCV	E + D	V	SwinTransformer
Pix2Struct [247]	2023	ICML	E + D	V	ViT+variableres
UDOP [443]	2023	CVPR	E + D	T + V + S	ResNeXt 101
Hi-VT5 [451]	2023	PatRecog	E + D	T + V + S	ViT
FormNetv2 [246]	2023	ACL	E	T + V + S	3-layer CNN
LayoutMask [458]	2023	ACL	E	T + S	-
UReader [510]	2023	ACL	D	V + S	CLIP-ViT
DocLLM [480]	2024	-	D	T + S	-
Gramformer [44]	2024	-	E + D	T + V + S	Linear
InstructDoc [442]	2024	-	E + D	T + V + S	CLIP-ViT

Table 2.2. Adapted from [16]. A summary of DU prior art is presented with their architecture (E: Encoder, D: Decoder), the input (T: text, V: vision, S: spatial features), the vision features branch and core extensions.

Llava, CogVLM, that could potentially be applied to DU tasks is presented in [512].

2.3.4 Challenges in Document Understanding

To tease the contributions of our works, we will highlight some of the most important challenges in DU, which are shared by all chapters in this thesis.

2.3.4.1 Long-Context Modeling

An important challenge for most SOTA DU models based on the Transformer architecture is long document processing, which is not yet solved satisfactorily, as it is the focus of Chapters 4 and 5.

We illustrate the extent of the problem with the most popular DU model, LayoutLMv3 [187][2] in Figure 2.5, pointing to the quadratic complexity of attention, which cannot be parallelized over pages with encoder-only models. Hi-VT5 [451] is the only model that is by design usable for multipage documents, yet it requires a lot of memory and depends on compressing page information into learnable embeddings.

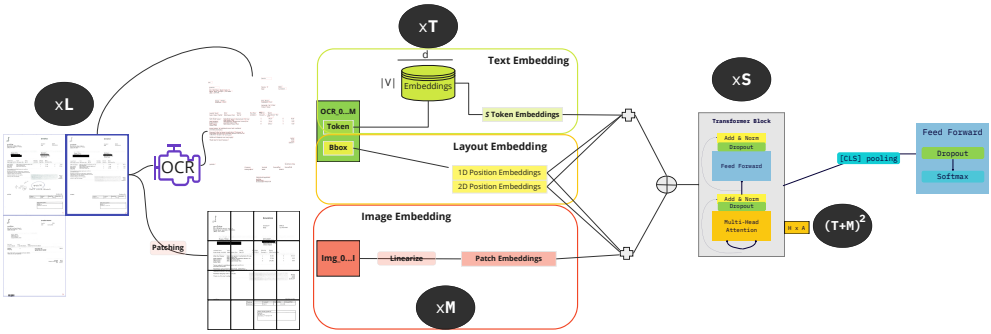


Figure 2.5. Inefficiency of document foundation models for processing multipage documents, illustrated with LayoutLMv3 [187]. Notation: L pages, T text tokens, M linearized visual patches, S Transformer layers

While a page is the modeling unit of preference to maintain computational efficiency in Transformers' processing sequences of tokens, it is not the natural appearance of a document. Some tasks require the global document context and treating each page contextually independent is suboptimal, as argued in our works on multipage document classification (Chapter 4) and DocVQA (Chapter 5) with multi-hop question answering.

Figure 2.6 illustrates how a prototypical multimodal architecture, Hi-VT5 [451], is used for the task of multipage ExVQA.

In principle, every LLM can perform multipage document processing depending on the ability of the LLM to extrapolate to longer context windows, given the position representation method (barring *absolute* positional encodings), and performance relying on also having trained on long sequences, *e.g.*, by

[2] >8.6M model weights downloads in January 2024

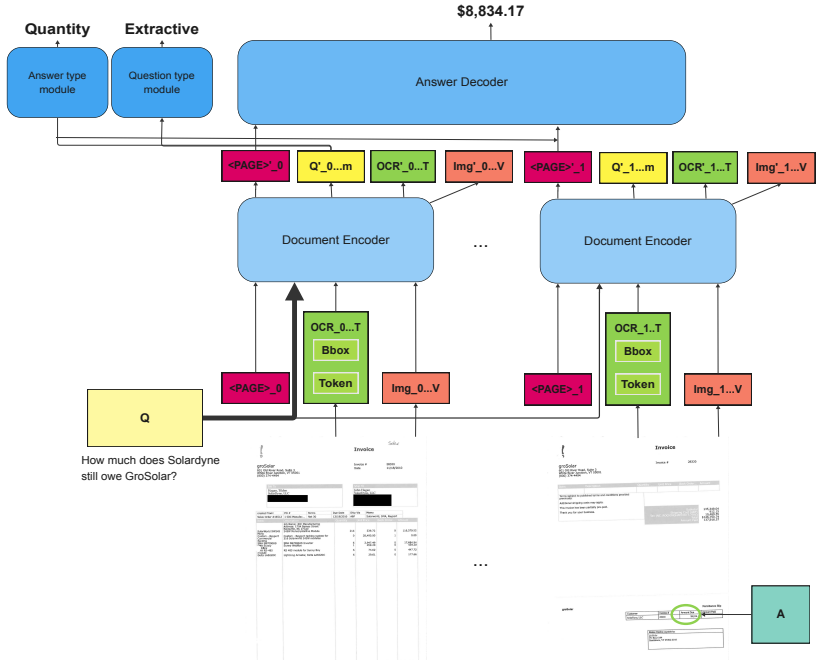


Figure 2.6. Hi-VT5 architecture for multipage, extractive DocVQA.

instruction-tuning on long-context data. Naturally, the computational cost will increase with the length of the input data, yet recently many advances have made subquadratic complexity feasible (e.g. relative positional encodings [382], ALiBi [374], Flashattention [82], multi-query attention [9] *etc.*). [102] provides an overview of the SOTA in long-range Transformers for DU tasks. A recent approach [44] proposes a hierarchical architecture to model both local page-level attention and global document-level attention on learnable document-level tokens, with an additional compression module to scale to 100+ pages while keeping latency low.

2.3.4.2 Document Structure Modeling

Representing structured documents as plain text resulting from OCR is not congruent with how humans perceive documents [294], which is the focus of Chapter 6. Document layout is a valuable cue to navigate a document's structure and find information more efficiently, but it is not always modeled properly, with most methods relying on geometric features (1D/2D absolute positional

encodings) that are not robust to OCR errors, nor are they able to capture the semantic complexity of document layouts.

There are great recent advances from better layout modeling, *e.g.*, modeling relative positions with polar coordinates and layout attention with Gaussian biases [555], and DocLLM [480] ignoring visual features to focus on disentangling the layout structure from the document text, which are promising directions for future research.

2.4 Intelligent Automation

Automation is the use of technology to perform tasks with reduced human assistance. Throughout history, humankind has experienced waves of automation, from the invention of the wheel to the steam engine, the assembly line, and the computer. Manual labor in particular, performed by *blue-collar* workers, has been increasingly automated since the 20th century. When applied to knowledge work as performed by *white-collar* workers, more through the use of software than hardware, it is referred to as **Intelligent Automation** (IA, not to be confused with the French acronym of 'intelligence artificielle') [1].

IA is a rapidly growing field, with the market for hyperautomation-enabling technologies projected to have reached nearly $ 600 billion in 2022, a 24% increase from 2020 [392]. A recent survey [135] does show that IA adoption is lagging behind expectations, with only 19% of organizations having deployed their automation programs and 38% in the planning stage.

[48] identified 5 key trends in IA: 1) the rise of the digital workforce, 2) the emergence of the digital twin, 3) the importance of data, 4) the need for orchestration, and 5) the rise of the citizen developer. The first three trends are particularly relevant to the work presented in this thesis.

IA is a subset of Artificial Intelligence (AI) specifically designed for the automation of knowledge work. It encompasses several technologies, including **Robotic Process Automation** (RPA), which can be thought of as software to automate routine tasks, and Workflow & Business Process Management (BPM). When combined with people and organizations, these technologies are capable of solving major world problems [48].

The goal of IA is to create a software-based digital workforce by mimicking the four main human capabilities required to perform knowledge work: vision, language, thinking & learning, and execution. This allows for the construction of **straight-through** business processes, which are more *efficient* in terms of productivity, processing speed, and cost, and often more *effective* in terms of

quality and logic. The ultimate aim is not to replace human workers, but to take the robot out of the human, augmenting human intelligence, creativity, and productivity.

IDP/DU is a prototypical example of an IA use-case, as it frees workers from paperwork, allowing them to focus on more value-adding tasks, thereby providing a clear perspective on the future of work. Finally, we provide an overview of the requirements for setting up IA, linking back to all technical concepts introduced before.

Enabling IA requires well-defined CSFs and either operational thresholding to determine the trade-off between automation and risk, or a selective prediction setup. When a system is deployed in production, it also requires robustness to distribution shifts, both expected and unexpected, and the ability to detect and predict a wide variety of failures.

Measuring IA is performed using calibration metrics and confidence ranking metrics. Calibration is the degree to which a model's predicted probabilities match the true probabilities of the events it predicts. Confidence ranking is the degree to which a model's predicted probabilities are ranked in accordance with the true probabilities of the events it predicts. If the i.i.d. assumption becomes violated, the model's confidence ranking will be affected, and the model will be overconfident on OOD inputs. As part of the deployment process, it is important to monitor the model's performance and to detect when it starts to fail, where other metrics are more appropriate.

Improving IA Improvements to IA can be made by inducing calibration through post-hoc strategies or designing calibrated loss functions, as well as through predictive uncertainty estimation for model selection and capturing issues with the data or model before deployment, and all investments in failure prediction will be rewarded with more robust and reliable systems.

Part I

Reliable and Robust Deep Learning

Chapter 3

Benchmarking Scalable Predictive Uncertainty in Text Classification

The contents of this chapter come from two publications [465, 466]:

Jordy Van Landeghem, Matthew B Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Benchmarking Scalable Predictive Uncertainty in Text Classification. *IEEE Access*, 2022

The first publication started as a reproduction of [500] with a deeper focus on text classification, and the second publication is a large journal extension of the first publication.

This chapter focuses on how to quantify uncertainty in text classification tasks, which is a prerequisite to trust a model's predictions in real-world applications such as intent classification in automated document processing based on the document text. We conduct a benchmarking study of uncertainty

estimation methods applied on 6 real-world text classification datasets, including both multi-class and multi-label classification, with 1-D convolutional neural networks and pretrained transformers. The experiments empirically investigate why popular scalable uncertainty estimation strategies (*Monte-Carlo Dropout*, *Deep Ensemble*) and notable extensions (*Heteroscedastic*, *Concrete Dropout*) underestimate uncertainty, and how to improve their performance. We motivate that uncertainty estimation benefits from combining posterior approximation procedures, linking it to recent research on how ensembles and variational Bayesian methods navigate the loss landscape.

We find that our proposed method combination of *Deep Ensemble* with *Concrete Dropout*, by analysis of in-domain calibration, cross-domain classification, and novel class robustness, demonstrates superior performance, even at a smaller ensemble size. Our results corroborate the importance of fine-tuning dropout rate to the text classification task at hand, which individually and as an ensemble impacts model robustness. We observe in ablation that pretrained transformers severely underperform in novelty detection, limiting the applicability of transfer learning when distribution shift from novel classes can be expected.

Supporting context: As the publications were written at the start of my PhD, we take the opportunity here to give an update on the state of the art and the relevance of our work in uncertainty estimation research.
The journal extension was motivated as a survey and benchmark of scalable Bayesian Deep Learning methods, in which we introduced novel hybrid models and evaluated uncertainty estimation quality under distribution shift configurations. We also provide a convenient entry point for practitioners, as our benchmarking software is available online (`https://github.com/Jordy-VL/uncertainty-bench`). Our work has also been re-used as the basis of a conference tutorial [524, `https://sites.google.com/view/uq-tutorial`].

In similar spirit as our work, new benchmarks have put different aspects of reliability and robustness to the test: Shifts [306] focuses on the robustness of uncertainty methods to real distribution shifts in large-scale tasks across overlooked modalities such as tabular, audio or sensor data, WILDS [220, 401] curates a collection of labeled and unlabeled datasets exhibiting distribution shifts in the wild, OpenOOD [527] generalizes a comprehensive benchmark for out-of-distribution detection, anomaly detection and open-set recognition, and finally, PLEX [455] probes pretrained models on their ability to estimate uncertainty, exhibit robustness under shifts, and adapt in settings of active, few-shot and life-long learning.

The supremacy of ensemble methods has been challenged by the recent publication of [346], which proposes a new method for uncertainty estimation in

NNs, called *EpiNet*. The authors claim that their non-Bayesian method is able to discern the difference between ambiguity or lack of data. Key ingredients are a dyadic sampling procedure, which creates interesting data pairs that are used to train a NN to predict the epistemic uncertainty, and a small architecture that can supplement any conventional NN to improve OOD detection and active learning [413]. Another competitive method [326] concentrates on feature-space density estimation under the assumption of smoothness and sensitivity, with their efficient baseline disentangling epistemic (Gaussian Mixture Model fit on training features, with a separate covariance matrix per class) and aleatoric uncertainty (entropy of softmax distribution). Other promising methods target aleatoric uncertainty, such as [75, 474] which focus on label noise or ambiguous tasks such as toxicity detection.

An important observation on the benefits of Bayesian NNs concerns the dataset and model size, particularly Bayesian modeling shines in dynamic settings where the size of the model/data are unknown or change over time [346], *e.g.*, online, continual, active and life-long learning. In static settings with high accuracy on a fixed test set, the benefits of Bayesian modeling are less pronounced [215].

Next to PUQ, alternative approaches have sought to learn explicit scoring functions [200, 351] or assess the similarity of inputs to the training distribution [54, 271, 285, 379, 487]. All efforts have recently increased in popularity, as uncertainty estimation has become even more important for safe deployment of LLMs in user-facing applications [111].

3.1 Introduction

Reliable uncertainty quantification is indispensable for any machine learning system trusted in decision-making in many application domains such as medical diagnosis, self-driving cars and automated document processing. In any typical industrial application, we desire predictive uncertainty to communicate on the model's lack of in-domain knowledge due to either training data scarcity or model design errors, or its ability to flag potentially noisy, shifted or unknown input data (see [136] for more detail on sources of uncertainty).

Supervised Deep Learning (DL) algorithms have been found to provide "catastrophically overconfident predictions" [116] under data distribution shift. Specifically, novel class distributions can emerge at inference time [367], which desirably should be detectable in a model's uncertainty. To this end, scalable Bayesian DL (BDL) methods for uncertainty estimation have been recently developed, generating increased interest from practitioners in need of practical solutions. BDL comprises an increasingly large range of theoretically well-

motivated predictive uncertainty methods (PUQ), yet only some are able to scale in network architecture and dataset size. Additionally, most surveys and research output on predictive uncertainty is based on multi-class image classification or regression experiments. We argue that predictive uncertainty methods and how well they scale in Natural Language Processing (NLP), for text classification tasks, is still an under-explored question.

The context of our study is a production-level text classification system for automatically handling incoming communications in information-intensive industries (e.g. legal, banking, insurance). Imagine a digital-first company where each department has its own document classifier operating under a closed world assumption. However, whenever a client mistakenly sends a document (car purchase invoice requesting a loan) to the wrong department (say underwriting or medical claims), this can generate high-confidence false positives that trigger the wrong action (insurance or claim settlement instead of loan application). Similarly, if an insurance broker suddenly decides to completely change the document template that clients use to apply for a car loan, the production model might not find previously salient features which it had learned to rely on for accurate classification. This shows that detection of anomalous inputs and shifting distributions is critical to keep errors in automation low.

We investigate different techniques and procedures for incorporating uncertainty into DL models for text classification, analyzing the degree to which they can reliably capture uncertainty under extrapolation (outside the support of the training set), both individually and combined in an ensemble. Our findings for individual predictive uncertainty methods are overall consistent with benchmarks in other modalities, with Deep Ensemble reporting greater robustness than approximate Bayesian methods. However, we discover from empirical findings that our newly proposed combinations, particularly *MC Concrete Dropout Ensemble*, can push the bounds by exploiting the in-domain calibration effect of Concrete Dropout and all-round ensemble qualities for increased out-of-domain and novel class robustness.

We intend our work to be used as a survey and benchmark of scalable BDL methods, where the architectures and datasets are drawn from NLP, thereby covering a void in the literature on uncertainty estimation in this field. Next to proposing a well-motivated evaluation methodology, this chapter also provides a convenient entry point for practitioners.[1]

Our key contributions can be summarized as follows:

- We conduct a benchmarking study of established uncertainty estimation

[1] Our benchmarking software [TensorFlow 2] is available at https://github.com/Jordy-VL/uncertainty-bench

methods applied on real-world text classification datasets. Our analysis focuses on model robustness and uncertainty quality in realistic data distributions. We propose a practical methodology to test the above, resulting in a better understanding of the individual shortcomings of predictive uncertainty methods.
- We motivate and introduce novel combinations of predictive uncertainty methods, providing empirical evidence for their complementary benefits. Through statistical analyses and ablation experiments we discern the importance of certain prior, model or hyperparameter influences on the reliability of predictive uncertainty.

Organization The paper is organized as follows. Section 3.2 overviews related work in uncertainty benchmarking, distribution shift, and uncertainty estimation in NLP. We present core concepts of BDL in Section 3.3 to build up a thorough understanding of predictive uncertainty in theory and practice. We include this introductory text for readers less familiar with uncertainty methods. Section 3.3.5 critically analyzes the practice of evaluating uncertainty under distribution shift. Sections 3.3.4 and 3.4.1 stand central in our work, connecting recent research on how neural networks navigate the loss landscape with posterior approximation procedures, followed by our work's hypotheses on complementary benefits between predictive uncertainty methods.
Section 3.4 details our methodological setup from datasets, model architectures, uncertainty estimation and evaluation, to experimental settings. We present in Section 3.5 the results of 3 large benchmarking experiments, followed by 4 smaller ablation studies on important hyperparameters. After closing the discussion in Section 3.6 with take-home messages targeting researchers and practitioners interested in uncertainty prediction in text classification, Section 3.7 details additional experiments, and Section 3.8 draws up some limitations of our research. Finally, we synthesize our contributions in Section 3.9 and propose directions for future work on uncertainty research in NLP.
The Appendices support the main text by detailing implementation (A), practical considerations (compute, timings) (B), and detailed evaluation data for full transparency (C).

3.2 Related Work

In this Subsection, we overview recent literature on benchmarking the quality of uncertainty quantification in DL and more specifically research on uncertainty estimation for NLP tasks.

Increasingly, there are efforts from the research community to help BDL methods scale to real-world scenarios [205]. Benchmarks are an important tool to help researchers prioritize the right approaches and to inform practitioners which methods are suited for their applications [276]. There is a growing demand for benchmarking in BDL, since methods must be scored both for task performance and uncertainty quality [411, 496]. Rigorously evaluating the latter is considerably more difficult, since depending on the problem setting no direct uncertainty ground-truth exists, requiring a well-defined experimental setup [323].

A standard benchmark in BDL is *UCI* [176], a set of curated regression datasets, which allows to judge uncertainty quality with the predictive log-likelihood metric. However, its general applicability and validity has been criticized on multiple accounts [113, 323, 360].

More recently, [19, 113, 301, 348, 462] presented large-scale evaluation studies of BDL methods with benchmarking on real-world datasets. These studies motivate data retention and distribution shift as generic protocols for evaluating predictive uncertainty. Similarly, we argue that even mild shifts of data are unavoidable in real-world applications and, conditional to specific distribution shift assumptions (see Section 3.3.5), this provides a good testing ground for uncertainty evaluation.

[348] consider two types of distribution shift: (a) *out-of-distribution* (OOD) data from separate datasets, and (b) *adversarial shift*, where the test distribution consists of perturbed or corrupted ground truth data isolated from training. In our work we propose novel class detection as an alternative to a), which we motivate to be a more representative experimental setup for testing uncertainty in text classification (more detail in Subsections 3.3.5 and 3.4.5.3). [142] bring a similar argument against b) that adversarial examples are often overly synthetic and disconnected from real-world performance concerns, which we assert to be especially true for perturbations applied to text data. Therefore, we derive a challenging experimental setup for b) (more detail in Section 3.4.5.2) inspired by the extensive literature in NLP on the problem of domain shifts and domain adaptation [45, 84, 129, 203, 388, 557]. Domain adaptation approaches aim to mitigate performance degradation that occurs when transferring a classifier from a source domain to a target domain. Learning under domain shift presents a complex challenge in text classification since linguistic patterns can be highly different across domains, even harder to tackle when domains are unknown a priori [388]. While out-of-domain generalization is the ultimate objective [18], we believe that accurate uncertainty prediction has a major role to play in the detection of out-of-domain data, which is currently under-explored. [488] is a notable exception where predictive uncertainty methods are leveraged to learn domain-invariant features in unsupervised fashion.

In this work we only consider methods that directly estimate the predictive posterior and aim at obtaining high quality uncertainty estimates by discriminative models without any additional OOD components. However, there exists a large number of alternative OOD detection and generalization approaches. We surmise that these can be more effective in handling the above distribution shifts, yet they have different modeling assumptions which complicates a direct comparison, for instance, access to (auxiliary) OOD data [271, 285], generative modeling [334], focus on abstention mechanisms [138], or characterization of dataset shifts with a two-sample-testing approach [379]. We recommend [54, 414] for an overview of these approaches.

While previous BDL benchmarks have helped standardize protocols, metrics and analysis tools, the effort is not spent equally across all modality and problem settings (as can be observed in the survey of [4]). Arguably, most research on uncertainty estimation focuses on regression and image classification tasks as they offer visual validation on uncertainty quality, *e.g.*, [214].

Tasks in the NLP field involve discrete natural language units (word, sentence, paragraph) as input, which requires a translation to the continuous domain by embedding discrete units to form high-dimensional distributed representations [321]. This presents additional complexity compared to image or time-series data which as continuous signals can be directly fed into a Neural Network (NN). Furthermore, specialized algorithms (*e.g.*, dealing with long sequences, attention for larger memory [473]) and progressively more complex architectures [27] are being created to tackle this unique challenge in NLP, which can affect the performance of predictive uncertainty techniques. With our work, we start the exploration into effects of field characteristics, notably different NLP architectures, inherent task complexity, and properties of language in text processing (*e.g.*, ambiguity [397], document length [478], pre-defined vocabulary [68]) that could cause problems when predicting uncertainty. More specifically, we seek to answer how uncertainty research translates to a prototypical language task such as *text classification*, which more frequently than vision tasks is characterized by non-mutually exclusive labels [312], a problem setting ignored by existing BDL benchmarks.

BDL research on NLP tasks is generally limited, certainly when considering quantitative evaluation of predictive uncertainty quality. While we draw inspiration from the uncertainty estimation methods of [500], their study focuses on the performance increase of non-probabilistic measures (mean-squared error) and only reports sentiment regression results. Moreover, we find no quantitative evaluation of the quality of the uncertainty scores and comparison to simpler measures of uncertainty, for instance, softmax score or predictive entropy. [174] does focus on the robustness of pretrained Transformers to distribution shift, yet without application of any predictive uncertainty methods. [322, 533] present

similar setups applying Monte Carlo Dropout to regular NLP architectures in an active learning setup, yet they only aim to increase overall predictive performance by relying on in-domain calibration. Our work benchmarks individual and joint predictive uncertainty methods in multiple text classification task settings over two well-motivated uncertainty evaluation setups, testing robustness to distribution shift for NLP problems.

3.3 Uncertainty Methods

The first Subsection formally presents how to quantify uncertainty in BDL and how popular methods approach inference differently. Section 3.3.2 treats predictive uncertainty methods with a focus on the algorithmic procedure, followed by representative method extensions for more reliable uncertainty estimation. Section 3.3.3 describes from what sources uncertainty originates and how to quantify uncertainty at test-time. In Section 3.3.4 we present the rationale of our study, connecting recent research on how NNs navigate the optimization landscape with the posterior approximation procedure of methods from Section 3.3.2. Section 3.3.5 provides a critical note on how distribution shift impacts uncertainty estimation and the evaluation thereof.

3.3.1 Quantifying Uncertainty in Deep Learning

In modern Deep Learning, two common uncertainty (or inversely "confidence") estimates are the maximum posterior class probability, known as *softmax-score*, and the *predictive entropy* over posterior class probabilities [415, 522]. However, [156]'s work on confidence calibration demonstrated these to be unreliable estimates of Neural Networks' uncertainty. While post-hoc calibration methods such as Temperature or Vector Scaling [156, 419] can easily calibrate classifier uncertainty in-domain (further discussed Section 3.3.5), they have been found to be less effective under increasing distribution shift [19, 348].

Bayesian Deep Learning (BDL) methods build on solid mathematical foundations and hold promise for more reliable learned uncertainty estimates [496]. Drawing on the ground-laying works of [91, 179, 299, 300, 337], the "second-generation" in BDL [140] is geared towards finding practical and scalable approximations to the analytically intractable Bayesian posterior (Equation (3.1)). Inferring a prediction and the associated uncertainty for a new test input x^* (with its associated label vector y^*) requires computing the

conditional probability of x^* given the training data $\mathcal{D} = \left\{\left(x^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$,

$$P(y^* \mid x^*, \mathcal{D}) = \int P(y^* \mid x^*, \mathcal{D}, \theta) \underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} d\theta, \tag{3.1}$$

with θ representing all Bayesian Neural Network (BNN) parameters: weights w, biases b.

In our study we will focus on two strategies with representative methods that circumvent the *inference problem* and have seen more widespread adoption given their ability to scale both in network architecture and dataset size.

I. **The weight snapshots direction**, *Deep Ensemble* [238], which aims to find different sets of model parameters. Snapshots can be collected during different stages of training [133, 186, 301], or by using a sampling process such as Markov Chain Monte-Carlo (MCMC) [141, 180, 530]. II. **The stochastic computation-graph direction**, *Monte Carlo Dropout* [124], involves introducing noise over weights during training and estimating uncertainty with multiple stochastic forward passes. Recent works [283, 464] have proposed "single-model" uncertainty methods that ideally compute posterior uncertainty in one forward pass.

Our work benchmarks representative methods from both categories (denoted by cursive), motivating a cross-category comparison and analyzing their individual-joint effectiveness in modeling predictive uncertainty.

Additionally, we later experimented with alternative scalable uncertainty methods, namely stochastic gradient MCMC methods, *cyclical SG-MCMC* (cSG-MCMC) [530], and a single forward pass uncertainty method incorporating a Gaussian Process (GP) output layer, *Spectral-normalized Neural Gaussian Process* (SNGP) [283]. Results and discussion for these are included as a self-contained subsection Section 3.7.

3.3.2 Predictive Uncertainty Methods

We will first introduce each method by explaining the algorithm, followed by advantages or identified shortcomings, with subsequent method extensions from the same procedure category. Finally, we will zoom in on how to quantify uncertainty using each method.

3.3.2.1 Monte Carlo Dropout

The seminal work of [124] on Monte Carlo Dropout (MC Dropout, MCD) proposes efficient model uncertainty estimation by exploiting dropout regularization as an approximate Variational Inference (VI) method. In practice, the MCD procedure boils down to (i) applying dropout on all non-linear layers' weights, and (ii) activating dropout both during training and evaluation. Quantifying "epistemic" *model uncertainty* using MCD involves sampling T stochastic weight sets from the variational Bernoulli distribution $\hat{\theta}_t \sim q(\boldsymbol{\theta})$ to calculate the lower-order moments of the approximate Gaussian posterior, respectively the predictive mean and variance (Equation (3.2)).

$$\begin{aligned} \hat{\mu}_{pred}(x^*) &= \frac{1}{T}\sum_{t=1}^{T} P(y^*|x^*, \hat{\theta}_t), \\ \hat{\sigma}^2_{pred}(x^*) &= \frac{1}{T}\sum_{t=1}^{T} [P(y^*|x^*, \hat{\theta}_t) - \hat{\mu}_{pred}]^2 \end{aligned} \tag{3.2}$$

MCD's simplicity and computational tractability, i.e., dropout training is a standard DL practice and prediction only requires 1 model to sub-sample from, has made it one of the most popular predictive uncertainty methods. However, an important shortcoming of VI, and in consequence MCD in [124]'s formulation, is that it is known to underestimate predictive variance [459]. We will touch on a selection of method extensions in Sections 3.3.2.3 and 3.3.2.4.

3.3.2.2 Deep Ensemble

Deep Ensemble [238] (DE) involves independently training multiple NNs with different random weight initializations and aggregating predictions from individual models. An ensemble of NNs trades off computational resources, due to the need to train and store M models, for uncertainty estimation and robustness to dataset shift [163, 348, 489]. In comparison to MC Dropout, DEs are treated as a uniformly-weighted Gaussian Mixture model, to which the formula for predictive variance is adapted:

$$\begin{aligned} \hat{\sigma}^2_{pred}(x^*) &= \frac{1}{M}\sum_{m} \left(\sigma^2_{\theta_m}(x^*) + \mu^2_{\theta_m}(x^*)\right) - \mu^2_*(x^*), \\ \mu_*(x^*) &= \frac{1}{M}\sum_{m} \mu_{\theta_m}(x^*) \end{aligned} \tag{3.3}$$

The empirical performance increase of ensembles can be attributed to the diversity of uncorrelated errors between ensemble members [225]. Without functional diversity in sets of model parameters, posterior approximation quality will be lower (zero variance) and for this reason, ensemble diversity promotion is a promising avenue for further improvements [49, 196]. Alternatively, the interplay between ensembling and regularization, "the effect of a prior", warrants more thought, since not regularizing risks overfitting, while too strong regularization risks constraining diversity (see Section 3.3.4).

3.3.2.3 Concrete Dropout

[125] proposes a **Con**tinuous-dis**crete** distribution relaxation to adapt and optimize the dropout probability p as a variational parameter using standard gradient descent. This overcomes the limitations of uncertainty underestimation, miscalibration, and the computational complexity of manually tuning layer-wise dropout probability in deeper models [345]. By taking advantage of the reparametrization trick, the Concrete distribution approximation $\tilde{z}$ of the original Bernoulli random variable z conveniently parametrizes to a simple sigmoid distribution (ϕ = sigmoid) allowing for gradient-based optimization. Given a uniform random noise variable u and a temperature r, the expression varies with respect to the dropout probability p, which for $p \to 0.5$ produces by a rate of $\frac{1}{r}$ values approaching 1.

$$\tilde{z} = \phi \left(\frac{1}{r} \left(\log p - \log(1-p) + \log u - \log(1-u) \right) \right) \tag{3.4}$$

Since the dropout probability characterizes the overall posterior uncertainty, Concrete Dropout can positively influence in-domain calibration at an almost negligible cost.

3.3.2.4 Heteroscedastic Extensions

[213, 236, 500] proposed similar approaches to extend MC Dropout to allow measuring uncertainty information from different sources. Estimating input-dependent, "heteroscedastic aleatoric", *data uncertainty* (detail Section 3.3.3) requires slightly modifying the model's architecture and objective function following [213].

Firstly, the output layer of model $f_{\hat{\theta}}$ is extended with a set of learnable variance variables $\boldsymbol{\sigma}^2$ per unique class output. The model's output logits, $\mathbf{v}$, are sampled from the stochastic output layer parametrized by $\mathcal{N}(f_{\hat{\theta}}(x), diag(\boldsymbol{\sigma}^2(x)))$. This

model adaptation will be referred to as the *heteroscedastic model*. *Fig.* 3.1 visualizes the difference in output layer design.

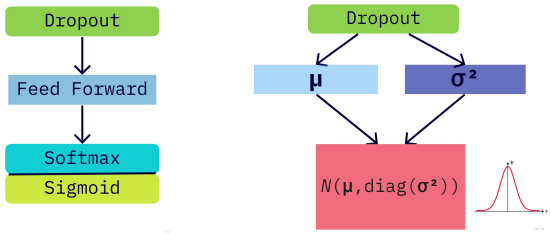


Figure 3.1. Visualization of output layer blocks. The left block denotes standard *softmax* (multi-class) or *sigmoid* (binary/multi-label) output. On the right, the *heteroscedastic* model outputs a normal distribution $\mathcal{N}(\boldsymbol{\mu}(x), diag(\boldsymbol{\sigma}^2(x))$ parametrizing mean and variance by the logits coming from two separate preceding feedforward layers.

Next, it requires incorporating a *heteroscedastic loss*:

$$\mathcal{L}_{\text{HET}}(\hat{\theta}) = \sum_{i=1}^{N} \log \frac{1}{T} \sum_{t=1}^{T} \exp \left(\mathbf{v}_{i,c}^{(t)} - \log \sum_{k}^{K} \exp \mathbf{v}_{i,k}^{(t)} \right) + \log T \quad (3.5)$$

with N the number of training examples passing through an instance t of the model $f_{\hat{\theta}_t}(x) + \boldsymbol{\sigma}^{(t)}$ ((^{2} omitted for sampling superscript) to generate for example i a sampled logit vector $\mathbf{v}_i^{(t)} \in \mathbb{R}^K$, where predicted value for class k, $\mathbf{v}_{i,k}^{(t)} \in \mathbb{R}$, and c the index of the ground truth class. The above loss formulation shares notation with a categorical cross-entropy objective, although the loss is computed over T sampled logits $\mathbf{v}_i^{(t)}$ perturbed with parameterized Gaussian noise. By learning to predict log variance over T dropout-masked samples, the model will be able to output high variance (uncertainty) for inputs where the predictive mean is far removed from the true observation, which by design has a smaller effect on the total loss.

3.3.3 Uncertainty Estimation

In this Subsection, we will introduce sources of uncertainty, a categorization of uncertainty measures, and how uncertainty is quantified in practice.

Total Uncertainty Classification models trained by minimizing negative log-likelihood quantify global uncertainty over class outcomes with entropy (H) over

logits. Therefore, the entropy of the posterior predictive distribution provides a measure of the total uncertainty, which is a combination of model and data uncertainty [190]. Instead of entropy, posterior predictive variance can also be decomposed into model and data uncertainty using the law of total variance [92]. Decomposing total uncertainty into the different sources is beneficial for determining actions to evaluate the room for improvement.

Model Uncertainty *Epistemic uncertainty* presents the inherent ignorance [345] of the model with regards to the true values for its parameters and structure after having seen the training data. Next to predictive variance, *Mutual Information* (MI) [426] has been proposed as a measure of epistemic uncertainty, as intuitively it captures the amount of information that would be gained about model parameters through "knowledge" of the true outcome [305].

Data Uncertainty *Aleatoric uncertainty* captures the inherent stochasticity and noise in data. It can be further decomposed into a *homoscedastic* component, which represents constant noise over inputs such as the numerical accurateness of a measuring device, and *heteroscedastic* uncertainty representing input-dependent noise generated by class overlap, complex decision boundaries or label noise [92]. Heteroscedastic data uncertainty allows for the expression of instance-level uncertainty together with the best possible prediction.

Uncertainty categorization Here follows a categorization of the uncertainty measures from methods (and combinations) of Section 3.3.2. We directly provide estimators for the theoretical quantities that are defined as either arising from entropy or variance-based uncertainty decomposition in [92]. To estimate for a new test sample x^* the prediction and uncertainty of model $f_{\hat{\theta}}(x^*)$ we typically seek to obtain the predictive posterior distribution $P(y^*|x^*, \hat{\theta})$ over class membership probabilities with $y_k^* \in \{1, \ldots, K\}$.

For MC Dropout at inference time, we presume $P(y^*|x^*, \hat{\theta}) \approx \frac{1}{T}\sum_{t=1}^{T} P(y^*|x^*, \hat{\theta}_t)$, with prediction obtained after applying softmax/sigmoid function for sample t, $\hat{p}_t = P(y^*|x^*, \hat{\theta}_t)$. For Deep Ensemble, the above notations would require a change from T to M, but for consistency over quantity formulas, we maintain T to denote posterior sampling. For ease of notation, we define a helper entropy function on $H(x^*, \cdot) = -\sum_{k=1}^{K} P(y_k|x^*, \cdot) \log P(y_k|x^*, \cdot)$ with $\cdot$ an input argument

to the function.

Quantity	*Formula*
Softmax-score	$S = \max_k \frac{\exp f_{\hat{\theta},k}(x^*)}{\sum_{j=1}^{K} \exp f_{\hat{\theta},j}(x^*)}$
Predictive Entropy	$H_{pred} = H(x^*, \hat{\theta})$
Mutual Information	$I = H_{pred} - \frac{1}{T}\sum_{t=1}^{T} H(x^*, \hat{\theta}_t)$
Model Uncertainty	$\hat{\sigma}^2_{model} = \frac{1}{T}\sum_{t=1}^{T} (\hat{p}_t - \hat{\mu}_{pred})^2$
Data Uncertainty	$\hat{\sigma}^2_{data} = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{K}\sum_{k=1}^{K} \text{var}_k^{(t)}(x^*)$

For any classification model, it is possible to compute the softmax-score and predictive entropy. For multi-label classification, the softmax-score does not take into account multiple winning classes and a standard approximation[2] would be to average over the sigmoid-scaled probabilities of predicted classes.

Model uncertainty can be quantified with Monte Carlo integration or the aggregation of individual models [461]. In practice, it is quantified by either (a) calculating the average sigmoid/softmax variance over the predictive mean from MC samples (Equation (3.2)) or (b) computing the total variance from an ensemble mixture distribution (Equation (3.3)). Changing to the heteroscedastic extensions allows to quantify data uncertainty. More specifically, data uncertainty is quantified with as "surrogate" [500] the average over variance logits $\text{var} = \boldsymbol{\sigma}^2$ (see *Fig.* 3.1). Whenever ensembling is applied where a single model estimates a quantity, one typically averages over the ensemble components' uncertainty.

[2]Intending to compare directly with multi-class results, averaging uncertainty estimates to obtain a single summary statistic for multi-label predictions is more straightforward than reporting class-wise results. In particular, the tested multi-label datasets share low average label cardinality, a high degree of label correlation, and a large set of unique classes ($K > 50$).

3.3.4 Motivating Hybrid Approaches

This Subsection will motivate the theorized complementarity of VI-based and ensembling methods for improved uncertainty estimation and robustness.

In light of the empirical success of Deep Ensemble, recent research [118, 496] raises an important question concerning the difference in function-space between variational Bayesian NNs (MC Dropout and extensions) and Deep Ensemble. Deep NNs are parametrized (typically non-linear) functions presenting a high-dimensional non-convex optimization problem, which may concern widely varying curvature and many flat regions with multiple locally optimal points within each [255]. Applying an optimization procedure to a maximum-a-posteriori (MAP) objective involves a search for parameter values (*hypotheses*) for which the loss function is low by navigating the high-dimensional loss landscape. Once model training converges, one ends up with a weight-space *solution*, representing a single *mode* of the parameter posterior . One such mode is a local optimum of the loss function $\mathcal{L}(\theta)$, representing unique functions f_θ as a set of NN parameters [133]. Each mode potentially marks a meaningfully different representation of the data.
The true posterior is generally a highly complex and multimodal distribution, with multiple possible but not necessarily equivalent parametrizations θ able to fit the training data. To accurately quantify posterior uncertainty, we wish to capture as many modes or separated regions as possible [117, 496].

Correspondingly, the common goal is to achieve reliable uncertainty and, following the BDL paradigm, one resorts to modeling a Bayesian posterior. What differs among the selected predictive uncertainty methods, is the form of the prior $P(\theta)$ over model parameters and likelihood $P(\mathcal{D}|\theta)$ [336], from which to determine a procedure. Below we expound on the **difference in posterior approximation procedure**:

- MC Dropout is a common VI procedure with Bernoulli dropout and Gaussian (L2) priors on weight-space, assuming a posterior Gaussian distribution from which to draw stochastic samples. VI-based methods tend to locally approximate uncertainty surrounding a single mode, **intra-modal** posterior approximation. Specifically, MC Dropout's procedure can be interpreted as imposing a spike-and-slab parameter prior with peaked variance [333], which offers a plausible explanation for approximated uncertainty centered tightly around 1 mode.
- An ensemble of NNs makes no direct assumptions on the form or distribution of the prior and just "obtains" different samples from the parameter posterior. It generates a series of MAP estimates which through inherent stochasticity in weight initialization and optimization end up at different regions in weight space, leading to functionally dissimilar but more or less equally accurate

modes of the solution space. Due to randomness in the optimization, some solutions may be significantly worse than others as measured by different metrics (*e.g.*, accuracy vs. calibration). Ensembles are effective at exploring the weight-space and by solving the MAP estimation problems converge to multiple modes [117, 149], allowing for **inter-modal** posterior approximation. Furthermore, by considering more possible hypotheses they will be better at approximating multimodal posterior distributions and avoid the collapse to a single mode [496].

Combining both procedures is to generate a mixture over priors [119], which in itself is again a prior, all under the same likelihood function. There is no guarantee that a combination of methods from both procedures captures the true posterior, yet in our work we will empirically analyse if combining inter and intra-modal posterior approximation offers the hypothesized complementary benefits.

3.3.5 Uncertainty Calibration under Distribution Shift

In this Subsection, we motivate the meaningfulness of evaluating uncertainty methods under distribution shift and what restricted assumptions one should reasonably specify to guarantee useful empirical results.

We consider the problem of detecting out-of-distribution data from a trained classifier's uncertainty. Let $P^S(x, y)$ and $P^T(x, y)$ denote two distinct distributions, respectively *in-domain* and *out-of-domain*. Further we assume the classifier $f \rightarrow [0, 1]$ trained on P^S, whereas in the experimental setup we test on a mixture distribution $\mathbb{P}^{(S,T)}(x, y)$. Given an input x from the mixture, we test if the classifier's uncertainty can be exploited to distinguish from which distribution the sample comes. To be clear, in this setting we expect to detect uncertainty arising from distribution shift and not from a lack of training data. It can be argued that there is a relationship between both, as having few in-domain samples complicates generalization, in turn increasing the chance of flagging a new data point as OOD.

Uncertainty estimation is generally well-defined in the context of in-domain data with the standard assumption that samples are independent and identically distributed (i.i.d.). In this setting, evaluation is typically expressed in terms of **calibration** (Definition 8), particularly as statistical error with respect to the conditional expectation (Definition 7).

To obtain a reliable probabilistic classifier in the traditional i.i.d. setting, explicit in-domain re-calibration approaches are effective [156, 229, 490]. However, there

is no general principle which states that a classifier, however calibrated on P^S, would be calibrated on OOD data from P^T. Infinitely many possible shifts can violate the standard i.i.d. assumption at varying degrees of severity, affecting calibration and uncertainty estimation in unpredictable ways. With the aim of still being able to rely on a classifier's uncertainty calibration to predict future generalization, there is a need to relax the i.i.d. assumption. An important condition for meaningful uncertainty estimation is to impose realistic, yet sufficiently restrictive assumptions on the nature of distribution changes and how P^S and P^T relate. The **covariate shift** [34, 418] assumption may be the most widely studied when the real-world data distribution differs from the training distribution.

Recently, [354] formalized the problem of calibrated prediction under covariate shift with theoretical bounds on calibration transfer over domains. Critically, related works [104, 145, 335, 349, 483] prove with importance weighting that shared structure and high overlap in distribution support (or conversely, low domain divergence) is crucial to upper bound the increase of calibration error due to covariate shift. To put it plainly: while one cannot guarantee calibration on OOD data in the general case, if domains are reasonably close one can expect to retain (some if not most) benefits from in-domain calibration.

Specific to our work, we consider two experimental settings (Section 3.4.5) with different distribution shift [320] between domains. Here we characterize each with the related distribution shift assumptions. (i) *Cross-domain classification*, where covariates differ $P^T(X) \neq P^S(X)$, but label distributions are identical $P^T(Y|X) = P^S(Y|X)$ [418]. (ii) *Novelty detection*, where label distributions disagree $P^T(Y|X) \neq P^S(Y|X)$, since the label sets differ between domains $[Y]^T \neq [Y]^S$ [307]. Whereas (i) is a clear case of covariate shift, we reasonably assume for (ii) that covariates are generally close $P^T(X) \approxeq P^S(X)$ and that the overall conditional shift will be small. Rather than interpreting novelty as a shift in label sets, one might define the probability of seeing some labels under S as exactly zero, while under T their probability is $\varepsilon > 0$. In practical text classification settings, novel class inputs will typically start occurring with small frequency in the real-world data distribution, as well as not having completely different syntax and semantics. This implies that 'excess' calibration error (defined as an expectation over the mixture) will only be impacted slightly.

Clearly specifying distribution shift assumptions is quintessential for reliably benchmarking uncertainty methods, since the calibration of each tested method can be affected in different ways and produce results biased towards an evaluation configuration. In our selected experimental settings, we can justify uncertainty calibration under distribution shift as a reasonable methodology, without making further claims on the general applicability of this evaluation procedure.

3.4 Experimental Methodology

In this work, our objective is to reliably benchmark both existing and novel combinations of predictive uncertainty methods in order to draw conclusions for text classification applications. This Section describes our study's experimental methodology with which we generate the empirical evidence presented in Section 3.5. Section 3.4.1 introduces our hypotheses on complementary benefits for uncertainty estimation and details the hybrid methods. Provided the focus on text classification tasks, Section 3.4.2 motivates a set of representative datasets, with a specification of different text problem characteristics. Section 3.4.3 documents two pre-selected text classification architectures, the first a simple and more controllable configuration for uncertainty benchmarking, the second a more complex NLP architecture for which we will compare relative gains in robustness. To ensure correct performance benchmarking, Section 3.4.4 summarizes the metrics used for evaluating calibration and robustness. Finally, Section 3.4.5 expounds on the model setups and experimental settings devised to compare predictive uncertainty methods.

3.4.1 Proposed Hybrid Approaches

This Subsection stands central in our work in which we motivate combinations of predictive uncertainty methods. We build hypotheses on complementary benefits from combining multiple uncertainty methods, for which we present an overview of hybrid methods in scope of our experiments (*Table* 3.1).

Given the obvious parallels and differences between both procedures presented in Section 3.3.4, we hypothesize **complementary benefits** for uncertainty estimation and robustness.

A. Whereas ensembles are adept at capturing multiple modes, they do not approximate uncertainty surrounding a single mode in solution space. However, since there is a lot of redundancy in function space, local neighborhood uncertainty approximation might make only a minimal contribution to the overall posterior uncertainty. [118] validated that applying subspace sampling on an optimized solution improves in-domain accuracy and calibration. They note improvements relatively lower than increasing ensemble size (M), yet they did not analyze for joint effectiveness.

B. A procedure can only be as good as the prior and the likelihood function, which in approximation of the intractable parameter posterior is limited by computational constraints (number of MC samples T, number of ensemble

models M). By lack of any specific prior constraining the optimization of independent ensemble members, the regularization effect from VI-based priors such as dropout may introduce smoothness [110, 369], inducing a simpler optimization landscape with less (possibly weak) hypotheses present. In turn, by modeling an ensemble of VI approximate posteriors less ensemble members could be required to reach the same in/out-of-domain performance as measured by the size and quality of captured solutions. [118] already observed that ensembles saturate after reaching peak in-domain performance, with suboptimal models taking over the benefit.

C. Important to note is that the influence of the prior and variational parameters requires fine-tuning, since over-regularization will reduce the optimization problem to one with an over-smooth, possibly unimodal landscape [117, 133]. This eliminates any functional diversity for whatever ensemble size, where the solution will be overconfident. Alternatively, since the hypothesis space for a NN is often so large, with many possible likely models for finite data, that some posterior collapse will often be desirable to reduce the number of considered hypotheses. [496].

Table 3.1 summarizes all model setups and hybrid methods considered for our experiments. The most complete combination is *MC Concrete Dropout Heteroscedastic Deep Ensemble*, where each member m of the ensemble has optimized the layer-wise dropout rate p and heteroscedastic loss $\mathcal{L}_{\text{HET}}$, with the final predictive distribution over K classes deriving from M times T stochastic MC Dropout samples (M x T x K).

Table 3.1. In total, we consider 18 model setups, based on combining methods and options from each column. (*) Deterministic dropout can only combine with Deep Ensembles. CE stands for cross-entropy loss.

Dropout	**MC sampling**	**Heteroscedastic**	**Deep Ensemble**
$p = 0^*$	$T = 1$	$\mathcal{L}_{\text{CE}}$	$M = 1$
$p = 0.5$	$T = 10$	$\mathcal{L}_{\text{HET}}$	$M = 5$
Concrete			

We admit two baselines, *Unregularized* and *Regularized*.
Unregularized ($p = 0$) offers a clean comparison, discounting any influence of sparsification (dropout) or normalization of weight magnitude (weight decay). However, it possibly overfits parameters to training data. In practice, one would always apply some combination of regularization (dropout, weight decay, batch normalization, data augmentation, ...) to counter overfitting. *Regularized*

($p = 0.5$) gives an alternate point of comparison over uncertainty methods, such that we can exclude that performance increase for an uncertainty method does not only come from regularization, which some such as MC Dropout rely upon.

Adhering to good practices and since we build ensembles with default $M = 5$, we report the mean (and standard deviation) for all individual models, making the results more statistically reliable than comparing to 1 independently trained model.

3.4.2 Datasets

We use six well-studied real-world text corpora characterized by a different number of classes, classification task, and size of the documents *(Table 3.2)*.

Table 3.2. D denotes the number of documents in the dataset, K the number of classes, I the class imbalance ratio [444], W the average number of words per document, V the total vocabulary size respectively.

corpus	task	D	K	I	W	V
20news	newswire topic	18,848	20	5e-4	240	212,267
IMDB	movie review	348,415	10	0.03	325.6	115,073
CLINC-OOS	intent detection	22,500	150	0	8	6,188
Reuters ApteMod	newswire topic	10,786	90	0.14	125.2	65,035
AAPD	academic paper subject	55,840	54	0.04	145.4	66,854
Amazon Reviews (#4)	product sentiment	8,000	2	0	189.3	21,514

The first three datasets share the task of multi-class classification in three different text domains.

`20News` [239] is a collection of 20K newsgroup documents with balanced samples for 20 different newsgroups. To allow for direct comparison, we use the dataset in the benchmark format of [172].

`IMDB movie reviews` [97] (imdb) is a large sentiment classification dataset which links user-based reviews of movies with labels on an ordinal scale between 1 and 10. Since there are no standard splits for this dataset we generate randomized (seed 42) stratified splits of 65% for training, 15% validation and 20% for testing.

`CLINC-OOS` (CLINC150) [240] is a recently become popular intent detection dataset comprising 150 training sentences for each of the 150 system-supported services. Next to this, it offers a separate Out-of-Scope (OOS) subset with 1200 natural sentences which can be used for Out-of-Domain (OOD) detection, more specifically detecting novel class instances. This dataset differs from the previous two through very short "intent" sentences requiring classification in a large output space. For training and evaluation, we use the predefined splits of TensorFlow Datasets.

We include two popular multi-label text classification datasets, since they are often not considered for uncertainty experiments. We argue that they should be included since their multi-label nature is very common in text classification where not all labels have to be mutually-exclusive, *e.g.*, topic categorization, subject attribution, ...

`Reuters ApteMod` [17] is a multi-label news topic categorization dataset with 90 possible topics and an average low label cardinality (C) of 1.24. We use the standard ApteMod splits.

`Arxiv Academic Paper Dataset` (AAPD) [505] comprises 55,840 computer science paper abstracts that have been labeled with corresponding multiple subject matters. Each academic paper has on average 2.41 subject targets with a minimum of 2. For reproducibility purposes, we use the same preprocessing steps and splits as in [5, 505] with 1K dev and 1K test samples.

`Amazon Reviews` [45] is a widely-used benchmark for domain adaptation research in NLP. It consists of binary sentiment classification datasets from four different domains: Books, DVDs, Electronics and Kitchen appliances. Each domain dataset contains 1K positive and 1K negative labeled instances. Following the convention of previous works [103, 557], we construct 12 balanced cross-domain sentiment analysis tasks, where for each source dataset we randomly hold out 400 test instances to evaluate in-domain and always predict on the full target dataset. We reserve this dataset for cross-domain experimentation only (Section 3.4.5.2).

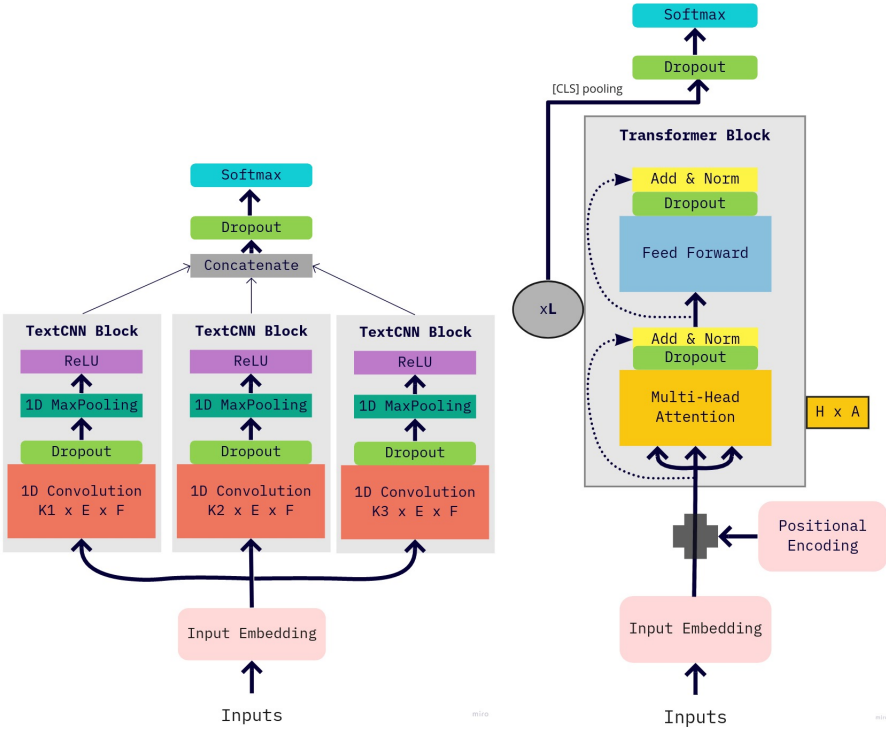
3.4.3 Architecture

This Subsection motivates the two NLP architectures in scope for the experiments.

TextCNN architecture We use a 1-D Convolutional NN for text classification (TextCNN), following the model structure of [218]. We chose this architecture for its comparative simplicity and solid out-of-the-box performance on a range of text classification tasks. Even as a light-weight model, it can deal with feeding in text sequences of varying sizes and learning n-gram-like structures over word embeddings, allowing a fair comparison across text datasets. An extensive hyperparameter study determined that regularization does not impact performance much [537].

Transformer architecture Models in NLP have become increasingly deeper and more complex with the advent of the Transformer architecture [473]. [94] have combined multiple bidirectional Transformers with wordpiece tokenization and self-supervised pretraining objectives —masked language modeling and next sentence prediction— to create the contextual representation modeling

architecture BERT. It allows for fine-tuning on downstream tasks where BERT has outperformed task-specific architectures even in low resource settings. In our experiments we use BERT_{base} (uncased, English): 12 layers, 768 hidden dimensions, 12 attention heads, with a total number of 110M parameters.



(a) TextCNN (b) BERT

Figure 3.2. Simplified block-diagrams for each of the NN architectures, demonstrating on which layer weights dropout is applied.
(a) The TextCNN model architecture with 3 kernels ($K1-3$), E word embedding dimensionality and F number of feature maps per kernel.
(b) The BERT model architecture with L Transformers blocks, hidden size H and number of self-attention heads A.

Complexity TextCNN comprises only 6M parameters with most parameters residing in the embedding matrix. However, it is restricted to a fixed window size with the downside of not being able to determine long-distance dependencies in text. BERT, on the other hand, has already captured prior language modeling knowledge thanks to pretraining. Nevertheless, our experiments already involve

significant computational complexity, which is why we decided not to run all variations with BERT. TextCNN presents a more controllable configuration, achieving decent performance and satisfying for the evaluation of predictive uncertainty in text classification. We include an ablation study (Section 3.5.4.2) comparing specifically selected models trained with BERT as base architecture.

3.4.4 Evaluation metrics

Since no single metric measures all desirable properties of predictive uncertainty, we use a variety of conventional metrics to evaluate our models' performance, (a) *calibration metrics*, b) *proper scoring rules* and c) *classification scores*.

The metrics are defined in detail in Section 2.2.3, here we will only provide a brief description.

For **in-domain evaluation**, we use the following metrics: (*a*) **Expected Calibration Error (ECE)** [156, 332], (*b*) **Brier Score** [50] and (*b*) **Negative Log Likelihood (NLL)** [378]. We use the same metrics for **out-of-domain evaluation**, with the addition of (*c*) **AUROC** and (*c*) **AUPR** for distribution shift detection following [172].

When evaluating a model trained in a source domain on a target domain with a similar task, we denote accuracy in the target domain as **OOD accuracy** as opposed to accuracy in the source domain, which we denote as **ID accuracy**.

3.4.5 Experimental design

We have determined three logical settings in text classification to evaluate predictive uncertainty for each model setup. We present experiments on in-domain uncertainty to form baseline results, followed by cross-domain classification with a focus on out-of-domain detection, and finally we propose novelty detection as a new protocol to evaluate predictive uncertainty.

While there is no gold standard procedure for comparing multiple (uncertainty) methods over multiple (text classification) datasets, we opted for an established procedure with statistical testing via multiple comparisons [89, 109]. Since we present an exhaustive list of model setups, we present our results in terms of rank and critical difference diagrams in order to analyze relative performance of each method over different experimental settings.
Concretely, each dataset concerns independent measurements, for which we rank each method, then compare average ranks, and in the event that we can

reject the null-hypothesis ($\mathcal{H}_0$: all methods have the same rank), we calculate post-hoc tests with critical differences over methods. However, only reporting ranks does not allow future researchers to compare to our work, which is why we include detailed absolute number results in the Appendix C.

3.4.5.1 In-domain Setting

To evaluate in-domain (ID) uncertainty, we will focus on measuring calibration and prediction quality with proper scoring rules (see Section 3.4.4). The ID setting assumes that the train and test examples are i.i.d.. To capture all details, we compare per task-setting, multi-class and multi-label, and finally zoom in on dataset-specific observations. For the in-domain evaluation, we focus on unique contributing effects per predictive uncertainty method and the relation between method combinations and evaluation metrics.

- When evaluating with proper scoring rules, does an absolute increase in combination size (higher T or M) correlate with better performance?
- What effect —equal over all tasks, datasets or architectures— can be discerned per unique predictive uncertainty method?

3.4.5.2 Cross-domain Setting

Since we test over sentiment classification datasets from multiple domains (Amazon Product Reviews), we seek to analyze uncertainty reliability across domains. However, learned knowledge from a source domain can often transfer to classification in the target domain. Provided this setting we need to account for cross-domain generalization next to out-of-domain detection, the latter which is the focus of our experiments.

Cross-domain generalization - *how well does a classifier trained in a source domain perform on a dissimilar target domain sharing a similar task?* The aim of cross-domain generalization is to learn a robust classifier, which can perform well in multiple domains even if there is limited labeled data in some of the domains. Domain discrepancy is a major challenge where, for instance, linguistic sentiment expressions used in one domain can be different from that of the source domain. For example, “garbage disposal” is neutral in kitchen appliances whereas a “garbage movie” is strictly negative. This domain discrepancy challenge is often approached by adaptation [497, 557] or encouraging domain-agnostic feature representations [103, 129]. We propose to test out-of-domain detection with predictive uncertainty as a viable fallback strategy when achieving generalization over domains is difficult.

Out-of-domain detection - *how reliably can a classifier trained in a source domain communicate uncertainty in a target domain provided good/bad generalization?* Whenever a model does not generalize to OOD examples, we would expect a model to be uncertain, allowing detection in order to abstain or trigger conservative fallback strategies [108]. As a proxy to good/bad generalization we measure the gap between in-domain and target domain accuracy as evidence of train-test skew. We argue that our current setting is more realistic than benchmarking OOD detection in totally disparate domains such as evaluating a newswire classifier on movie reviews.

Our analysis will be centered on the following question:

- How does domain similarity affect out-of-domain detection with uncertainty methods? Is there a clear increase of uncertainty given a higher OOD generalization gap?

3.4.5.3 Novelty Detection Setting

Novelty detection - *how well can the model identify and communicate uncertainty on samples of a novel class?* In the worst case, classifiers "fail silently" and wrongly attribute high confidence to an in-distribution class [11, 146]. In the best case, the model either lowers its confidence or signals uncertainty. Prior work hypothesizes model uncertainty to be mostly impacted [213, 250].
With this experiment we simulate the conditions of novel class data by removing a single or multiple classes during training. The resulting distribution shift is not too far from the original domain and cannot be considered fully out-of-distribution (as detailed in Section 3.3.5).

We determine diverse novelty detection strategies adapted per dataset. For `20news`, we follow [172, 348] and take out all odd-numbered classes to simulate novel distribution shift. Since `imdb` is a sentiment classification dataset, we isolate the middle class, rating "5" out of the 10 ratings, from training and expect the models to allocate prediction mass to a label close to the holdout class (ratings "4" or "6"). `CLINC-OOS` provides a separate out-of-scope intents set on which we assess novel class robustness.

We devise a new strategy for the multi-label classification datasets, where we would isolate a class that is very distinct from the remaining classes, i.e., (i) by not appearing often in the originally multi-label annotated dataset jointly with the remaining classes, and (ii) occurring frequently enough to guarantee representative results. We draw statistics on the label co-occurrence rates of each dataset, and find that for `Reuters` "Acquisitions" (id:0) occurs in 94% of

documents as a single topic, making it an ideal candidate for testing novel class detection. For `AAPD` we apply the similar strategy and find the frequent label "CS.it" (id:0) to have relatively low label- co-occurrence (2.49), even when there are at least 2 labels to be predicted per sample. We isolate all examples where the novel class appears, either alone or in combination with other labels.

We focus our analysis around three specific questions concerning predictive uncertainty under distribution shift, and compare generally to other modality benchmarks:

- Do hybrid predictive uncertainty methods incrementally or critically improve detection of unseen class instances?
- Does calibration in the in-domain setting translate to calibration under distribution shift?
- Do we see the same trends as in benchmarks from different modalities (Section 3.2)?

3.5 Results

We will present the experimental results in a step-wise manner to avoid confusion on the conclusions to be drawn. We start with general and task-specific trends observed for the in-domain setting, followed by the distribution shift experiments, cross-domain classification and novelty detection. Finally, we present 4 ablation studies on critical, learned or empirically set hyperparameter values.

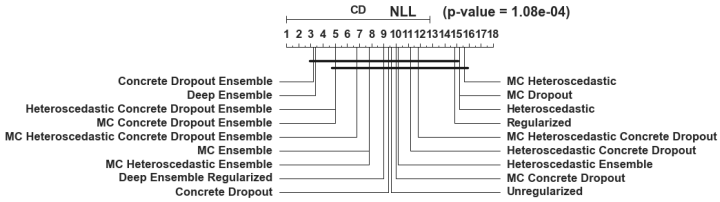


Figure 3.3. *In-domain* results with critical difference diagram comparing all methods by average rank, with the calculated critical difference in the top-left and Friedman χ^2 p-value top-right. *Concrete Dropout Ensemble* achieves the highest NLL rank. While comparing over 5 datasets, the critical difference is large, with only the two aforementioned methods significantly differing from MC Dropout.

3.5.1 Experiment: In-domain

Naively combining predictive uncertainty methods will not give any absolute performance increase, as proper scoring rules show no correlation (-0.01) with the absolute number of predictive uncertainty methods combined. This requires deeper analysis to identify which singular or hybrid methods do significantly outperform baselines.

First, we visualize **general results** with critical difference diagrams comparing all methods by average ranking over datasets (*Fig.* 3.3). *Critical difference* (CD) can be interpreted as the smallest difference between methods which is likely to indicate a significant improvement. In short, the null hypothesis —there is a significant difference between the methods— cannot be rejected for all methods connected by a dark bar. We also report *Friedman* χ^2 , which is a non-parametric statistical test that considers ranking methods over different attempts, in our case datasets, requiring a minimum of 3 methods in comparison. This test checks whether the measured average ranks are significantly different from the mean rank that is expected under the null-hypothesis.

Table 3.3. *In-domain* (left) combined *Brier* and *NLL* proper scoring rule pairwise comparison counts of wins/draws/losses and (right) *ECE* metric reported for comparing in-domain calibration. For in-domain predictive accuracy, ensembles clearly are superior. Considering only miscalibration, Concrete Dropout generally adds calibration to predicted probabilities. The combination with MC Dropout gives unpredictable ranking results.

	ref	wins	draws	losses
9	Deep Ensemble	142	0	28
12	Concrete Dropout Ensemble	135	1	34
16	Heteroscedastic Concrete Dropout Ensemble	130	4	36
15	MC Heteroscedastic Ensemble	114	2	54
17	MC Heteroscedastic Concrete Dropout Ensemble	114	2	54
11	MC Ensemble	111	3	56
13	MC Concrete Dropout Ensemble	102	0	68
10	Deep Ensemble Regularized	90	1	79
14	Heteroscedastic Ensemble	82	2	86
0	Unregularized	79	4	87
5	Concrete Dropout	77	1	92
7	Heteroscedastic Concrete Dropout	70	3	97
8	MC Heteroscedastic Concrete Dropout	65	2	103
6	MC Concrete Dropout	58	0	112
4	MC Heteroscedastic	40	5	125
2	MC Dropout	39	6	125
1	Regularized	34	0	136
3	Heteroscedastic	30	0	140

	ref	wins	draws	losses
5	Concrete Dropout	68	1	16
12	Concrete Dropout Ensemble	58	1	26
4	MC Heteroscedastic	52	1	32
8	MC Heteroscedastic Concrete Dropout	52	0	33
2	MC Dropout	49	2	34
15	MC Heteroscedastic Ensemble	48	1	36
16	Heteroscedastic Concrete Dropout Ensemble	48	0	37
7	Heteroscedastic Concrete Dropout	46	0	39
9	Deep Ensemble	45	1	39
0	Unregularized	40	2	43
6	MC Concrete Dropout	40	0	45
11	MC Ensemble	38	2	45
17	MC Heteroscedastic Concrete Dropout Ensemble	37	1	47
1	Regularized	32	0	53
3	Heteroscedastic	29	2	54
14	Heteroscedastic Ensemble	27	2	56
10	Deep Ensemble Regularized	24	2	59
13	MC Concrete Dropout Ensemble	23	0	62

Table 3.3 shows more detailed pairwise comparison scores, demonstrating that if both proper scoring rules are considered, plain ensembles and hybrid methods based on deep ensembles are overall superior to single model uncertainty prediction methods. However, the benefit resides more in accuracy than calibration, where some single model predictive uncertainty methods rank higher, specifically *Concrete Dropout*.

For a most complete answer to unique effects per predictive uncertainty method, we need to analyze **dataset-specific results**. Detailed results per dataset and metrics (Appendix C.1 *Fig.* A.1) reconfirm that a method's superiority (i.e., for the whole application domain of in-domain text classification) should not be concluded based on 1 single dataset. Each dataset has specific problem characteristics, which affect method ranking differently at varying magnitudes. However, the comparative performance of each method is not fully dependent on the dataset tested, with Deep Ensemble performing reliably in-domain as evidenced by rank.

3.5.2 Experiment: Cross-domain

This Subsection is dedicated to analyzing predictive uncertainty methods under domain shift. We first present results on cross-domain generalization, followed by a challenging OOD detection setting. Finally, we draw parallels between both settings' experimental results.

We conduct extensive experiments on the benchmark Amazon product review datasets on a total of 12 source-target domain configurations. Each domain is abbreviated by its first uppercase letter: (**B**)ooks, (**D**)VD, (**E**)lectronics, (**K**)itchen. *Fig.* 3.4 reports on the lowest **cross-domain generalization** gap between ID and OOD domain datasets. We observe higher ID accuracy for Kitchen and Electronics, which can indicate a relatively lower complexity of domain sentiment. Importantly, the gap between Kitchen - Electronics and Books - DVD are smallest overall, coinciding with our intuitions on domain similarity. Remarkably, *regularized Deep Ensemble* trained on Book reviews even scores higher accuracy (+1.8%) on its target domain (B→D).

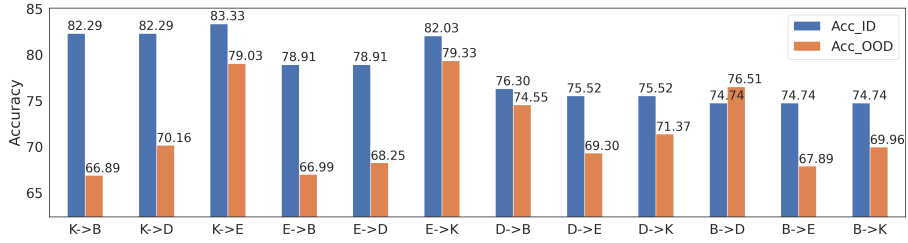


Figure 3.4. Lowest accuracy generalization gap, in-domain (Acc_ID) minus out of domain (Acc_OOD) accuracy (y-axis), of all predictive uncertainty methods per source→target domain combination (x-axis).

To analyze the cross-domain performance of predictive uncertainty methods we report average rank ID NLL and OOD accuracy (*Fig.* 3.5). *Heteroscedastic*

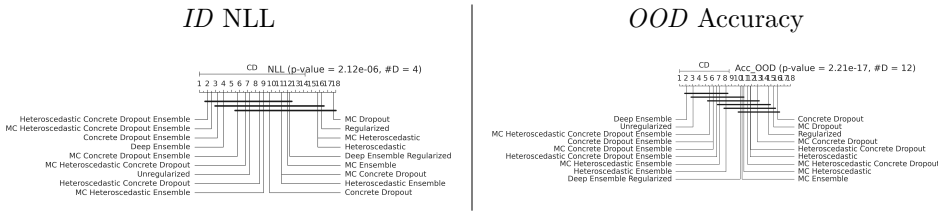


Figure 3.5. Average rank of in-domain NLL for the 4 source datasets (left) and out-of-domain accuracy over 12 source-target configurations (right) for all tested predictive uncertainty methods.

Concrete Dropout Ensemble ranks highest in-domain when evaluated with a proper scoring rule. Models without any regularization achieve higher OOD accuracy scores, with *Deep Ensemble* significantly outperforming more than half of the predictive uncertainty methods (first black bar). A possible explanation could be that most target domain data is more similar to the source domain than expected, effectively giving an edge to methods that achieve high ID accuracy.

To evaluate **Out-of-domain detection**, we report AUROC ranks in *Fig.* 3.6 and additionally plot OOD detection over generalization scores in *Fig.* 3.7. *Concrete Dropout Ensemble* and variations outrank other methods on OOD detection. Nevertheless, we must nuance the ranking results since the magnitude of AUROC is generally low, close to random (50-54%) with no class imbalance, over all 12 cross-domain settings. These results might indicate that from the perspective of the methods tested, there are no salient differences between the different domains. More specifically, Books and DVD as a source have AUROC scores on target OOD domain data centered around 51% and Kitchen and Electronics as a source have comparable AUROC scores with 1 higher AUROC (54%) cluster for OOD Books and DVD targets.

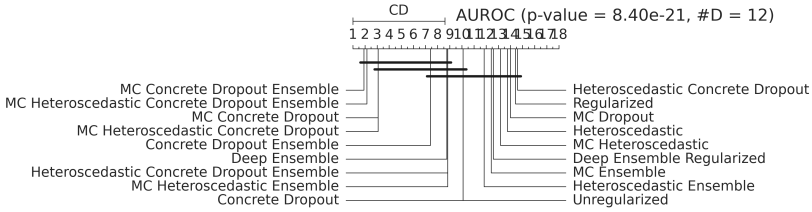


Figure 3.6. Average rank of OOD AUROC over 12 cross-domain settings for predictive uncertainty methods.

Additionally, *Fig.* A.2 in Appendix C.1 demonstrates a similarly clear difference

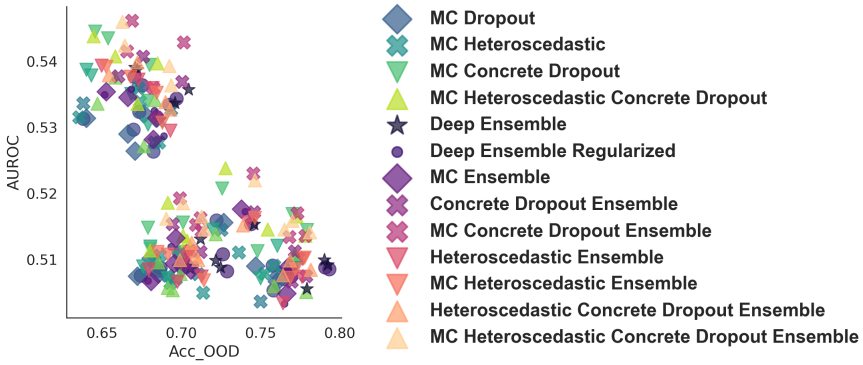


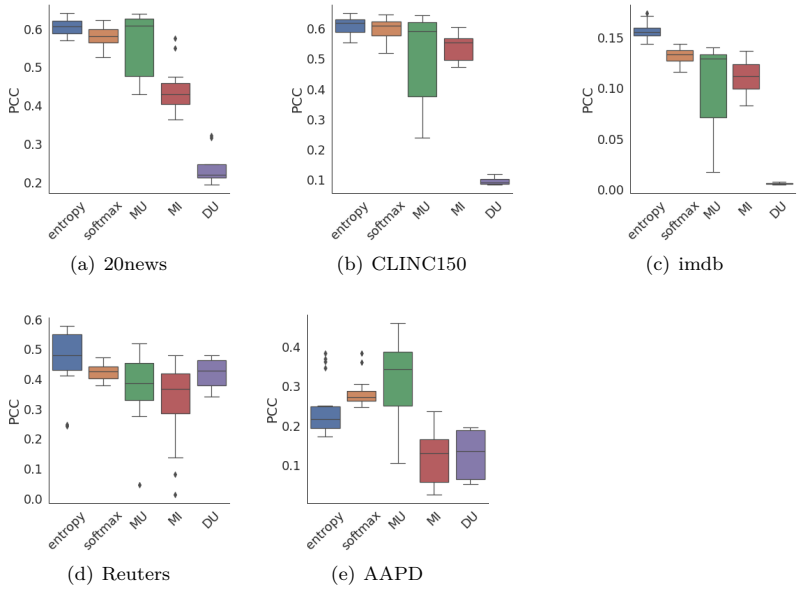
Figure 3.7. AUROC detection magnitude (y-axis) mapped over OOD accuracy (x-axis) with a legend on the right for methods that support uncertainty estimation.

in correlation effect size of uncertainty quantities with ID-OOD data depending on the target domain, *e.g.*, high overall mean correlation (0.3) for Kitchen source evaluated on the disparate domain of Books, whereas uncertainty correlation on Electronics averages around 0.1 for the most correlated quantities.

3.5.3 Experiment: Novelty Detection

Before analyzing which predictive uncertainty methods provide better detection of instances of an unseen class, we report on how uncertainty metrics (cf. Section 3.3.3) correlate with novel class data.
In *Fig.* 3.8 the final rank over datasets confirms the superior robustness of predictive entropy as an uncertainty metric. Logically, it is closely followed by maximum softmax score. Next, model uncertainty correlates generally well with novel class data. Interestingly, model uncertainty outperforms entropy on `AAPD`, with most methods showing the need for learning from more data to better approximate the model parameters.

Similarly to the evaluation of in-domain performance, we use CD diagrams (*Fig.* 3.10) with binary detection metrics *AUPR* and *AUROC* to provide a ranking of predictive uncertainty methods over datasets.
The absolute pairwise comparisons (*Table* 3.9) confirm that hybrid predictive uncertainty methods improve detection of novel class data. Quite surprisingly, *Deep Ensemble* which ranked absolute highest for in-domain, drops multiple ranks in favour of combination ensembles (*Heteroscedastic Ensemble* or even *MC Concrete Dropout*). The in-domain calibration effect from *Concrete Dropout*



(a) 20news (b) CLINC150 (c) imdb

(d) Reuters (e) AAPD

Figure 3.8. We report the Pearson Correlation Coefficient (PCC) between uncertainty values and binary variable ID-OOD for 5 benchmark datasets. Higher absolute correlation score points to stronger association of uncertainty and novelty detection. *Model Uncertainty (MU), Data Uncertainty (DU), Mutual Information (MI).

appears to pass over to this novelty detection setting. More importantly, it also helps boost the novelty detection performance of Deep Ensembles when jointly used (*e.g.*, *MC Concrete Dropout Ensemble*).

While comparing over 5 datasets, there is no critical difference between the average ranking of methods, which can point to task or dataset-specific interactions. *Fig.* 3.11 shows the variation of AUROC performance for the different methods, from which we can observe that (non-finetuned) dropout sampling (*MC Dropout*) under-performs in most datasets, most clearly on `AAPD`, by severly underestimating uncertainty on samples of a novel class. We also observe relative benefits of the *Heteroscedastic* loss function for multi-class text classification, which most clearly is represented in the `CLINC150` results. The same visualization allows us to evaluate the quality of uncertainty quantification for each method. Generally, epistemic uncertainty derived from ensembles offers higher quality detection of novel class data than single model predictive uncertainty. This effect is clearly visible for multi-class classification where the ensembles clearly group on top, as opposed to the results for the multi-label

ref	wins	draws	losses
MC Concrete Dropout Ensemble	121	1	48
Heteroscedastic Ensemble	119	1	50
MC Concrete Dropout	109	1	60
MC Heteroscedastic Ensemble	102	0	68
Deep Ensemble Regularized	100	0	70
Concrete Dropout	90	1	79
MC Heteroscedastic Concrete Dropout Ensemble	89	2	79
MC Heteroscedastic Concrete Dropout	86	1	83
Concrete Dropout Ensemble	83	0	87
Regularized	81	1	88
Heteroscedastic	80	0	90
Deep Ensemble	80	0	90
Heteroscedastic Concrete Dropout Ensemble	75	2	93
MC Heteroscedastic	75	0	95
MC Ensemble	71	2	97
Unregularized	69	0	101
Heteroscedastic Concrete Dropout	47	1	122
MC Dropout	46	1	123

Figure 3.9. *Novelty detection* AUROC and AUPR pairwise comparison counts of wins/draws/losses.

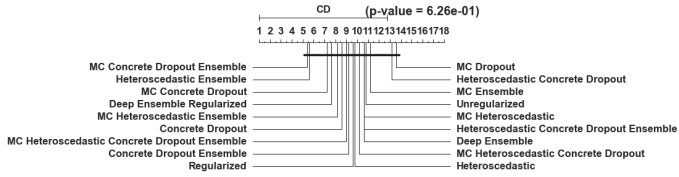


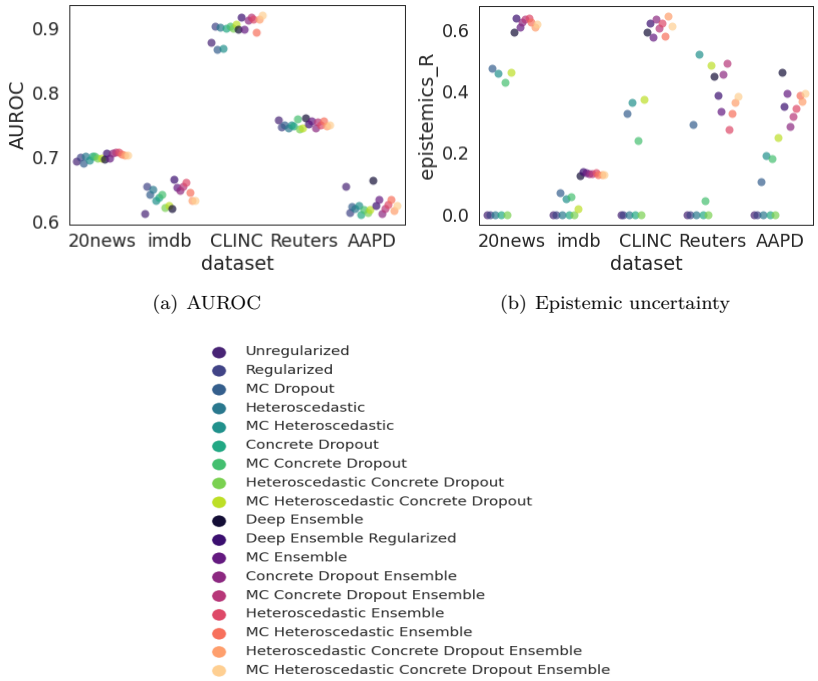
Figure 3.10. *Novelty detection* CD diagram of AUROC.

datasets.

Additionally, we visually detail in Appendix C.1 *Fig.* A.3 density estimates for uncertainty quantities with respect to in-domain versus novel data with most hybrid ensemble methods demonstrating better separable densities.

3.5.4 Experiment: Ablations

In this Subsection, we zoom in on the best performing uncertainty prediction methods relative to the complementary benefits hypothesized for hybrid approaches (Section 3.4.1), provide explanations for results specific to an architecture (TextCNN vs. BERT, Section 3.4.3), and present ablations on critical hyperparameters.



(a) AUROC (b) Epistemic uncertainty

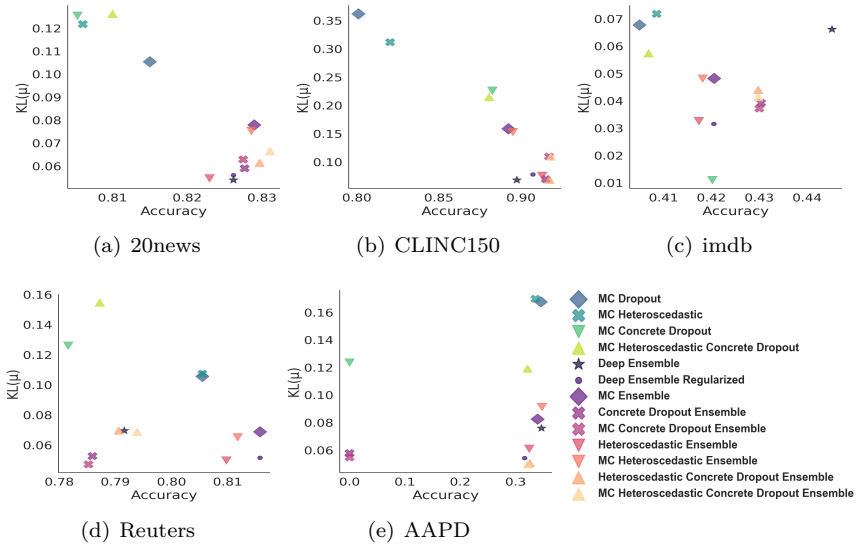
Figure 3.11. Comparison with AUROC(↑) and Epistemic uncertainty PCC(↑) for task and dataset-specific differences in novel class detection. Methods with 0 correlation do not support model uncertainty quantification.

3.5.4.1 Diversity

Diversity of samples drawn from a posterior, either via T MC samples and/or M ensemble components, is an important condition for efficient uncertainty estimation. If each sample presents a similar function, the overall prediction can be overconfident, and increasingly drawing samples will not reduce this. We derive a small experimental setting from [118] to measure function-space diversity for all predictive uncertainty methods involving posterior sampling.

In *Fig.* 3.12 we analyze the relation between accuracy and diversity as measured by Kullback-Leibler divergence between a sampled prediction and the predictive mean, $\frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(p(y^*|x^*,\hat{\theta}_t)||\bar{p}(y^*|x^*,\hat{\theta}))$. For a fair comparison, we calculate diversity at the ensemble level if a predictive uncertainty method consists of multiple models, else at the dropout sample level.

While the diversity-accuracy plane does not provide a one-on-one linear



(a) 20news (b) CLINC150 (c) imdb

(d) Reuters (e) AAPD

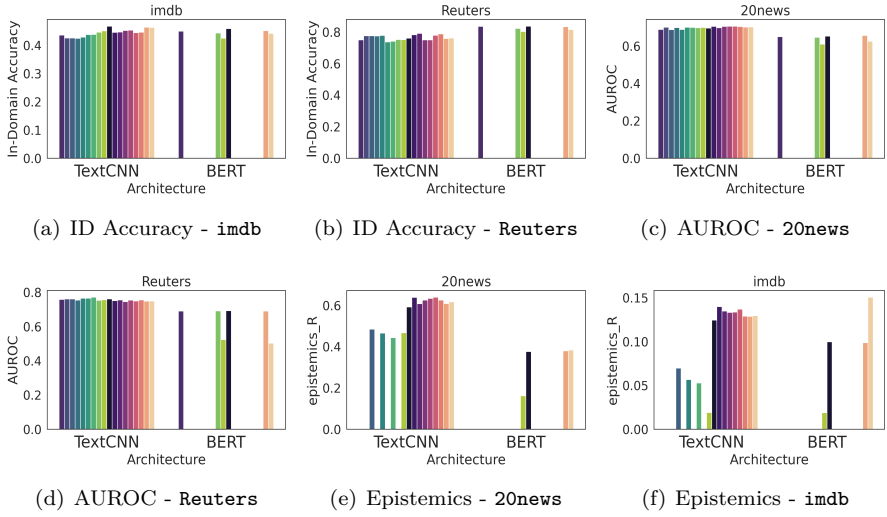
Figure 3.12. Detailed accuracy scores mapped over diversity measured by average KL divergence for each of the benchmark datasets.

relationship, we note in *Fig.* 3.12 (a,b,d) promising results for hybrid ensemble methods, which with higher diversity improve on accuracy over Deep Ensemble. The visual of `imdb` (c) registers overall low diversity, even for simple predictive uncertainty methods which generally achieve higher diversity, albeit by capturing multiple dissimilar yet weaker functions. For `AAPD` (e), most methods are tied for exact accuracy even with different diversities.

3.5.4.2 NLP Architecture

We selected specific representative predictive uncertainty methods on the basis of our previous experiments to run with the Transformer BERT as base architecture. We argue that the chosen architecture can have a non-negligible impact on uncertainty estimation, and we compare with the simple yet controllable TextCNN architecture in order to investigate whether the same conclusions hold for novelty detection.

The separate Out-of-Scope set of `CLINC150` allows us to easily evaluate novelty detection with BERT. We observe in *Fig.* 3.14 on `CLINC150` that BERT does increase novelty detection over all metrics. Even without any hyperparameter



(a) ID Accuracy - `imdb` (b) ID Accuracy - `Reuters` (c) AUROC - `20news`

(d) AUROC - `Reuters` (e) Epistemics - `20news` (f) Epistemics - `imdb`

Figure 3.13. Novelty detection scores mapped per architecture for the benchmark datasets without dedicated OOD split. The legend of Fig. 3.11 applies here.

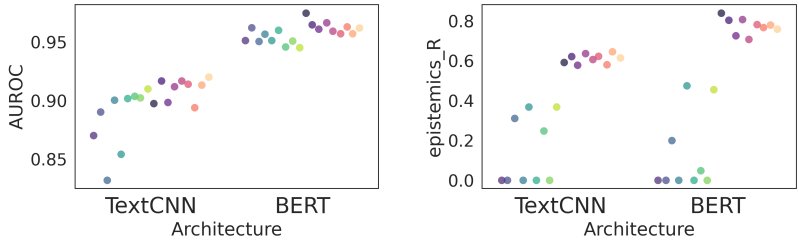


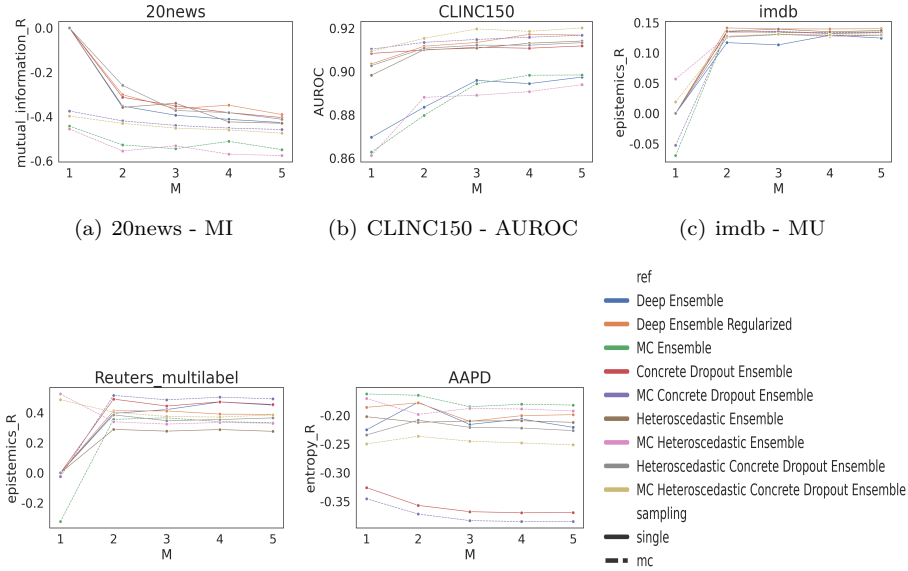
Figure 3.14. Detailed AUROC-epistemics (PCC) scores mapped per architecture on `CLINC150`. Best performance: upper-right corner. The legend of Fig. 3.11 applies here.

tuning *Unregularized* BERT outperforms all TextCNN models. Overall, we register the same ranking of predictive uncertainty methods, albeit a Deep Ensemble with BERT is superior to hybrid ensembles. Crucially, we note that the correlation of epistemic uncertainty with novelty detection is higher for each TextCNN ensemble than for every single BERT model.

Most notably, results on all other datasets are inconsistent with the above. For

comparison, we have trained an informed sub-selection of predictive uncertainty methods with BERT as base architecture (*Fig. 3.13*).

Generally, we observe in (a,b) higher ID accuracy for BERT with relatively slighter gains when ensembling. AUROC scores (c,d) are well below even single TextCNN models, pointing to a crucial deficiency with BERT in a novelty detection setting. The correlation of epistemic uncertainty with novel class samples draws a similar picture (e,f). *MC Heteroscedastic Concrete Dropout Ensemble* on `imdb` does produce more correlated epistemic uncertainty than all other methods.



(a) 20news - MI (b) CLINC150 - AUROC (c) imdb - MU

(d) Reuters - MU (e) AAPD - H

Figure 3.15. Visualization of representative dataset-quantity/metric combinations mapped over stepwise increasing ensemble size M. Note that positive and negative correlations are corollary to the quantity reported. Given the small relative differences, plots are best viewed online.

3.5.4.3 Ensemble size M

Combining models to an ensemble generally benefits performance both in and out-of-domain. Previous research [118, 238] worked out that ensembling benefits stagnate with larger model size M. *Fig.* 3.15 selectively reports novelty detection

metrics or uncertainty correlation scores for all ensemble-based methods of different sizes.

AUROC score for `CLINC150` (3.15b) is a representative example of the expected effect of ensembling. Importantly, it provides crucial evidence for our general hypothesis, demonstrating that ensembling over predictive uncertainty methods gives complementary benefits in novelty detection settings. What is similarly interesting is that the relative benefit of ensembling shows slightly different curves in certain cases. Epistemic uncertainty for `imdb` (3.15c) already attains similar performance at M=2, again showing comparatively slower (since less required) increase at larger M for hybrid ensembles. `AAPD` (3.15e) shows more stagnant behavior for the reliability of entropy with growing ensemble size, irrespective of the predictive uncertainty method.

3.5.4.4 Concrete Dropout p

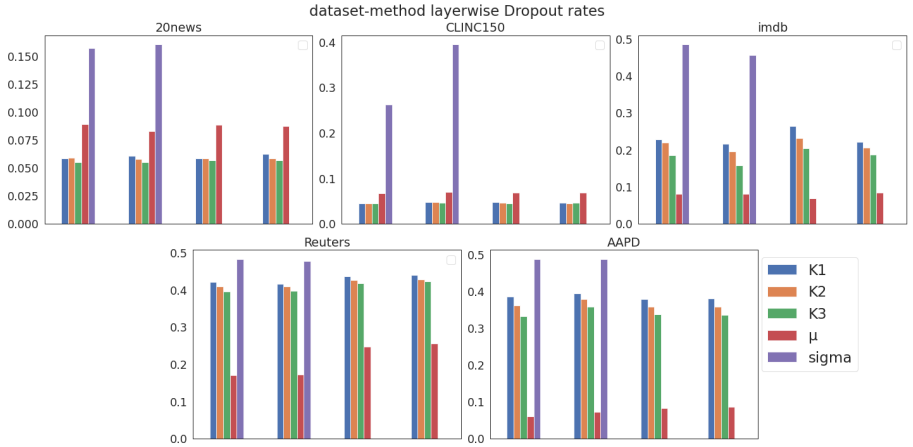


Figure 3.16. Learned layer-wise dropout probability per layer for each method with Concrete Dropout. The first 3 layers are the CNN kernels ($K1-3$), followed by the penultimate layer μ, possibly with σ for modeling heteroscedasticity. The legend of Fig. 3.17 applies here.

Fig. 3.17 relays an important observation on the dataset-wise adaptation of Concrete Dropout: increasing the learned dropout rate as is required for the problem at hand. This reinforces the argument against fixed-rate dropout. [125] remarked that practitioners started to adopt the strategy of fine-tuning dropout with a bottleneck pattern, i.e., start with a higher dropout rate in early layers and decrease the deeper you go in the network. Our results (*Fig.* 3.16)

shows discrepancy with this practice, specifically for `20news` and `CLINC150`. We do note that both converged to low dropout rates, which can provide the basis for this differing behavior.

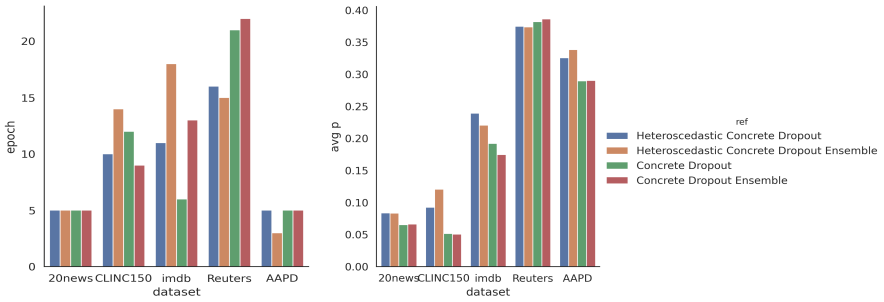


Figure 3.17. Top: Average epoch of convergence per dataset. Bottom: Average learned Concrete Dropout probability per dataset over predictive uncertainty methods. We observe very dataset-dependent dropout rates.

3.6 Discussion

Our study investigates both scalable and hybrid procedures for incorporating uncertainty into DL models for text classification. Next to baseline in-domain uncertainty evaluation, we have designed two experimental settings, novelty detection and cross-domain classification, to analyze the reliability of uncertainty. Additionally, we devised ablation studies to analyze important hyperparameters in connection to our three hypotheses (Section 3.4.1) on complementary benefits for hybrid uncertainty prediction methods.

Benchmarking uncertainty methods We summarize our findings succinctly and discuss the results of each experimental setting.

We find that individually ($>$ indicating "outperforms" over all experiment settings):
Deep Ensemble $>$ *Concrete Dropout* $>$ *(MC) Heteroscedastic* $\geq$ *MC Dropout*

We find that jointly, by considering method combinations:
(MC) Concrete Dropout Ensemble $\geq$ *(MC) Heteroscedastic Ensemble* $>$ *MC Concrete Dropout* $>$ *Deep Ensemble* $>$ *Deep Ensemble Regularized* $>$ *MC Dropout*

In-domain results (Section 3.5.1) corroborate the superiority of *Deep Ensemble* with high accuracy and proper scores (NLL, Brier). *Table* 3.3 demonstrates that the improvements come from accuracy as opposed to calibration, where Concrete Dropout-based methods rule.

Cross-domain experiments (Section 3.5.2) give differing conclusions: cross-domain generalization results are similar to in-domain, whereas out-of-domain detection follows novelty detection results. Our evaluation of uncertainty quantities (*Fig.* A.2) demonstrate reliably higher correlation of uncertainty with domain discrepancy. We do take note of relatively low magnitude AUROC (*Fig.* 3.6), which underlines how challenging out-of-domain detection is in a domain adaptation setting with comparably similar linguistic patterns.

Novelty detection (Section 3.5.3) in text classification gives reverse results: Hybrid ensemble methods with *Concrete Dropout* rank highest scored by AUROC, AUPR and model uncertainty correlation, followed by other method combinations that induce calibration. We do note that specific method performance is often tied to task and dataset characteristics, with results averaged over the 5 benchmark sets showing statistically non-significant differences between methods. As shown in *Table* 3.9, standard Deep Ensemble, i.e., without any regularization or prior from combining methods, perform worse outside the in-domain setting. The case for standard MC Dropout is even worse with novel class robustness (AUROC and AUPR) lower than the *Unregularized* point-estimate model.

Remarkably, BERT performs worse than the simpler TextCNN model at detecting distribution shift in the form of novel class data (*Fig.* 3.14). Results on the OOS set of `CLINC150` differ from results obtained on all other datasets, which we believe can be attributed to the short, in-domain intent commands differing strongly in vocabulary with the OOS samples, resulting in a comparatively less challenging novelty detection setting. We contend that novelty detection is actually more challenging for BERT despite of its pretrained language modeling knowledge and because of the strict requirement to fine-tune the task-specific final layer with new supervision. Its ability to detect (and overly rely on, *e.g.*, [162]) statistically relevant yet possibly spurious cues in language data will make it overconfident with transfer to a new task when the i.i.d. assumption cannot be maintained.

Validating hybrid approaches We have empirically analyzed individual-joint effectiveness in modeling predictive uncertainty and will answer our three hypotheses on complementary benefits from combining inter and intra-modal posterior approximation.

Firstly [**A**], ensembling (increasing M) proves to give relatively higher performance benefits than stochastically sampling predictions from an optimized solution (T). The effect is clearest in the in-domain setting (*Table.* 3.3) and is less pronounced in the out-of-domain settings. For a given predictive uncertainty method, we cannot provide solid evidence that uncertainty reliability always improves when subspace sampling (increasing T, "MC"). AUROC and AUPR rankings (*Figs.* 3.10 and 3.6) present evidence in favour, although *Fig.* 3.11 depicts a more fine-grained comparison over datasets and uncertainty methods. Our analysis of diversity (*Fig.* 3.12) shows promising results for hybrid ensemble methods, which exhibit higher diversity in posterior samples resulting in improved accuracy.

Secondly [**B**], our newly proposed hybrid uncertainty estimation methods improve effectively over singular methods, both in novelty detection (*Table* 3.9 and *Figs.* 3.10, 3.11) and out-of-domain detection (*Fig.* 3.6). Additionally, in ablation studies we find (*Fig.* 3.15) that combining predictive uncertainty methods in an ensemble attains higher performance with a lower number of models ($M < 5$) compared to a Deep Ensemble ($M = 5$).

Thirdly [**C**], *Table* 3.3 demonstrates that MC Concrete Dropout improves over MC Dropout (p=0.5) on ECE and proper scoring functions. The out-of-domain experiments (detail: *Fig.* 3.11) similarly show that not fine-tuning dropout to the dataset and task at hand is detrimental even when combining models into an ensemble (*e.g.*, MC Ensemble vs. MC Concrete Dropout Ensemble). Ablation on Concrete Dropout (*Fig.* 3.17) points to very dataset-dependent learned probability rates, which vary strongly layer-wise (*Fig.* 3.16). We link the empirical superiority of *MC Concrete Dropout Ensemble* to balanced posterior collapse, thanks to the VI-based optimization of the dropout prior. We tentatively claim that the former provides constrained hypothesis support and a more fine-tuned influence of prior.

Benchmark comparison When comparing our results to existing BDL benchmarks, most observations are consistent for in-domain and out-of-domain performance.
Our in-domain results are most similar to [348], where Deep Ensemble outperforms most methods, —albeit in their survey they did not compare combinations of predictive uncertainty—, in our benchmark closely followed by hybrid ensemble methods. When evaluating over various data retention rates [113] observed that "an ensemble of MC Dropout models" (our *MC Ensemble)* consistently outperforms all other methods. This survey offers the closest point of comparison, although our experimental settings vary. While we cannot directly compare cross-domain detection with other benchmarks, we argue that our cross-domain classification setting mimics their low data regime

experiments.

Across different modalities and tasks, Deep Ensemble has been reported to consistently outperform VI-based methods, most specifically MC Dropout, with/without distribution shift (image classification [348], molecule prediction [409], and pendulum physics [56]). However, for a binary image classification problem, [113] report higher accuracy for MC Dropout compared to Deep Ensemble, whereas our results suggest that MC Dropout can induce positive calibration, yet score lower on accuracy and with proper scoring rules. In their experiments they use a fixed dropout rate of 0.2 and fine-tuned weight decay rate, making them fitting for their task at hand and explaining possibly optimistic results. Another uncertainty quantification benchmark [462] reports strong results on image classification for various Monte Carlo methods, although we cannot make a direct comparison. For further discussion, we refer the reader to Appendix 3.7.1.

Our results suggest that BERT performs worse in a novelty detection setting, whereas [174] concludes that Transformers are considerably more robust when compared across domains, *e.g.*, detection of news samples with a sentiment classifier. We point out below that both settings are in fact incomparable. We evaluate detection on novel samples which have alike vocabulary characteristics to the source domain albeit they are excluded from training supervision. Their setting evaluates detection between very disparate domains where linguistic patterns are significantly different and BERT will most probably fallback to its pretrained knowledge for detection. In short, we do believe that pretrained Transformers could perform better under varying distribution shifts, yet with our results underpinning the exception of novel class detection. More research is needed into how the inductive bias from given NN architectures influences approximate inference.

Take-homes For predictive uncertainty in text classification, we derive a number of take-homes from the benchmarking evidence, centered around practical facets to consider for applications.

One has to consider (i) ease and cost of implementation, (ii) computational and memory complexity, comprising training compute, test compute and storage/memory constraints, (iii) the degree of fine-tuning required, (iv) type of supervision; multi-class with low/high number of classes (K) or multi-label with low/high cardinality (C), (v) expectation of distribution shift; in the form of novel class data or unseen language patterns, and (vi) support for uncertainty quantification by source.

For a prototypical low K multi-class text classification task, we advise *Deep*

Ensemble for solid in-domain performance and adequate distribution shift robustness. In the case of memory or storage constraints, for example if your base model already has high complexity, using *(MC) Concrete Dropout* will provide calibration benefits both in and out-of-domain, albeit at a slightly larger implementation cost. Similarly, to constrain computational complexity, it can be more sensible to rely on a TextCNN ensemble (5*6M parameters) rather than BERT (110M parameters). Considering time complexity, we have added detailed compute, time and storage statistics for evaluated methods (Appendix Appendix B.2). We would advise against using *MC Dropout* if the dropout rate and weight regularization are not fine-tuned for the problem at hand. Our benchmarking experiments demonstrate the unpredictable behavior of fixed-rate *MC Dropout*, compared to *Concrete Dropout*, which we used as a proxy for models with fine-tuned dropout ratio. This (mal)practice should be highlighted as it has substantial impact on uncertainty estimation and robustness.

If K starts to increase, it warrants the effort to implement the *Heteroscedastic* loss function, which will make the model more calibrated in-domain. Additionally, it enables data uncertainty estimation for possible noisy ground truths, which can happen more frequently with a larger number of classes.

If C grows larger, reliable epistemic uncertainty estimation becomes more important, since the problem is made more complex given the larger number of label combinations. Our evidence is slightly contradicting, with results obtained on `Reuters` suggesting *MC Concrete Dropout Ensemble* and on `AAPD` warranting *Deep Ensemble*. What should be clear, is that any form of ensembling is valuable in multi-label classification to boost performance.

Under the expectation of distribution shift in the form of novel class data, adding *Concrete Dropout* with stochastic sampling to an ensemble, *MC Concrete Dropout Ensemble*, gives relatively strong benefits compared to a regular Deep Ensemble. Ablations also show that less models (M) would be required to reach similar performance. Generally, in-domain calibration inducing methods are more robust when applied in the tested out-of-domain settings. For the in-domain setting, the incorporation of data uncertainty incrementally improves multi-class text classification. Ablation on NLP architectures (Section 3.5.4.2) points to a deficiency of BERT for detecting novel class data and would similarly be advised against in favour of simpler text classification architectures.

3.7 Additional Uncertainty Approaches

Next to the method combinations benchmarked in the main work, we acknowledge two alternative approaches to uncertainty estimation with appealing

properties such as training scalability and cheaper inference.

3.7.1 Stochastic Gradient MCMC Methods

There exists a wide range of sampling-based inference methods in the stochastic gradient MCMC (SG-MCMC) literature, which have become increasingly more tractable and empirically successful for uncertainty estimation. Specifically, we re-implemented an exemplary approach [530], *cyclical SG-MCMC* (cSG-MCMC), which uses a cosine cyclical learning rate schedule [292] to (i) better explore the highly multimodal loss landscape and (ii) sample more efficiently from the posterior. While this appealing approach reduces computational complexity by only training a single model, we experienced that it is very tricky to finetune with many hyperparameters interplaying. Instead of benchmarking these methods and reporting scores over ranges of hyperparameters, we provide a discussion of the perceived gap in theory and practice for this family of uncertainty methods.

While the stochastic MCMC setting, estimating parameter updates from minibatches, is computationally convenient, it induces several theoretical challenges: i) minibatch noise introduced from small subsets of data [297], ii) omission of the Metropolis-Hastings correction step provides fundamentally biased estimates of posterior expectations [192], and iii) the suggested practice of temperature tempering implies an approximation to the exact posterior instead of proper convergence [122, 491].

Closer to practice, [530]'s methods have been successfully benchmarked [462, 491] with reported performance on OOD detection for image classification datasets comparable to or better than Deep Ensembles. An important caveat is that all hyperparameters have been meticulously finetuned to the task at hand. This is non-trivial given the additional specification of the number of cycles as guided by a training budget, proportion of burn-in steps, and finding an appropriately tempered posterior. The original work [530] mentions little dependence of results on these modifications to the optimization procedure, yet we observed similar to [122] "the complexity and fragility of hyper-parameter tuning, including the learning rate schedule and those that govern the simulation of a second-order Langevin dynamics". Additionally, making combinations of uncertainty methods with cSG-MCMC is non-trivial, since regularization in any form influences the large scale curvature of the regions the optimizer explores.

With regards to re-implementation, we experienced issues with the indexing of sparse gradient updates for the embedding lookup, an operation pervasive in NLP architectures. Our original baseline models were trained with Adam

optimizer, which consistently outscored any of our cSG-MCMC experiments built upon SGD modifications.

There is an unmistakable complexity with how to sample appropriately from the true posterior, as we now rely much on the training data, a "weak" regularizer, on how to add noise for parameter space exploration. Concurrently, the overparametrized regime is becoming commonplace in DL, especially in NLP with the advent of Transformers, which calls for more sensible priors for more than millions of parameters [453] and a better understanding of how output functions are affected [107]. We believe stronger priors are available, not only over parameters $P(\theta)$ but rather over functions $P(f_\theta(x))$ as specified by the choice of architecture [192], which can make this family of methods an even more competitive challenger.

3.7.2 Spectral-normalized Neural Gaussian Process

[283] propose with *Spectral-normalized Neural Gaussian Process* (SNGP) a principled, scalable approach to uncertainty estimation for deep NNs. They promote "distance awareness" as a necessary condition, which they accomplish via spectral weight normalization and a GP output layer. Thanks to the mean-field approximation [295] only a single forward pass suffices without MC sampling to estimate the predictive distribution. Empirically, SNGP was shown to outperform Deep Ensemble by some margin on OOD detection for both image and text data. By demonstrating the relative importance of the decision boundary of a single model $f_\theta(y|x)$ versus averaging over multiple models, we are inspired to analyze the combination of SNGP with alternate uncertainty methods.

We have re-implemented SNGP using components of `edward2` [454], Laplace approximation, random feature GP and spectral normalization. In our experience, the most crucial hyperparameters to finetune were the number of inducing points ($\iota \leq 1024$) and spectral norm multiplier s. For the latter, we follow the recommended tuning procedure to find an appropriate value in the range $\{1, 2, 5, (10, 15)\}$, where we heuristically increased the search space.

For simplicity and computational reasons, we use TextCNN as base architecture. However, in order to correctly apply spectral normalization to convolutional filters [151], we had to re-implement TextCNN(v2) with 2D convolutions and maxpooling. This in turn requires specifying a fixed sequence length in advance, which invalidates directly comparing to the experiment results of Section 3.5. We additionally re-train base models with TextCNN(v2) and combine SNGP with our Regularized baseline (Reg), with MC Dropout (MCD), Concrete Dropout

(CD) and Ensemble (Ens). For SNGP ensembles, we empirically selected $s = 15$ for the base model.

3.7.2.1 SNGP Results

First, we present critical difference analyses for in-domain classification (*Fig.* 3.18) and novelty detection (*Fig.* 3.19). Ensembling SNGP models, *Deep Ensemble SNGP*, proves superior in-domain, followed by *Concrete Dropout Ensemble* with and without SNGP. For novely detection, *(MC) Deep Ensemble* is most successful with small differences between next high-ranked methods.

To our surprise, *SNGP* ranks quite low on the text classification tasks, although in the original work it demonstrated OOD detection superior to *Deep Ensemble*. In what follows, we analyze the novelty detection ranking of SNGP, specifically per dataset and for multiple values of s.

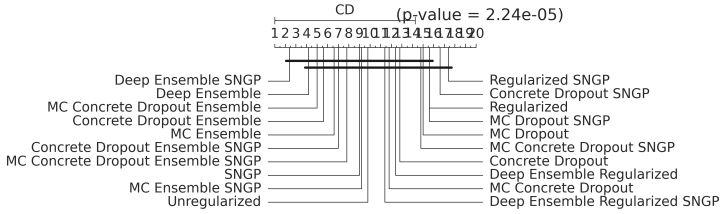


Figure 3.18. CD diagram of NLL for base and SNGP method combinations with a TextCNNv2 backbone.

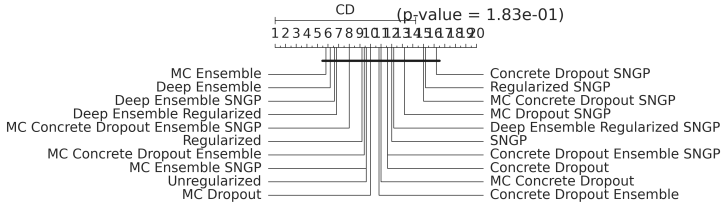


Figure 3.19. CD diagram of AUROC for base and SNGP method combinations with a TextCNNv2 backbone.

In order to zoom in on the relative ranking of SNGP (combination) methods, we plot in *Fig.* 3.20 AUROC detection scores for datasets with interesting trend changes. Overall, SNGP underperforms on `CLINC-OOS`, with the exception of *Deep Ensemble SNGP*. For `20news`, *SNGP* and *Deep Ensemble SNGP* rank

high, although any additional regularization with *SNGP* worsens detection, even as ensemble. For `Reuters`, we observe the exact opposite to `20news`, with *SNGP* reporting high detection scores only when regularization is added, e.g. *Regularized SNGP*. Remarkably, this trend is reversed for the base model, with *Unregularized* scoring particularly good.

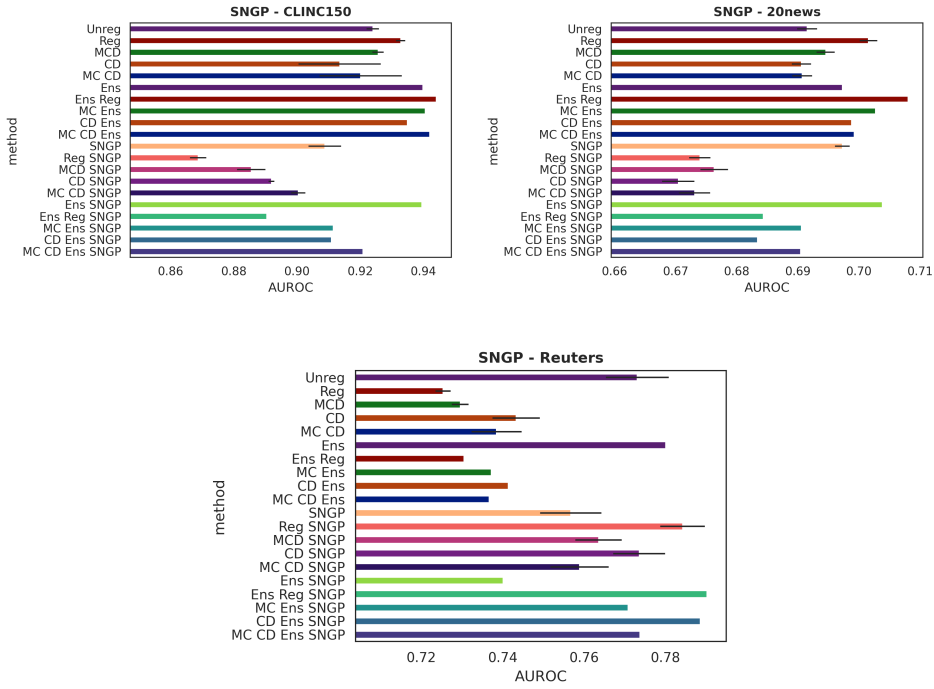


Figure 3.20. AUROC scores over unique (abbreviated) methods per dataset. Error bars are computed over multiple runs (5 seeds) for non-ensembles.

Finally, *Fig.* 3.21 reports on how novelty detection varies for different values of the spectral normalization multiplier s. As the trend lines indicate, larger values of s generally improve novelty detection, although AUROC varies more (larger shading) between methods and datasets. This observation prompts us to investigate the optimality of s per dataset. The right subplot shows that spectral norm multipliers are very dataset-dependent and that searching further than the originally suggested range can give great performance boosts.

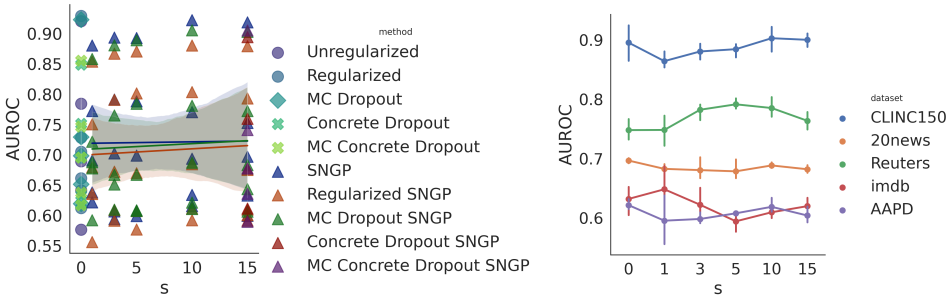


Figure 3.21. Left: AUROC scores (y-axis) over all datasets with unique runs plotted for base ($s = 0$) and *SNGP* TextCNNv2 models with varying spectral normalization multipliers (x-axis). Lines with shading indicate the trend observed between AUROC and s. Right: AUROC mean and stddev over runs, sampling and datasets.

3.7.2.2 SNGP Discussion

While *SNGP* was reported to outperform Deep Ensemble in the original CLINC OOD detection experiments [283], our results do not deliver the same ranking. While investigating the interaction of *SNGP* with different uncertainty methods, we observe the nontrivial role of spectral normalization, specifically setting the norm multiplier s to an appropriate value. Additionally, we contribute the analysis of the interplay with additional regularization mechanisms, which was missing in the literature. The original work mentions that given an approximation with the power iteration method, there is not a precise control of the true spectral norm. Whereas spectral normalization keeps the magnitude of updates to weights in check, Dropout regularization and weight decay may rescale layers' spectral norm in unexpected ways. We hope our experimentation demonstrates the need for deeper understanding of how to combine multiple regularization mechanisms and maintain a good spectral norm approximation for effective posterior approximation.

3.8 Limitations

As with the majority of benchmarking literature in Bayesian Deep Learning, the design of the current study is subject to limitations.

The first limitation concerns selection bias for text classification datasets. We benchmark 6 prototypical text classification datasets covering binary, multi-class, and multi-label classification by topic, sentiment and intent. The task

domain of text classification is very large with additionally interesting variations of (i) short social media or long business document text, (ii) hierarchical or extreme multi-label text classification, and (iii) challenging task settings such as fake news detection or reading comprehension. Since these present open sub-problems in text classification we did not consider them for our benchmarking study, yet encourage analysis for future research.

The second limitation is related to the representativeness of uncertainty quantification methods. We specifically opted for scalable procedures which have been increasingly gaining attention by practitioners. In total we derive 18 method combinations from two competing predictive uncertainty procedures, for which we already resort to statistical summaries and rank-based evaluation to present results. Due to computational constraints, retraining min. 5 ensembles of size $M = 5$ per dataset and per experiment setup, we did not consider a natural Bayesian extension of Deep Ensemble, *Bayesian Ensemble* [360] where all weight initialization is shared around a single prior. Additionally, 3.7 includes preliminary experiments with two new uncertainty approaches, *cyclical SG-MCMC* [530] and *SNGP* [283], which are less practical to benchmark, but bring promising ideas for improved, high-quality uncertainty estimation.

Finally, evaluating the quality of uncertainty quantification is an open problem in BDL, typically approached with proxy setups, as is the case in our benchmark with a focus on novelty detection and cross-domain generalization. Section 3.3.5 presents a nuanced view of this evaluation practice. In addition, evaluating reliable uncertainty estimation in NLP as opposed to other modalities is complicated due to the discrete nature of language. Ideally, we would have extended our benchmark with more probing setups covering situations where we expect predictive uncertainty to be crucial, for instance, when dealing with noisy supervision/inputs or low data regimes.

3.9 Chapter Conclusion

In general, while seeking to optimize for a well-approximated (whether or not Bayesian) posterior, current predictive uncertainty methods are imperfect and very often practically not useful. However, the need for practical and scalable solutions to both incorporating and evaluating the quality of uncertainty is huge, as it is a prerequisite to reliable automation. Uncertainty quantification requires modality to task-specific benchmarking to help practitioners safely rely on them and inform researchers to prioritize the right approaches.

In this work, we have presented empirical evidence from benchmarking uncertainty methods in text classification, contributing and calling attention

to the under-explored study of uncertainty quality and model robustness in realistic NLP data distributions.

Interestingly, we find that general behavior of predictive uncertainty methods does not hold over different datasets, with method performance often tied to the text classification task. Overall, we cannot discern a clear winning predictive uncertainty procedure, yet some methods clearly perform worse. Although a universal methodology is absent, we observe that there are specific correlations between a method's performance and the problem setting representing text classification task characteristics, formulated in practical take-homes.

An important contribution is the proposed novel combinations of predictive uncertainty methods. Our benchmarking experiments have revealed *MC Concrete Dropout Ensemble* to be overall superior at novel class and out-of-domain detection in text classification, even with a lower ensemble size. Most notably, it outperforms Deep Ensemble which has leading performance in recent BDL surveys on image data. We linked complementary benefits of hybrid uncertainty estimation methods to ongoing research on NN diversity in function-space and have provided more evidence in support of hybrid approaches. We have determined in an ablation study that M, ensemble size, T, number of Monte Carlo samples, and p, dropout probability rate, are crucial hyperparameters to take into consideration for improved robustness and uncertainty estimation. Finally, we experimentally validated predictive uncertainty methods on real-world text classification tasks, including multi-label targets, coupling our hypotheses and results to the NLP problem space. Crucially, we found an important deficiency of BERT, compared to a more simple NLP architecture TextCNN, with respect to novel class robustness, limiting the applicability of transfer learning from pretrained Transformers under the expectation of uncertainty and novel class instances.

To further improve calibration and robustness in the text classification domain, and by extension uncertainty in NLP, we need to better understand what will make existing or novel uncertainty estimation techniques successful. This requires the development of well-motivated tooling and protocols to reliably assess the quality and fidelity of posterior approximation. Generally, the role of priors in increasingly larger models deserves more attention. While our work focused on posterior geometry and weight-based priors in the form of regularization, stronger, more meaningful functional priors exist, which should be exploited to encourage desirable predictive behavior such as robustness to specific distribution shifts. Particularly for NLP, more focused research is required into what aspects —language data characteristics, inherent task difficulty or ambiguity, architecture design, learned representations, objectives, and effective parameter usage— render NLP pipelines more complex to imbue with reliable uncertainty and guarantee future out-of-distribution robustness.

Part II

Realistic and Efficient Document Understanding

Chapter 4

Beyond Document Page Classification: Design, Datasets, and Challenges

The contents of this chapter comes from a publication [470] that was presented as an **oral** presentation at WACV 2024 ($\frac{53}{2042} \approx 2.5\%$):

Jordy Van Landeghem, Sanket Biswas, Matthew Blaschko, and Marie-Francine Moens. Beyond Document Page Classification: Design, Datasets, and Challenges. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2962–2972, 2024

Disclosing the work done:

I conceptualized the work, implemented the experiments, and wrote the manuscript. Sanket Biswas helped with related work and polishing the writing, and we acknowledge help in data collection from Ruben Perez Tito and Stefan Larson.

This chapter focuses on moving beyond the (self-imposed) restrictions of page limits, and exploring the full potential of DL for document processing. A major highlight is the need to bring document classification benchmarking closer to real-world applications, both in the nature of data tested (X: multi-channel, multipaged, multi-industry; Y: class distributions and label set variety) and

in classification tasks considered (f: multipage document, page stream, and document bundle classification, ...). We start by introducing the problem of document classification (DC) and its importance in the larger scope of document understanding, for which we emphasize visually-rich documents, adopting the acronym VDU instead. Moreover, we identify the lack of public multipage document classification datasets, formalize different classification tasks arising in application scenarios, and motivate the value of targeting efficient multipage document representations.

An experimental study on proposed multipage document classification datasets demonstrates that current benchmarks have become irrelevant and need to be updated to evaluate *complete* documents, as they naturally occur in practice. This reality check also calls for more mature evaluation methodologies, covering calibration evaluation, inference complexity (time-memory), and a range of realistic distribution shifts (*e.g.*, born-digital vs. scanning noise, shifting page order). This chapter ends on a hopeful note by recommending concrete avenues for future improvements, pertaining to document dataset construction efforts and suggested methodologies.

The work in this chapter was the trigger for the next chapter (Chapter 5), in which we propose a new, comprehensive DU benchmark, DUDE, that is more aligned with real-world applications and practices, naturally including multipage documents that satisfy many of this chapter's recommendations.

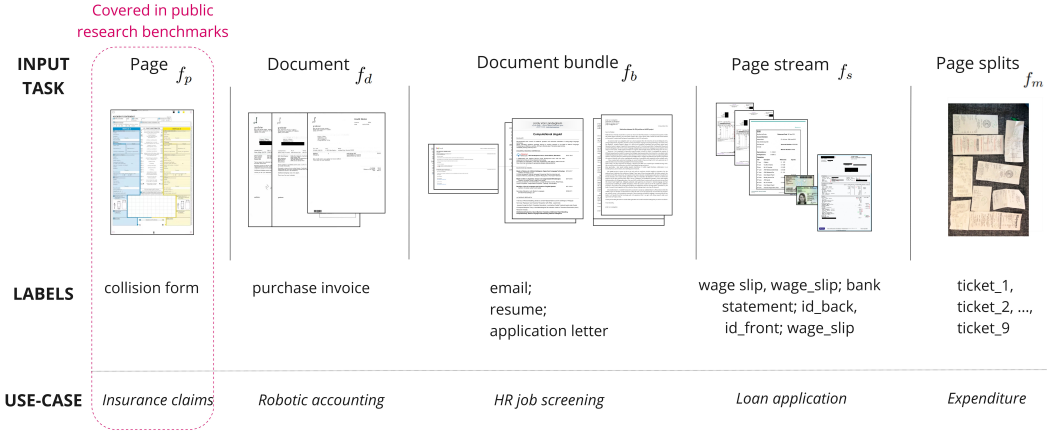


Figure 4.1. Overview of different classification tasks that can be found in real-world VDU applications, that are not sufficiently addressed in DC research. The classification task notation and definitions are introduced in Section 4.2.

4.1 Introduction

Visual Document Understanding (VDU) comprises a large set of skills, including the ability to holistically process both textual and visual components structured according to rich semantic layouts. The majority of efforts are directed toward the application-directed tasks of classification and extraction of key information (KIE) in visually-rich documents (VRDs). **Document classification** (DC) is a fundamental step in any industrial VDU pipeline as it assigns a semantically meaningful category, routes a document for further processing (towards KIE, fraud checking), or flags incomplete (*e.g.*, missing scans) or irrelevant documents (*e.g.*, recipe cookbook in a loan application).

Documents are intrinsically multipaged, explaining (partly) why PDF is one of the most popular universal document file formats.[1] While DC in information management workflows typically involves multipage VRDs, current public datasets [165, 233] only support single-page images and constitute too simplified benchmarks for evaluating fundamental progress in DC.

With the advent of deep learning, the VDU field has shifted from region-based analysis to whole-page image analysis. This shift led to substantial improvements in processing document images with more complex layout variability, exposing the limitations of template-based methods. Our work highlights the opportunity and necessity of moving *beyond the page* limits toward evaluation on *complete* document inputs, as they prevalently occur (multipage documents, bundles, page streams, and splits) across various practical scenarios within real-world DC applications, demonstrated in Figure 4.1.

The practical task of long document classification [372] is largely underexplored due to challenges in computation and how to efficiently represent large multimodal inputs. Additionally, the proximity to applications involves a larger community for conducting research, yet innovations may happen in isolation or are kept back as intellectual property, lacking evaluation on public benchmarks [147, 148], consequently hindering reproducibility and fair comparisons.

Existing DC methodology is limited to single-page images, and independently and identically distributed (i.i.d.) settings. We propose an improved methodology that extends its scope to multipage images and non-i.i.d. settings. We also reflect on evaluation practices and put forward more mature evaluation protocols. To better capture the complexity of real-world document handling, we align DC benchmarking closer to practical applications and task formulations.

[1] PDF is the 2nd most popular file format on the web (after HTML and XHTML) following detected MIME types in CommonCrawl.

Our key contributions can be summarized as follows:

- We have redesigned and formalized multipage DC scenarios to align fragmented definitions and practices.
- We construct and share two novel datasets RVL-CDIP_MP[2] and RVL-CDIP-N_MP[3] to the community for evaluating multipage DC.
- We conduct a comprehensive analysis of the novel datasets with different experimental strategies, observing the promise from best-case analysis (+6% absolute accuracy) by targeting multipage document representations and inference.
- We overview challenges stalling DC progress, giving concrete guidelines to improve and increase dataset construction efforts.

4.2 Problem Formulation

We propose to use formal definitions to better align DC research with real-world document distributions and practices. This will help to standardize DC practices and make it easier to compare different methods.
Let $\mathcal{X}$ denote a space of documents, and let $\mathcal{Y}$ denote the output space as a finite set of discrete labels. Document page classification is a prototypical instance of classification [472], where the goal is to learn an estimator $f : \mathcal{X} \rightarrow \mathcal{Y}$ using N supervised input-output pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from an unknown joint distribution $P(X, Y)$.

A **page** p is a natural classification input that consists of an image $\boldsymbol{v} \in \mathbb{R}^{Q \times H \times W}$ (number of channels, height, and width, respectively) with T word tokens $\{t_i\}_{i=1}^{T}$ organized according to a layout structure $\left\{\left(x_i^1, y_i^1, x_i^2, y_i^2\right)\right\}_{i=1}^{T}$, typically referred to as bounding boxes, either coming from Optical Character Recognition (OCR) or natively encoded.

Note that in practical business settings, VRDs are presented at inference time to a production VDU system in different forms:

I. Single page (often scanned or photographed)
II. Single document
III. Multiple documents
IV. Multiple pages (often bulk-scanned to a single PDF)
V. Single image with multiple localized pages

[2] huggingface.co/datasets/bdpc/rvl_cdip_mp
[3] huggingface.co/datasets/bdpc/rvl_cdip_n_mp

Classification tasks In a unification attempt, we formalize the different classification inputs and tasks that arise in practical scenarios, as visualized in Figure 4.1.

Definition 9 [*Page Classification*]. (I) A page (as defined above) is categorized with a single category. When only considering the visual modality, the literature refers to it as 'document image classification' [165]. An estimator for page classification with the input dimensionality ($\mathcal{X}_p$) relative to a page (viz., number of channels, height, and width) is defined as:

$$f_p : \mathcal{X}_p \rightarrow \mathcal{Y}, \\ \text{where } \mathcal{Y} = [C] \text{ for } C \text{ mutually exclusive categories.} \tag{4.1}$$

Definition 10 [*Document Classification*]. (II) A **document** d contains a fixed number of $L \in [1, \infty)$ pages, which do not necessarily have the same dimensions (height and width). Albeit a design choice, the input dimensionality is normalized across pages (*e.g.*, $3 \times 224 \times 224$). Assuming a fixed input dimensionality ($\mathcal{X}_d$) relative to a document ($L \times Q \times H \times W$), a document classifier is defined as:

$$f_d : \mathcal{X}_d \rightarrow \mathcal{Y}, \\ \text{where } \mathcal{Y} = [K] \text{ for } K \text{ mutually exclusive categories.} \tag{4.2}$$

Note also the difference in label space between the two previous classification tasks, which can have some overlap for document types that are uniquely identifiable from a single page (*e.g.*, an accident statement form).

Definition 11 [*Document Bundle Classification*]. (III) A bundle b can contain a variable number of B documents, each with a potentially different amount of L pages. A bundle classifier models a sequence classification problem over multiple documents:

$$f_b : \mathcal{X}_b \rightarrow \mathcal{Y}, \text{ where } \mathcal{Y} \text{ is a product space of } B \text{ documents,} \\ \mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_B, \text{ with } \{\mathcal{Y}_j = [K] : j \in [B]\}. \tag{4.3}$$

Definition 12 [*Document Stream Classification*]. (IV) A page stream s is similar to a document in terms of input (number of pages L), albeit typically more varied in content and page formats. Page streams can implicitly contain many different documents, with pages not necessarily contiguous or even in the

right order, as illustrated in Figure 4.1.

$f_s : \mathcal{X}_d \to \mathcal{Y}$, where $\mathcal{Y}$ is a product space of L pages,

$$\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_L, \text{ with } \{\mathcal{Y}_j = [C] : j \in [L]\}. \quad (4.4)$$

A very concrete example of how the label sets $[C]$ and $[K]$ can differ is in a loan application use-case where national registry proofs need to be sent: If two pages are sent with the front and back of the ID-card, f_s requires two labels (*id_front, id_back*), whereas f_d requires a single document label (*id_card*).

A critical note is due to differentiate page stream segmentation (PSS) [128, 328, 494] and page stream classification as defined above (f_s). PSS treats a page stream as a binary classification task to identify document boundaries, without classifying the identified documents afterward. f_s considers the task in one stage where C is constructed in a way to send atomic units such as a wage slip in Figure 4.1 for individual downstream processing or it can be combined to a single document label from $[K]$ based on assigned page labels. Two-stage processing is possible by applying PSS as an instance of a f_s classifier with $[C] = \{0, 1\}$ where 1 indicates a document boundary, followed by f_d.

Definition 13 [*Page Splitting*]. (V) A multipage image m contains multiple page objects of similar types which can have multiple orientations, page dimensions, and often physical overlap from poor scanning [132]. A standard example involves multiple receipts to be analyzed for reclaiming VAT. While a complete approach will consist of localizing pages (using edge/corner detection, object detection, or instance segmentation) and identifying page types, we will only focus on the latter. For instance, multipage splitting can be defined as a preliminary check on how many page types are present in a multipage image (with input dimensionality similar to a single page p):

$$f_m : \mathcal{X}_p \to \mathcal{Y}, \text{ where } \mathcal{Y} = \mathbb{Z}^C. \quad (4.5)$$

Payment proofs such as tickets and receipts more often are packed together due to their compactly printed sizes, which would require splitting the unique documents from within a page to send individually for further processing. Following the national registry example. another rare yet “economical" variation for f_d occurs when a single page contains both the front and back of the ID card stitched together. These edge cases (rightmost example in Figure 4.1) should be dealt with on a case-by-case basis for how to set up $[K]$ (*e.g.*, specific label: multi-tickets).

The formalisms defined above establishes a taxonomy of DC tasks, which will be retaken in the discussion of challenges to align DC research and applications (Section 4.5).

4.3 Balancing Research & Applications

Having established a taxonomy, we further sketch the role of DC in the larger scope of VDU, both in the applications and research context. We point to related VDU benchmarks and describe current DC datasets with their relevant (or missing) properties using the task formalizations. Next, we link to related initiatives in dataset construction and calls for reflection on DU practices. Finally, we introduce the curated DC datasets to support multipage DC (f_d) benchmarking, which will be used in a further experimental study.

General Benchmarking in VDU: In any *industrial application context* where information transfer and inbound communication services are an important part of the day-to-day processes, a vast number of documents have to be processed. To provide customers with the expected service levels (in terms of speed, convenience, and correctness) a lot of time and resources are spent on categorizing these documents and extracting crucial information. Complex business use cases (such as consumer lending, insurance claims, real estate purchases, and expenditure) involve processing bundles of different documents that clients send via any communication channel. For example, obtaining a loan typically entails sending the following documents to prove solvency: a number of monthly pay stubs, bank statements, tax forms, and national registry proofs. Furthermore, not all documents are born-digital (BD), and as an artifact of the communication channel (bulk scans/photographs, digitization of physical mail), a single client communication can contain an arbitrary amount of document page images in an unknown order, requiring an f_s classifier. Figure 4.1 provides an overview of the different DC tasks that arise in application scenarios, which are scarcely covered by DC research benchmarks (see Table 4.2). As RVL-CDIP is the only large-scale non-synthetic DC benchmark, we discuss it in more detail, other dataset descriptions can be found in Supplementary.

Current state-of-the-art DU research based approaches [15, 187, 259] leverage the "pretrain and fine-tune" procedure that performs significantly well on popular DU benchmarks [165, 188, 197, 544] (see Table 4.1). However, their performance drops significantly when exposed to real-world business use cases mainly due to the following reasons: (1) The models are limited to modeling page-level context due to heavy compute requirements (*e.g.*, quadratic complexity of

Dataset	Size	Data Source	Domain	Task	OCR	Layout
IIT-CDIP [252]	35.5M	UCSF-IDL	Industry	Pretrain	✗	✗
RVL-CDIP [165]	400K	UCSF-IDL	Industry	DC	✗	✗
RVL-CDIP-N [241]	1K	Document Cloud	Industry	DC	✗	✗
TAB [328]	44.8K	UCSF-IDL	Industry	DC	✗	✗
FUNSD [197]	199	UCSF-IDL	Industry	KIE	✓	✗
SP-DocVQA [308]	12K	UCSF-IDL	Industry	QA	✓	✗
OCR-IDL [40]	26M	UCSF-IDL	Industry	Pretrain	✓	✗
FinTabNet [543]	89.7K	Annual Reports S&P	Finance	TSR	✗	✓
Kleister-NDA [432]	3.2K	EDGAR	US NDAs	KIE	✓	✗
Kleister-Charity [432]	61.6K	UK Charity Commission	Legal	KIE	✓	✗
DeepForm [435]	20K	FCC Inspection	Forms broadcast	KIE	✓	✗
TAT-QA [550]	2.8K	Open WorldBank	Finance	QA	✓	✗
PubLayNet [544]	360K	PubMed Central	Scientific	DLA	✗	✓
DocBank [261]	500K	arxiv	Scientific	DLA	✓	✓
PubTabNet [545]	568K	PubMed Central	Scientific	TSR	✗	✓
DUDE [468]	40K	Mixed	Multi-domain	QA	✓	✗
Docile [422]	106K	EDGAR & synthetic	Industry	KIE	✓	✗
CC-PDF [460]	1.1M	Common-Crawl (2010-22)	Multi-domain	Pretrain	✗	✗

Table 4.1. **DU Benchmarks** with their significant data sources and properties. Acronyms for tasks DC: Document Classification DLA: Document Layout Analysis KIE: Key Information Extraction QA: Question Answering TSR: Table Structure Recognition

Dataset	**Purpose**	**#d**	**#p**	$\lvert\mathcal{Y}\rvert$	**Language**	**Color depth**
NIST [98]	f_s		5590	20	English	Grayscale
MARG [290]	f_s		1553	2	English	RGB
Tobacco-800 [553]	f_s		800	2	English	Grayscale
TAB [328]	f_s		44.8K	2	English	Grayscale
Tobacco-3482 [232]	f_p		3482	10	English	Grayscale
RVL-CDIP [165]	pretraining, f_p		400K	16	English	Grayscale
RVL-CDIP-N [241]	f_p, OOD		1002	16	English	RGB
RVL-CDIP-O [241]	f_p, OOD		3415	1	English/Mixed	RGB
`RVL-CDIP_MP`	f_d	±400K	$\mathbb{E}[L] = 5$	16	English	Grayscale
`RVL-CDIP-N_MP`	f_d, OOD	1002	$\mathbb{E}[L] = 10$	16	English	RGB

Table 4.2. **Statistical Comparison** of public and proposed extended multipage `DC` datasets. OOD refers to out-of-distribution detection. $\#d$ and $\#p$ refer to number of documents or pages, respectively. For the novel MP datasets, we report the average number of pages.

self-attention [473]), effectively treating each document page as conditionally independent and potentially missing out on essential classification cues. (2) The methods are heavily reliant on the quality of OCR engines to extract spatial local information (i.e. mostly at word level) suitable to solve downstream benchmark tasks; but fail to *generalize* well on business documents. (3) Existing datasets used for pretraining [165, 252] are different in terms of domain, content, and visual appearance from many downstream `DC` tasks (detailed in Section 4.5.3).

Therefore, it can be challenging for industry practitioners to choose a specific model to fine-tune for the DC use cases and task specifics that they commonly encounter.

RVL-CDIP The Ryerson Vision Lab Complex Document Information Processing [165] dataset used the original IIT-CDIP (The Illinois Institute of Technology dataset for Complex Document Information Processing) [252] metadata to create a new dataset for document classification. It was created as the equivalent of ImageNet in the VDU field, which invited a lot of multi-community (Computer Vision, NLP) efforts to solve this dataset. It consists of low-resolution, scanned documents belonging to one of 16 classes such as *letter, form, email, invoice.*

Proposed Datasets RVL-CDIP_MP is our first contribution to retrieve the original documents of the IIT-CDIP test collection which were used to create RVL-CDIP. Some PDFs or encoded images were corrupt, which explains that we have around 500 fewer instances. By leveraging metadata from OCR-IDL [40], we matched the original identifiers from IIT-CDIP and retrieved them from IDL using a conversion. However, the same caveats for RVL-CDIP apply.

RVL-CDIP_MP-N can serve its original goal as a covariate shift test set, now for multipage document classification. We were able to retrieve the original full documents from DocumentCloud and Web Search. As no existing large-scale datasets include granular page-level labeling (in terms of $[C]$) for multipage documents, we could not create a benchmark for evaluating f_s. Appendix B points to visualizations from the proposed datasets.

Related Initiatives General benchmarking challenges have driven the VDU research community to set the seed for initiatives to create its own document-oriented "ImageNet" [399] challenge over which multiple long-term grand challenges can be defined (deepdoc2022, scaldoc2023). In another task paradigm, DocuVQA, there have been efforts in the same spirit to redirect focus to multipage documents [451, 467]. For the task of KIE, [424] launched a similar call for practical document benchmarks closer to real-world applications. While these initiatives demonstrate a similar-looking future direction, our contribution goes beyond introducing novel datasets and seeks to guide the complete methodology of DC benchmarking.

4.4 Experimental Study

To classify a multipage document, one might ask the question "*Why not just predict based on the first page? What would be the gain of processing all pages? What baseline inference strategies can be applied to classify a multipage document?*". This prompted us to put these assumptions to the test in a small motivating study[4].

As current public datasets only support page classification, we have extended some existing DC datasets to already enable testing a slightly more realistic, yet more complex document classification scenario (f_d).

We have reconstructed the original PDF data of the DC datasets in Section 4.3. The goal of this experiment is to tease some issues and strategies when naively scaling beyond page-level DC. Our baseline of choice is the document foundation model DiT-Base [259], which as a visual-only f_p is competitive with more compute-intensive multimodal, OCR-based pipelines [15, 187, 443].

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

Table 4.3. **Tested inference methods** to classify multipaged documents and simulate a true document classifier f_d. Scope refers to the independence assumption taken at inference time.

Table 4.3 overviews some straightforward inference strategies. Consider the simplest inference strategy is to *sample* a given page with index $l \in [L]$ (or in our case $\{1, 2, L-1\}$) from $\hat{y}^l = [f_p(x)]^l$. The *sequence* strategies mainly differ in how the final prediction $\hat{y}$ is obtained from predictions per page, assuming a probabilistic classifier $\tilde{f}_p : \mathcal{X}_p \to [0,1]^K$.

$$\text{MaxConf}(x, y) = \underset{\substack{l \in [L] \\ k \in [K]}}{\text{argmax}} [\tilde{f}_p(x, y)]_k^l \tag{4.6}$$

[4]Code provided at: https://huggingface.co/bdpc/src

$$\text{SoftConf}(x, y) = \underset{k \in [K]}{\text{argmax}} \sum_{l=1}^{L} [\tilde{f}_p(x, y)]^l \tag{4.7}$$

$$\text{HardVote}(x, y) = \underset{k \in [K]}{\text{argmax}} \sum_{l=1}^{L} e_{\hat{y}^l}, \tag{4.8}$$

with e a one-hot vector of size K. The *grid* strategy is intuitive as we tile all page images in an equal-sized grid that trades off the resolution to jointly consume all document pages. While results in this experiment with fairly low grid resolution (224 x 224) are poor, variations (with aspect-preserving [247] or layout density-based scaling) deserve to be further explored.

Strategy	Acc↑	F1↑	$F1_M$ ↑	ECE↓	AURC↓
f_p\$ [259]	93.345	93.351	93.335	0.075	0.010
first	91.291	91.286	91.271	0.073	0.014
second	87.295	87.305	87.277	**0.070**	0.029
last	85.091	85.060	85.028	0.072	0.038
MaxConf	**91.407**	**91.453**	**91.344**	0.124	0.006
SoftVote	91.220	91.185	91.236	0.134	**0.004**
HardVote	85.995	86.182	85.781	0.085	0.018
grid	72.642	72.045	73.266	0.109	0.042

Table 4.4. Base classification accuracy of DiT-base [259] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_MP` per baseline f_d strategy. Best results per metric are boldfaced. \$ refers to our reproduction of results.

Strategy	Acc↑	F1↑	$F1_M$ ↑	ECE↓	AURC↓
f_p [241]	78.643	81.947	60.564	0.105	0.076
first	**78.760**	**75.316**	**60.801**	0.144	**0.025**
second	64.939	58.741	50.773	0.132	0.071
last	64.228	58.192	48.859	0.128	0.074
MaxConf	76.321	72.855	57.470	0.180	0.042
SoftVote	73.984	69.163	56.486	0.183	0.039
HardVote	67.480	63.188	52.235	0.110	0.088
grid	47.755	40.645	38.584	**0.102**	0.170

Table 4.5. Base classification accuracy of DiT-base [259] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_N_MP` per baseline f_d strategy. Best results per metric are boldfaced.

Following similar calls in the `VDU` literature [468] to establish calibration and confidence ranking as default evaluation metrics, we include Expected Calibra-

tion Error (ECE) [156, 332, 340] to evaluate top-1 prediction miscalibration and Area-Under-Risk-Coverage-Curve (AURC) [138, 193] to measure selective (proportion of test set%) accuracy (cf. Section 2.2.3).

Results in Tables 4.4 and 4.5 demonstrate that classifying by only the first page is a solid strategy, with performance dropping when considering only later pages. Maximum confidence and soft voting require L (pages) times more processing, yet attain similar performance as the best single-page prediction. However, this could be attributed to two factors: i) dataset creation bias since [165] constructed RVL-CDIP from a page of each original .tiff file, for which the label was kept if it belonged to one of the 16 categories, whereas RVL-CDIP-N [241] consistently chose the first-page; ii) documents are fashioned in a summary-detail or top-down content structure over pages. To confirm the validity of the latter hypothesis, more robust experiments on more fine-grained labeled `DC` are needed.

The results from Table 4.4 and Table 4.5 can be interpreted as an upper bound (i.i.d.) and a loose lower bound (non-i.i.d., yet related), respectively. For the former, MaxConf is the most accurate, yet compared to SoftVote has worse AURC, potentially making SoftVote a better candidate for industry use where controlled risk is more valued. While this trend is not reproduced in `RVL-CDIP_N_MP`, it can be explained by the more consistent first-page labeling, adding distracting classification cues from later pages.

Dataset	**Strategy**	Acc↑	Δ
`RVL-CDIP_MP`	first+second$^{(*)}$	93.795	2.504
	first+last$^{(*)}$	93.675	2.384
	second+last$^{(*)}$	89.709	−1.583
	first+second/last$^{(*)}$	**94.454**	3.163
`RVL-CDIP_N_MP`	first+second$^{(*)}$	83.638	4.878
	first+last$^{(*)}$	83.130	4.370
	second+last$^{(*)}$	71.545	−7.215
	first+second/last$^{(*)}$	**84.553**	5.793

Table 4.6. Best-case classification accuracy indicated with $^{(*)}$ when combining 'knowledge' over different pages. Δ refers to the absolute difference with the first page only.

To answer what can be gained from processing a multipage document in a single shot, Table 4.6 reports a best-case error analysis, where a page prediction is counted as correct if the model would have had access to the other pages. This is calculated by using a bit-wise OR operation between the one-hot vectors

($\mathbb{I}[y == \hat{y}]$) expressing correctness for each strategy model. As a proof of concept, this shows that targeting multipage document representations and inference is a promising avenue to improve DC.

4.5 Challenges and Guidelines

Following the introduced task formalizations of Section 4.2, we claim that the distribution on which document classification is currently evaluated publicly and the real-world distributions have heavily diverged. Additionally, our experimental validation on the novel datasets demonstrated the potential of multipage DC, empirically reinforcing our call to action on improving DC methodologies. Let $P^A(X, Y)$ and $P^R(X, Y)$ denote those two distinct distributions, *real-world applications* and *research* respectively. Further, we will characterize the specific divergences with concrete examples and suggestions for better alignment.

4.5.1 Divergence of Tasks: f

The challenge of directly processing multipaged documents is typically avoided by current DC models which only support single-page images [15, 153, 187, 216, 247, 263, 371, 443]. Whenever a new DU model innovation happens, the impact for document classification is publicly only measured on the first task scenario (*e.g.*, f_p on RVL-CDIP), whereas production DU systems more often need to deal with the other settings (II,III,IV,V) in Figure 4.1. Moving beyond the limited page image context will test models' ability to sieve through potentially redundant and noisy signals, as the classification can be dependent on very local cues such as a single title on the first page or the presence of signatures on the last page. Without any datasets to test this ability, we also cannot blindly assume that we can simply scale f_p classifiers to take in more context or that aggregating isolated predictions over single pages is a future-proof (performant and efficient) strategy, as our experiments have shown.

While p is a natural processing unit for humans, acquiring supervised annotations for every single page can be more expensive than attaching a single content-based label (from $[K]$) to a multipage document. However, fine-grained labeling with f_s could allow for more targeted and constrained KIE, as knowing a certain page l has label $y^l = \text{id\_front} \in [C]$ will allow you to focus on specific entities such as *national registry number, date/place of birth*. Ultimately, these classification task formulations can also help one consider how to set up f directly and annotate document inputs, depending on the DC use-case.

4.5.2 Divergence of Label Space: Y

Current benchmarks often use simplified label sets that are difficult to reconcile with industry requirements. While RVL-CDIP is the de facto standard for measuring performance on f_p DC, recent research [242] has revealed several undesirable characteristics. It supports only 16 labels that pertain to a limited yet generic subset of business documents, which is far from the 1K classes in ImageNet on whose image it was modeled. Real-world DC use cases typically support a richer number of classes ($K \sim$ 50-400). RVL-CDIP suffers from substantial label noise, estimated to be higher than current state-the-art f_p error rates (see [242] for a detailed analysis) which are overfit to noise. Due to the absence of original labeling guidelines, the labels in RVL-CDIP can be ambiguous, containing disparate subtypes (*e.g.*, business cards in the *resume* category), and inconsistencies between classes (cheques present in both *budget* and *invoice*). Other errors include (near-)duplicates causing substantial overlap between train and test distributions, corrupt documents, and plain mislabeling. However, many common CV benchmarks are plagued by similar issues [31] and would benefit from relabeling campaigns [519] to maintain their relevance.

Considering the above, multi-label classification (not covered explicitly in Section 4.5.1) could be a solution to resolve label ambiguities, yet this requires absolute consistency in label assignments, which when lacking introduces even more label noise. The highest labeling quality could arise from consistent labeling at the page level and hierarchically aggregating page labels ($C \rightarrow K$), yet granular annotations are more expensive to obtain. Alternatively, it may be better to follow the mutually exclusive and collectively exhaustive (MECE) principle [72] to construct label sets at the document level.

Finally, an overlooked aspect of current benchmarks is that label sets $[K]$ can be constructed based on some business logic, where a very local cue can lead to a class assignment such as some checked box on page 26. Admittedly, this does conflate the tasks of document object detection, KIE, and DC within a single label set. However, the current focus on classes with plenty of evidence across a document, with more global classification cues, should be balanced with document types that rely on local cues.

Taking the above issues into account, the community should work together towards developing more effective and realistic DC datasets that better align with the needs of industry practitioners. While tackling the challenge of Y divergence was out-of-scope for the contributed datasets, the next Subsection gives systematic recommendations for obtaining better future DC benchmarks.

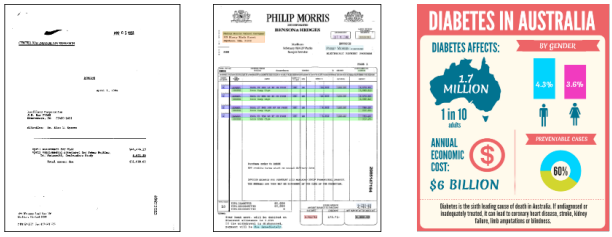


Figure 4.2. **Divergence of input data.** The first image is an example from DC benchmark RVL-CDIP [165], the second one from Docile [422] for KIE, while the third one comes from Info-VQA [310], illustrating the visual-layout richness of modern VRDs vs. the monotonicity of most DC document data.

4.5.3 Divergence of Input Data: X

We offer suggestions for future benchmark construction efforts such that they take into account what properties are currently unaccounted for, organically improving on our first pursuit towards multipage DC benchmarking.

We argue that current VDU benchmarks fail to account for many real-world document data complexities: multiple pages, the distinction between born-native, (mobile) scanned documents, accounting for differences in quality, orientation, and resolution. Additionally, the UCSF Industry Document Library (and in consequence all DC datasets drawn from this source) contains mostly old (estimated period 1950s to 2002), type-written black and white documents, while in reality, modern documents can have multiple channels, colors, and (embedded) fonts varying in size, typeface, typography. Recently, there have been efforts to collect more modern VRD benchmarks for tasks such as DocVQA [310, 468], KIE [422], DLA [362]. Modern VRDs contain visual artifacts such as logos, checkboxes, barcodes, and QR codes; geometric elements such as rectangles, arrows, charts, diagrams, ..., all of which are not frequently encountered with the same variety in current benchmarks. Future DC benchmarks should incorporate modern VRDs to bring more diversity and variability in input data.

When developing DU models, it is therefore important to consider the role of vision, language, and layout and how these are connected to the classification task. For example, current datasets are based on tobacco industry documents containing very domain-specific language, which a less robust classifier can overfit (*e.g.*, the spurious cue of a particular cigarette brand indicates an invoice). We highlight that document data can be multi-lingual, and code-switching is fairly common in document-based communications. For instance, an email may be in one language while the attachment is in another language.

In summary, future benchmarks must contain multipage, multi-type, multi-industry (*e.g.*, retail vs. medical invoice), multi-lingual documents with a wide range of document data complexities to build and test generic DC systems.

The community should explore potential solutions to the lack of adequate datasets for testing DC models such as i) leveraging public document collections, ii) synthetic generation, and iii) anonymization.

Public document collections: There are increasingly more (non-profit) organizations (*e.g.*, DocumentCloud), governments (SEC EDGAR), financial institutions (World Bank Documents & Reports), and charities (Guidestar) that make business-related documents publicly available for transparency in their operations and archival/research purposes. These collections provide datasets that are closer to real-world scenarios. However, these documents are typically unlabelled, although annotations could be crowd-sourced through combined funding from interested parties. Since most document data sources restrict automated crawling or document scraping, future dataset constructions will require some cooperation and creativity, whilst fulfilling licensing, ethical, and legal requirements. A specific highlighted initiative is CC-PDF [460], which collected modern, multi-lingual VRDs from CommonCrawl for future use.

Data synthesis: This alternative was suggested by prior work on KIE [30, 424] and DLA [37] for generating business and scientific documents. [422] followed up on this, delivering a large-scale KIE dataset with 6K real documents annotated and 100K synthetic examples. However, synthetic generation can be challenging to simulate real-world documents with similar data and classification complexity.

Anonymization can be a viable option to construct a DC dataset without compromising ethical guidelines and privacy regulations. This process involves removing, masking, replacing, or obfuscating data so that document content can no longer be attributed to an individual or entity. For example, one should remove names, addresses, and identifying information such as social security numbers or replace it with a textual tag ([SOCIAL-SECURITY-NUMBER]) or similar pattern (*e.g.*, Faker). While this process is not viable for creating KIE datasets, KIE can play a big role in semi-automatically anonymizing documents [143, 366]. Companies may be hesitant to make document collections public due to concerns about privacy, confidentiality and GDPR compliance. While anonymization can be an effective method, it should be approached with caution as potential risks of re-identification can make someone with originally good intentions legally liable. A potential side-step can be investing in privacy-preserving federated

learning (*e.g.*, PFL-DocVQA) to allow access to private industry document data.

4.5.4 Maturity of Evaluation Methodology

Most DC models are evaluated using predictive performance metrics such as accuracy, precision-recall, and F1-score on i.i.d. test sets. However, in user-facing applications, calibration can be as important as accuracy [156, 332, 340]. Even more so, when the confidence estimation of a DC is used to triage predictions to either an automated flow or manual processing by a human. Once a DC is in production, the i.i.d. assumption will start to break, which would recommend a priori testing of robustness against various sources of noise (OCR, subtle template changes, wording or language variations, ...) and expected distribution shifts (born-digital-scanning artifacts, shifting page order, page copies, irrelevant or out-of-scope documents, novel document classes, concept drift, ...).

Nevertheless, we observe only a few applications in DC (only reported on f_p) of more mature evaluation protocols [193] beyond predictive performance. Notable exceptions include covariate shift detection from document image augmentations [304], sub-class shift and generalization in [241, RVL-CDIP-N], out-of-distribution detection [241, RVL-CDIP-O], and cross-domain generalization [23, (RVL-CDIP $\leftrightarrow$ Tobacco-3482)]. However, the results on the latter can be misleading as both datasets are drawn from a similar source distribution. Another gap in DC benchmarking concerns evaluating selective classification [138, 193], which is closer to the production value evaluation of how many documents can be automated without any human assistance.

Another interesting evaluation protocol concerns *out-of-the-box* performance or how data-hungry/sample-efficient a certain model is. In practice, few-shot learning from minimal annotations is a highly valued skill. This few-shot learning evaluation protocol has been applied in [402] with different data regimes. Finally, inference complexity (time-memory) has been brought back to the attention of OCR-free models [216], which we believe will be the key to measuring when scaling solutions to multipage documents.

4.6 Chapter Conclusion

Our work represents a pivotal step forward in establishing multipage DC by proposing a comprehensive benchmarking and evaluation methodology. Thereby, we have addressed longstanding challenges and limitations (Section 4.5) that

have hindered progress in the field. As motivated in our experimental study, we have proven the need to advance multipage document representations and inference.

Following up on this, we provide recommendations for future DC dataset construction efforts pertaining to the type and nature of document data, variety in and quality of the classification label set, with a focus on particular DC scenarios closer to applications, and finally how future progress should be measured. Nonetheless, we are hopeful that the VDU community can come together on these shortcomings and apply the lessons from this reality check. Extending the applicability of current state-of-the-art models in VDU to multipage documents needs further exploration, which will go hand in hand with benchmark creation initiatives or incorporating multiple DC task annotation layers on a single dataset.

Chapter 5

Document UnderstanDing of Everything (DUDE 😎)

The contents of this chapter come from two publications [468, 469]:

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *International Conference on Document Analysis and Recognition*, pages 420–434. Springer, 2023

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Marie-Francine Moens, and Tomasz Stanisławek. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023

The first publication on the Document UnderstanDing of Everything (DUDE) competition was selected for **oral** presentation at ICDAR 2023. The second publication on the DUDE dataset and benchmark was featured as a poster presentation at ICCV 2023.
This multi-party collaboration (6 universities and 3 companies) with many brilliant researchers involved the creation of a new dataset and benchmark, the organization of a competition, and the publication of the results. For clarity, we will refer to the DUDE competition as the ICDAR 2023 competition, and the DUDE dataset and benchmark as the ICCV publication.
Author declarations: https://drive.google.com/file/d/1AmSxTOLk1Lo61sgWLd5FN5OMNQEgam_v
In short, I conceptualized the project, was responsible for the dataset creation, annotation, and benchmarking (encoder-only models, T5, HiVT5), designed evaluation and confidence estimation, and wrote the majority of the ICDAR and ICCV papers.

The dataset is available: https://huggingface.co/datasets/jordyvl/DUDE_loader.
Benchmark code is available: https://github.com/rubenpt91/MP-DocVQA-Framework.
The competition remains open for submissions at: https://rrc.cvc.uab.es/?ch=23.

Document **U**nderstan**D**ing of **E**verything (**DUDE**) is a concept rooted in both machine learning and philosophy, seeking to *expand* the boundaries of document AI systems by creating highly challenging datasets that encompass a diverse range of topics, disciplines, and complexities. Inspired by the philosophical 'Theory of Everything', which aims to provide a comprehensive explanation of the nature of reality, **DUDE** endeavors to stimulate the development of AI models that can effectively comprehend, analyze, and respond to *any* question on *any* complex visually-rich document (VRD).

Incorporating philosophical perspectives into **DUDE** enriches the approach by engaging with fundamental questions about knowledge understanding, and the nature of documents. By addressing these dimensions, researchers can develop AI systems that not only exhibit advanced problem-solving skills but also demonstrate a deeper understanding of the context, nuances, and implications of the information they process.

This chapter will present the Document UnderstanDing of Everything (DUDE) dataset, benchmark and competition. It will be presented in a similar form as the ICCV publication, extended with the results of the ICDAR competition. In line with the standpoint in the previous chapter, we call on the Document AI (DocAI) community to re-evaluate current methodologies and embrace the challenge of creating more practically-oriented benchmarks. This project aims to remediate the halted research progress in understanding visually-rich documents (VRDs). We present a new dataset with novelties related to types of questions, answers, and document layouts based on **multi-industry**, **multi-domain**, and **multipage** VRDs of various origins, and dates.

Moreover, we are pushing the boundaries of current methods by creating multi-task and multi-domain evaluation setups that more accurately simulate real-world situations where powerful generalization and adaptation under low-resource settings are desired. **DUDE** aims to set a new standard as a more practical, long-standing benchmark for the community, and we hope that it will lead to future extensions and contributions that address real-world challenges. Additionally, we present the results of the **DUDE** competition and discuss the innovations demonstrated by participants. The competition was structured as a single task with a multi-phased evaluation protocol that assesses the few-shot capabilities of models by testing generalization to previously unseen questions and domains, a condition essential to business use cases prevailing in the field. Under the newly studied settings, current SOTA models show a significant performance gap, even when improving visual evidence and handling multipage documents. We conclude that the **DUDE** dataset proposed in this competition will be an essential, long-standing benchmark to further explore for achieving improved generalization and adaptation under low-resource fine-tuning, as desired in the real world. To sum up, our work illustrates the importance of

finding more efficient ways to model language, images, and layout in DocAI.

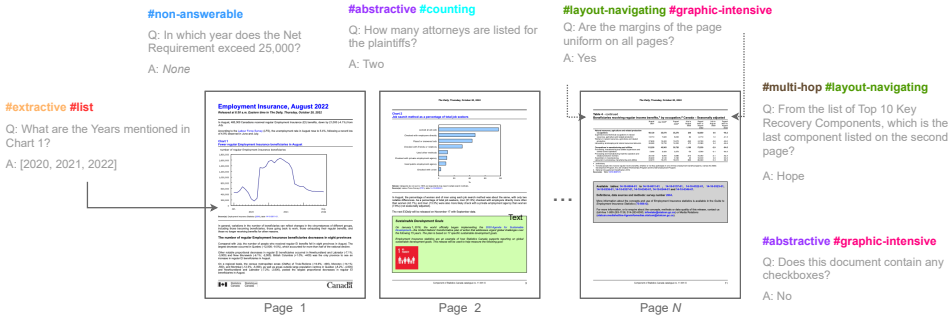


Figure 5.1. QA as a natural language interface to multipage VRDs.

5.1 Introduction

Early stages of research and growth in any field are characterized by enacting proof-of-concept and demonstrating the feasibility of the proposed solution. In the Deep Learning era, this is often echoed by building narrow and simplified datasets that do not reflect real-world complexity, leading to models that may not be suitable for practical use.

The field of Document Understanding (DU) is not an exception to the recent proliferation of deep architectures, which in this case are predominantly used for classification and information extraction from documents. However, the wide and complex nature of documents presents many challenges that remain unsolved or not yet addressed. One such challenge is domain generalization, where a model trained on medical documents may not be directly applicable to financial or tabular content. Another challenge concerns task-agnostic architectures, where a model must be able to adapt to various DU subtasks such as document classification, key information extraction (KIE), and question answering (QA). Lastly, the high variability of document contents and layouts often leads to highly imbalanced samples within document types, resulting in a long-tailed distribution with few or almost no samples to train a model.

Despite the importance of these challenges, there is currently no DU benchmark dataset that simultaneously addresses all of these issues. This paper proposes a novel dataset formulated as an instance of Document Visual Question Answering (DocVQA) to evaluate how well current DU solutions deal with multipage

documents, if they can navigate and reason over visual layouts, and if they can generalize their skills to different document types and domains.

The data collection and evaluation design of **DUDE** naturally motivates targeting models that can answer natural yet highly diverse questions (*e.g.*, regarding document elements, their properties, and compositions) for any VRD (*e.g.*, drawn from potentially unseen distributions of layouts, domains, and types). The presented problem setting relates to Multi-Domain Long-Tailed Recognition (MDLT) [507], which concerns learning from multi-domain imbalanced data whilst addressing label imbalance, divergent label distributions across domains, and possible train-test domain shift. Put plainly, since we cannot provide ground truth QA pairs for, *e.g.*, stamps, on every document type (domain), we expect a solution to transfer the subtask 'stamp detection' learned on document types where stamps naturally occur (and thus training QA pairs were created organically) to other domains. The DocVQA and MDLT formulations of **DUDE** allow us to create a longstanding, challenging benchmark that in the future can be easily extended with more subtasks formulated as QA pairs, and domains relating to document types (see Limitations).

The contribution of this work is twofold. First, we have created **DUDE**, a novel large-scale, multipaged, multi-domain, multi-industry DocVQA benchmark for evaluating DU progress. Second, we show that the zero-shot and fine-tuned performance of current SOTA models applied to DU lags far behind human baselines, explained in part by the need for more holistic and efficient modeling of language, vision, and richly structured layouts.

5.2 Related Work

Document Understanding encompasses datasets related to various subtasks like document layout analysis [261, 544], classification [165], key information extraction [197, 432], table extraction [427, 543, 545], and visual question answering [308, 315, 450]. These benchmarks lead to end-to-end DU architectures that have transformed common DocAI practices [15, 134, 153, 187, 263, 365, 371]. These task-specific benchmarks, however, are often tailored to a single domain, limiting the ability to create and assess how well DU models generalize to other document types and domains. To fill this gap, we adopt a visual question answering (VQA) approach, which has been crucial in the growth of the DU field.

The VQA paradigm provides a natural language interface for various tasks from both computer vision and natural language processing. In the latter, the question-answering approach has been successfully used in

several domains, including medicine [202, 209, 257, 318, 338, 352, 384], open-domain knowledge [281, 291, 313, 506], emotions [41, 155], code [7, 278], logical reasoning [282, 504, 516, 534], claim verification [185, 446, 523], and math [10, 65, 182, 316, 529]. As a result of its ability to function as a natural language interface for various forms of data, this paradigm has been applied to other domains. For example, the question-answering approach is combined with modalities such as images [13, 38, 39, 161, 353, 513], speech [237, 514], knowledge graphs [106, 206, 408, 429, 457], videos [58, 59, 74, 158, 249], and maps [60, 359].

Overall, the convergence of computer vision and NLP through the emergence of VQA tasks has also opened up new avenues for research in the DU field, with many DU datasets now including rich visual content alongside questions. Yet, prior study on document VQA has mainly focused on single-page documents [308, 310, 449] with rare exceptions such as MP-DocVQA [451]. However, [308, 449] pose only extractive questions where the answer follows the context on which the question is defined as in other question answering benchmarks [235, 386, 456]. Moreover, these datasets do not contain *non-answerable* questions as in established (natural language) QA datasets like [235, 387]. To the best of our knowledge there are no VQA datasets containing questions requiring lists as an answer. There are however few text-only QA datasets that contain such answer types [83, 256, 357]. Other datasets mainly related to our work are rather domain-specific like [310, 375, 440, 441, 551]. We give a detailed comparison of most related document VQA datasets in Table 5.1 highlighting the major contributions.

5.3 DUDE Dataset

While **DUDE** shares some similarities with existing VQA datasets, a closer comparison (see Table 5.1) highlights its unique features. We are confident that the model's proficiency in the areas introduced in this work will showcase its capability to handle the intricacy and diversity of document understanding tasks in real-world scenarios.

Documents. The dataset covers a wide range of document types, sources and dates, as shown in Table 5.1 and Figure 5.2 where its diverse nature is confirmed by the spread of document content representations. Moreover, it covers a broad range of domains, including medical, legal, technical, and financial, among others, to evaluate models' ability to handle diverse topics and the specific knowledge each requires. Furthermore, the dataset contains documents with varying layouts: diverse text arrangements, font sizes, and styles, to ensure that models can handle visually diverse documents.

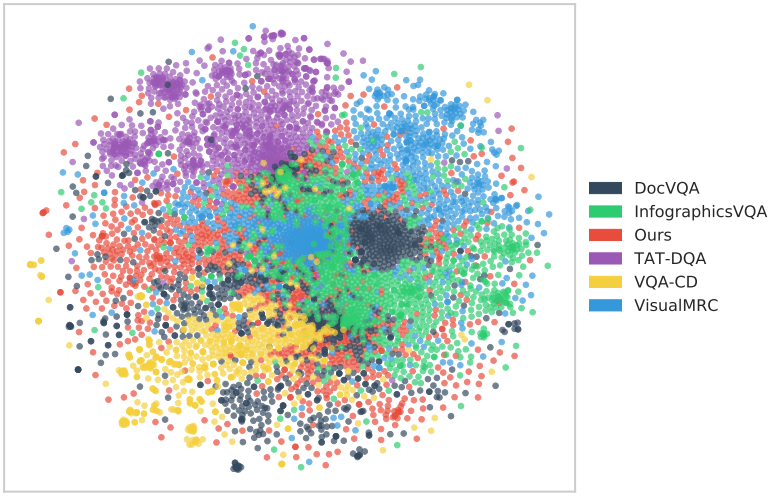


Figure 5.2. Visualization of inter-document similarities between samples from different datasets (t-SNE over TF-IDF representations of 1k passages from each source).

In contrast to our proposal, current VQA datasets often focus on homogeneous documents, such as invoices in VQA-CD [302] or financial reports in TAT-DQA [551]. Even when not restricted to a single domain or layout, these datasets share essential characteristics. For example, InfographicsVQA [310] demonstrates significant diversity in topics and designs, but still embodies a preference for visual aids over complex tables or long text passages. Moreover, VQA datasets are commonly restricted to either born-digital or scanned documents, which limits their ability to measure the robustness to mixed-origin files that one usually finds in real-world applications. In particular, this restriction makes it uncertain whether state-of-the-art performers on website fragments from VisualMRC [440] can be efficient on multi-column layouts and documents with OCR errors or incorrectly-detected reading orders. Finally, a typical dataset for document VQA contains documents from a limited period, i.e., a few years (Table 5.1).

Considering the properties mentioned above, the most diverse dataset to date is Single Page DocVQA (SP-DocVQA) [308], which contains mixed-origin documents of different types created over several decades. However, it is built exclusively on single-page document excerpts and is limited to several domains represented in the Industry Documents Library. As a result, it complements rather than serves as a touchstone for general-purpose DU systems. MP-DocVQA [451] extends this including previous and posterior pages of the documents. However, the questions are kept the same which makes the extra

pages mere distractors.

Questions. We use VQA as a natural language interface to VRDs, challenging the DU model with diverse questions, advanced operations, and multi-step reasoning to achieve real-world success.

Firstly, we assert that various layouts and visual elements must be comprehended semantically. As such, we introduce complex questions targeting these document elements, requiring comprehension beyond the document content, such as '*how many text columns are there?*', '*does the document contain words with diacritics*?' or '*which page contains the largest table in the document?*'. These layout-navigating questions bridge the gap between Document Layout Analysis and Question Answering paradigms.

Our unique and detailed compositional questions demand a model that comprehends semantics and generalizes to new questions in a zero-shot setting. For example, >90% of our questions are unique, while we target questions whose answer scope is much more diverse than in previous works.[1] Since neural networks are known to perform poorly at mathematical reasoning and symbolical processing, we provide training and evaluation questions demanding arithmetic and comparison operations on numbers and dates.

Moreover, we feature multi-hop questions that indicate a model's robustness to sequential reasoning and mimic how humans ask questions. They may be useful in real-world tasks such as '*If the checkbox on page 1 section 3a indicates that the company is incorporated, how much yearly revenue did it generate in 2022 (given the table on page 5)*?'

Answers. Even though some VQA datasets are deliberately limited to questions of exclusively extractive (SP-DocVQA) or abstractive (VisualMRC) nature, others do not obey such restrictions and include both question types (see Table 5.1). The dataset we provide includes both abstractive and extractive answers, covering various types such as *textual, numerical, dates, yes/no, lists, or no answer*.

This allows us to cover all possible business use cases and reveal major deficiencies of existing DU systems beyond typical textual answers. For instance, no existing VQA dataset includes not answerable questions and questions answered with a list. In turn, the models considered to date supposedly tend to make unreliable guesses on questions with an answer not entailed by the content [387]. Our

[1] Answer type comparison is included in supplementary materials.

dataset is designed to cover answers beyond plain extractive text such as a list of items or even 'None'.

The 'None' answer type demands that the model correctly identifies that the answer cannot be provided, as the question needs to be better formed, *e.g.*, it asks about the value of an empty cell in the table. In addition, list generation problems pose challenges to the model, as (1) more tokens need to be generated, (2) they may be sourced from different places in the document, and (3) OCR reading order may influence the element ordering.

5.3.1 Gathering Documents

A fundamental difficulty in gathering raw source files was ensuring dataset diversity while fulfilling strict licensing requirements. Therefore, rather than depending on initial sources of files, *e.g.*, libraries that originally published digitized materials, we resorted to aggregate websites.

The document collection process was manual and assumed formulating queries to archive.org (containing 36M books and texts), commons.wikimedia.org (with 86M media types of various types), and documentcloud.org (with around 5M public documents). The queries consisted of keywords relevant to some category of interest, *e.g.*, the *resume* category of our proposal consists of 'resume', 'cv', 'curriculum', and 'biography' keywords). Where necessary, a separate query parameter ensured that the resulting files belonged to the public domain or were released under a permissive license. Information on keywords and the search procedure is distributed as a part of the DUDE dataset.

From the resulting documents, we selected those representing the requested category and visually distinctive from the ones already gathered. Special care was put into removing examples that visibly expose controversial content or may be subject to privacy or legal concerns, despite the declared license. We collected five thousand, typically multipage, English documents using this methodology.

5.3.2 Annotation Process

The annotation process involved in-house annotators and Amazon Mechanical Turk freelancers. For the latter, there is limited control over the expertise, and where justified, we resorted to limiting task availability depending on the number of completed tasks and historical acceptance rate.[2] The former are five highly qualified people with a Ph.D. in Linguistics. These three annotation

[2]Approval above 97% over at least 5k HITs.

scenarios will be referred to as *All MTurkers*, *Best MTurkers*, and *Qualified Linguists*.

We estimate the total cost of annotation involving both *Linguists* and *MTurkers* as $20,000.

Phase 1. We started by providing *All MTurkers* documents described in Section 5.3.1 in separate batches aimed at collecting abstractive, extractive, and list QA pairs. Each freelancer was asked to propose up to five questions of a particular type, and in the case of extractive ones to provide an evidence bounding box. The exception to this process is the annotation of non-answerable questions previously shown to be particularly challenging [387]. These are predominantly annotated by *Qualified Linguists* and because of their quality promoted without passing through Phases 2-3.

Candidate QA pairs are semi-automatically filtered to exclude annotations that cannot be valid due to the length, use of non-typical character combinations, or type-specific criteria, such as non-list answers for list batches. Additionally, we cluster duplicate and near-duplicate question-answer pairs to ensure dataset diversity and promote them directly to Phase 3 after a manual review (the same QA pairs provided independently by several annotators indicate their validity).

Phase 2. The rest of the annotations promoted from Phase 1 were directed to *All MTurkers*, but this time instead of providing complete QA pairs, they were asked to answer the question from the previous round. Obtained triples of questions and two answer variants (one from each phase) were evaluated using inter-answer ANLS (defined in Section 5.3.5) promoted to the final dataset if the agreement was >0.8. Otherwise, QA triples were directed to Phase 3.

Phase 3. *Best MTurkers* were provided with document, question, and answer variants to decide the correctness of each answer and optionally overrule both variants if they are not correct. Outliers from decisions in this phase, such as repealing without a judgment on previous answers, were reviewed by *Qualified Linguists* and corrected if needed.

Optional Phase 4. Annotations of the test set were reviewed by *Qualified Linguists*. Given data from Phase 3, they corrected questions, answers and created metadata related to diagnostic categories described in Section 5.3.4.

5.3.3 Dataset Statistics

Dataset	Ours	SP-DocVQA	VisualMRC	InfographicsVQA	TAT-DQA
		Dataset-level properties			
Sources	Multi	Industry docs	Web pages	Infographics	Finance reports
Origin	BD, Scan	Mostly scans	BD	BD	BD
Period	1860-2022	1960-2000	Jan-Mar 2020	not specified	2018-2020
Documents	5,019	12,767	10,234	5,485	2,758
Pages (*avg±std*)	5.72±6.4	1.0±0.0	1.0±0.0	1.0±0.0	1.11±0.32
Tokens (*avg±std*)	1,831.53±2,545.06	183±149.96	154.19±79.34	287.98±214.57	576.99±290.12
Simpson coeff. (ResNet)	0.82	0.76	0.83	0.86	0.73
Simpson coeff. (Tf-Idf)	0.95	0.93	0.99	0.94	0.15
		Question-level properties			
Questions	41,541	50,000	30,562	30,035	16,558
Unique (%)	90.9	72.34	96.26	99.11	95.65
Length (*avg±std*)	8.65±3.35	8.34±3.04	9.38±4.01	11.57±3.71	12.51±4.18
Semantics	All	T, L, F, Ch	T, L, F, Ch	T, L, F, Ch, M	T, L
		Answer-level properties			
Unique (%)	70.7	64.29	91.82	48.84	77.54
Length (*avg±std*)	3.35±6.1	2.11±1.67	8.38±6.36	1.66±1.43	3.44±7.20
Extractive (%)	42.39	100.0	0.0	71.96	55.72
Abstractive (%)	38.25	0.0	100.0	24.91	44.28
List (%)	6.62	0.0	0.0	5.69	0.0
None	12.74	0.0	0.0	0.0	0.0

Table 5.1. Summary of the existing English document datasets and our challenge. `BD` stands for born-digital. Layout semantics are abbreviated as (T)able, (L)ist, (F)igure, (Ch)art, and M(ap). Comparison based on Azure Cognitive Services (3.2) OCR.

We conducted a statistical analysis of our dataset and found that the distribution of document length, question length, and answer type was much more diverse than in other datasets in the same domain. We also used the Simpson diversity coefficient [421] for analysis and summarized the results in Table 5.1. The following are the statistics for the data split:

	train	val	test (diagnostic)
documents	3,010	749	1,215 (530)
questions	23,728	6,315	11,448 (2,462)

Table 5.2. Data split counts.

The number of tokens in the document distribution is much more diverse compared to other datasets, a consequence of the more diverse distribution of pages (see Figure 5.4). Note some of the documents are more visual than textual (or even visual-only), making the left whisker essentially reach 0 ($\log_2$-scaling of x-axis).

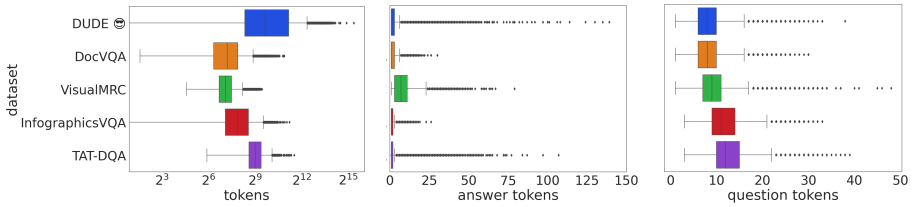


Figure 5.3. Distribution of the number of tokens in documents, answers, and questions.

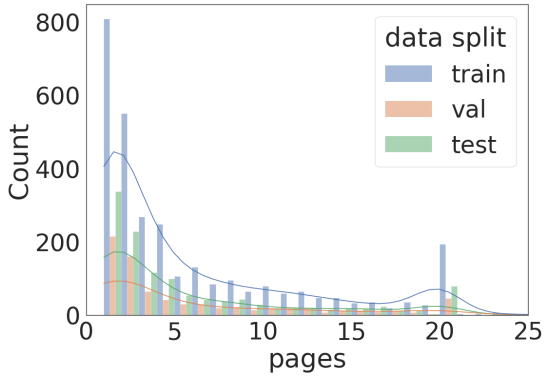


Figure 5.4. While other datasets are predominantly single-page only, the number of pages featuring in **DUDE** is more diverse, yet still biased towards shorter documents.

The distribution of the number of tokens in answers is heavy-tailed, to some extent this is also the property of the distribution of number of tokens in questions. Furthermore, 90.9% of questions are unique, and so are 70.7% of answers (taking answer variants into account).

We scrutinized the answer types by aggregating possible answers into classes representing the information they conveyed. The study used heuristics to determine if the answers fit into NER labeling scheme [20] or categories we anticipated, such as *yes/no* and *none*, or did not anticipate, such as *color*. This resulted in 25 different groups of answers, with the *other* answer type being the fourth largest group. Cramer's V coefficient was used to check for correlations between question types and answer types, and the results indicated that there were few correlations . The expected correlations, such as *none* answers with *not-answerable* questions or *yes/no* answers with *abstractive* questions, were present, but barely any correlation was significant. This suggests it is hard to guess the answer based on the question solely.

We study relative diversity measure, called Simpson coefficient [421, 546]. To

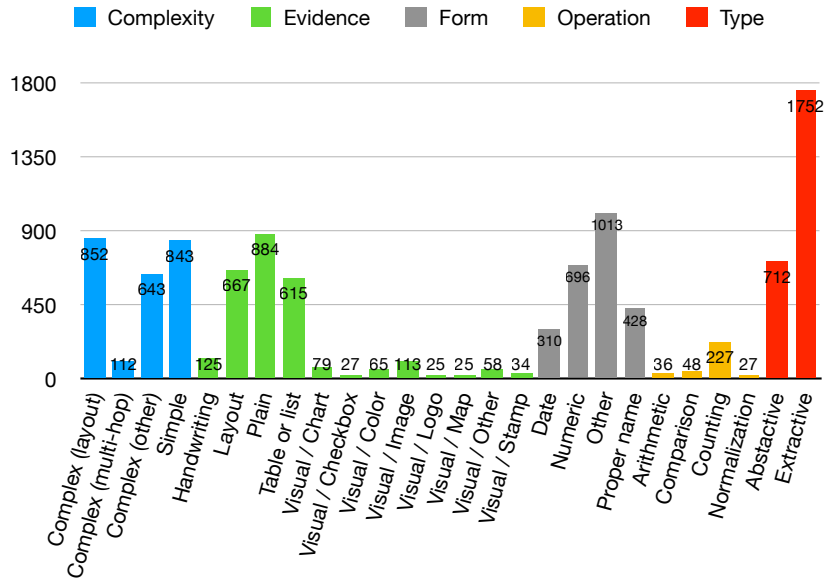


Figure 5.5. Count of particular diagnostic categories in a subset of 2.5k test set QA pairs annotated in detail to help analyze models' performance.

define it, consider a fixed distance function $d(a_1, a_2)$ defined for pair of documents $a_1, a_2 \in A$: the dataset. In our applications, it is the cosine similarity of a document embedding. Further, for an arbitrary number of datasets $A_1, \ldots, A_N$ the diversity of A_1 with respect to $A_2, \ldots, A_N$ is defined as

$$\text{Div}_{A_2,\ldots,A_N}(A_1) = 1 - p\Big(d(a_{11}, a_{12}) < \min_{i=2:N} d(a_{i1}, a_{i2})\Big)$$

where $a_{i1}, a_{i2} \in A_i$, are randomly selected, $i = 2 : Ni = 2 : N$. We report relative diversities of each of the datasets, relative to other datasets in the study, based on two embeddings: visual (ResNet-101 embeddings-based) and semantic (Tf-Idf embeddings-based), in Table 5.1. The results show that the probability that two random documents from **DUDE** are more similar than each random pair of documents from other datasets is small, meaning that documents in our dataset are well-distributed and diverse.

5.3.4 Diagnostic Subsets

Following previous DU datasets, we gather diagnostic metadata for close to half of the documents and QA pairs in the test set (see Figure 5.5). These

are intended to enable a fine-grained analysis of the models' performance. The taxonomy used is an extension of the one from earlier works [47, 308, 310], covering **DUDE**-specific questions and enables a more detailed examination of visual artifacts under consideration.

Question type and perceived complexity. We distinguish questions perceived as *simple*, i.e., those based on spotting value near a phrase mentioned explicitly as a part of the question. For example, "Who is the Secretary of the U.S. Department of Commerce?" when the document contains "Penny Pritzker, Secretary, U.S. Department of Commerce." Such could be guessed given an approximate string matching algorithm and does not require much comprehension beyond that. The remaining questions are marked as *hard* with distinguished categories of *hard multi-hop questions*, and *hard meta/layout-navigating questions*.

Answer evidence. We provide information on what types of elements have to be comprehended to provide an answer, including *free text*, *handwriting*, *table or list*, and *layout*, i.e., non-tabular spatial understanding of text placement. These follow the ontology established by previous works [47, 308, 310]. In addition, we supply hints on graphical artifacts one needs to consider for particular questions, such as *image/photo*, *plot/chart*, *checkbox*, and *annotation*.

Required operation. We distinguish *arithmetic*, *comparison*, *counting*, and *normalization* operations to provide information on the need for performing, respectively, arithmetic operations on extractable data, comparing numerical values or sizes, counting elements or converting data present in the document to another format (*e.g.*, rounding or date format conversion).

Answer form/shape. Finally, we provide information on the shallow form of the returned answer, including *date*, *numeric*, and *proper name*.

5.3.5 Evaluation

The evaluation process follows the typical paradigm of separate training, validation, and test splits. We provide both a standalone evaluator and a website[3] [467] to submit test set predictions.

[3] rrc.cvc.uab.es/?ch=23

To assess models' performance, we rely on the ANLS metric introduced by authors of the ST-VQA dataset [39]. Roughly speaking, it is a generalization of accuracy that does not penalize the system for an answer whose similarity to the gold standard measured with normalized Levenshtein similarity is above a specified threshold. Moreover, the metric assumes the presence of multiple, equally valid reference answers. The mentioned properties account for possible OCR errors or different phrasings, such as the same numerical answer represented as *two* and *2* by different annotators.

In practice, production DU systems provide an estimation of confidence in order to triage documents that do not need to be manually reviewed by a human. While the reliability of the automation ability of a DU solution is deemed quintessential for generating business value in practice [48], DU research rarely reports any confidence evaluation. Some exceptions are in closely related task domains like scene text recognition [425] and QA [208, 531].

With DUDE, we want to establish calibration evaluation and confidence ranking as a default evaluation methodology in DU, especially since the field is so close to applications.

To this end, we report (next to ANLS) two additional metrics, Expected Calibration Error (ECE) [156, 332, 340], and Area-Under-Risk-Coverage-Curve (AURC) [138, 193].

Calibration requires that the probability a model assigns to its predictions equals their true likelihood of being correct [86, 88, 520].

ECE approximates top-1 calibration error by a weighted average over the accuracy/confidence difference of histogram bins. Particularly in our evaluation setting, we consider a predicted answer correct if its ANLS to the ground truth answer is above a pre-defined threshold (τ=0.5). For consistency, not-answerable and list-answers both have confidence estimated for the answer as a whole (regardless of the number of answers). Following [342], we apply equal-size binning (with 100 bins, $\mathcal{L}_p = 1$), avoiding some pathologies of equal-range binning [231, 463].

AURC is a selective classification metric that evaluates how well an estimator prevents silent failures on an *i.i.d* test set. As an aggregate measure of estimator performance (ANLS) and confidence ranking, it provides a more practically useful estimate of overall performance when the estimator can abstain from (low-confidence) decisions and defer to a human for feedback.

By reporting the above metrics, we hope that in future work there will be contributions (*e.g.*, calibration methods for improved forecasting or metrics for better predictive uncertainty evaluation) that concretely target the empirical

observations of overconfidence/miscalibration in DU models.

5.4 DUDE Competition

Over the past few years, the field of Document Analysis and Recognition (DAR) has embraced multimodality with contributions from both NLP and CV. This has given rise to DU as the all-encompassing solution [15, 187, 371] for handling VRDs, where layout and visual information is decisive in understanding a document.

This umbrella term subsumes multiple subtasks ranging from KIE [197, 432], DLA [544], VQA [310, 450], table recognition [201, 376], and so on. For each of these subtasks, influential challenges have been proposed, *e.g.*, the ICDAR 2019 Scene Text VQA [38, 39] and ICDAR 2021 Document VQA (DocVQA) [308, 450] challenges, which in turn have generated novel ideas that have impacted the new wave of architectures that are currently transforming the DAR field.

Nevertheless, we argue that the DAR community must encompass the future challenges (multi-domain, multi-task, multipage, low-resource settings) that naturally juxtapose the previous competitions with pragmatic feedback attained via its business-driven applications.

5.4.1 Challenge Objectives

We aim to support the emergence of models with strong multi-domain layout reasoning abilities by adopting a diversified setting where multiple document types with different properties are present. Moreover, a low-resource setting (number of samples) is assumed for every domain provided, which formulated as a DocVQA competition allows us to measure progress with regard to the desired generalization (Section 5.4.3.1). Additionally, we strive for the development of confidence estimation methods that can not only improve predictive performance but also adjust the calibration of model outputs, leading to more practical and reliable DU solutions.

We believe that **DUDE**'s emphasis on task adaptation and the capability of handling a wide range of document types, layouts, and complexities will encourage researchers to push the boundaries of current DU techniques, fostering innovation in areas such as multimodal learning, transfer learning, and zero-shot generalization.

5.4.2 Challenge Contributions

DUDE answers the call for measuring improvements closer to the real-world applicability of DU models. By design of the dataset and competition, participants were forced to make novel contributions in order to make a significant impact on the DU task. Competitors showcased intriguing model extensions, such as combining models that learn strong document representations with the strengths of recent large language or vision-language models (ChatGPT [52] and BLIP2 [258, 260]) to better understand questions and extract information from a document context more effectively. HiVT5 + modules extended Hi-VT5 [451] with token/object embeddings for various DU subtasks, while MMT5 employed a two-stage pretraining process and multiple objectives to enhance performance. These innovative extensions highlight the ingenuity in addressing the complex challenges of document understanding.

5.4.3 Motivation and Scope

We posit that progress in DU is determined not only by the improvements in each of its related predecessor fields (CV, NLP) but even more by the factors connecting to document intelligence, as explicitly understood in business settings. To improve the real-world applicability of DU models, one must consider (i) the availability and variety of types of documents in a dataset, as well as (ii) the problem-framing methods.

Currently, publicly available datasets avoid **multipage** documents, are not concerned with **multi-task** settings, nor provide **multi-domain** documents of sufficiently different types. These limitations hinder real-world DU systems, given the ever-increasing number of document types occurring in various business scenarios. This problem is often bypassed by building systems based on private datasets, which leads to a situation where datasets cannot be shared, documents of interest are not covered in benchmarks, and published methods cannot be compared objectively. **DUDE** counters these limitations by explicitly incorporating a large variety of multipage documents and document types. Furthermore, the adaptability of DU to the real world is slowed down by a low-resource setting, since only a limited number of training examples can be provided, involving unpleasant manual labor, and subsequently costly model development. Anytime a new dataset is produced in the scientific or commercial context, a new model must be specifically designed and trained on it to achieve satisfactory performance. At the same time, transfer learning is the most promising solution for rapid model improvements, while zero- and few-shot performance still needs to be addressed in evaluation benchmarks.

Bearing in mind the characteristics outlined above, we formulated the **DUDE** dataset as an instance of *DocVQA* to evaluate how well current solutions can simultaneously handle the complexity and variety of real-world documents and all subtasks that can be expected. Optimally, a DU model should understand layout in a way that allows for zero-shot performance through attaining "desired generalization", i.e., generalization to *any documents* (*e.g.*, drawn from previously unseen distributions of layouts, domains, and types) and *any questions* (*e.g.*, regarding document elements, their properties, and compositions). Therefore, we incorporated these criteria while designing our dataset, which may stand as a common starting point and a cooperative path toward progress in this emerging area.

5.4.3.1 Desired Generalization.

The challenge presented by **DUDE** is an instance of a Multi-Domain Long-Tailed Recognition (*MDLT*) problem [507].

Definition 14 (Multi-Domain Long-Tailed Recognition)**. MDLT** *focuses on learning from multi-domain imbalanced data whilst addressing label imbalance, divergent label distributions across domains, and potential train-test domain shift. This framework naturally motivates targeting estimators that generalize to all domain-label pairs.*

A *domain* $D = \{(x_i, y_i)\}_{i=1}^{N}$ is composed of data sampled from a distribution P_{XY}, where $\mathcal{X}$ denotes an input space (documents) and $\mathcal{Y}$ the output space (QA pairs). Each $x \in \mathcal{X}$ represents a document, forming a tuple of (v, l, t), expressing a complex composition of **v**isual, **l**ayout and **t**extual elements. For simplicity, consider that each 'label' $y \in \mathcal{Y}$ represents a **q**uestion-**a**nswer pair, relating to implicit tasks to be completed (such as date KIE in *What is the document date?*). Due to the potentially compositional nature of QA, the label distribution is evidently *long-tailed*. During training, we are given M domains (*document types*) on which we expect a solution to generalize (Figure 5.6), both within (different number of samples for each unique task) and across domains (even without examples of a task in a given domain).

What sets apart domains is any difference in their joint distributions $P_{XY}^{j} \neq P_{XY}^{k}$. For example, an invoice is less similar (in terms of language use, visual appearance, and layout) to a contract than to a receipt or credit note. Yet, a credit note naturally contains a stamp stating information such as "invoice paid", whereas receipts rarely contain stamps. This might require a system to transfer 'stamp detection' learned within another domain, say on notary deeds.

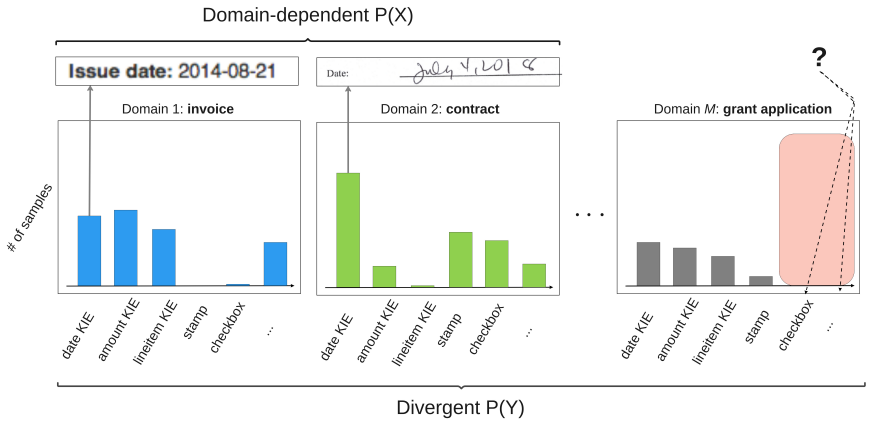


Figure 5.6. Illustration of MDLT as applicable to the **DUDE** problem setting. The y-axis aggregates skills related to specific KIE or reasoning tasks over document elements (checkbox, signature, logo, footnote, ...). The x-axis denotes the obtained samples (QA pairs) per task. Each domain has a different label distribution $P(Y)$, typically relating to within-domain document properties $P(X)$.This training data exhibits label distribution shifts across domains, often requiring zero-shot generalization (marked red).

Notably, it will be 'organic' to obtain more examples of certain questions (*tasks*) in a given domain. This should also encourage models to learn a certain skill in the domains where they have more training examples. Put plainly, it is better to learn checkbox detection on contracts than on invoices, which rarely contain any. This MDLT framework allows us to create a lasting, challenging benchmark that can be easily extended in the future with more tasks (formulated as QA pairs) and domains (relating to document types). In the first iteration of the **DUDE** competition, we have targeted specific skills by guiding annotators with focused instructions, which we share for future extensions.

5.4.4 DUDE Competition Protocol

The ICDAR 2023 competition on Document UnderstanDing of Everything took place from February to May of 2023. A *training-validation* set with 30k QA annotations on 3.7k documents was given to participants at the beginning of February. The 11.4k questions on 12.1k documents for the *test set* were only made accessible for a window between March and May. Participants were asked to submit results obtained on the public, blind test set documents rather than deliver model executables, although they were encouraged to open-source

their implementations. We relied on the scientific integrity of the participants to adhere to the competition's guidelines specified on The Robust Reading Competition (RRC) portal[4].

5.4.4.1 Task Formulation

Given an input consisting of a PDF with multiple pages and a natural language question, the objective is to provide a natural language answer together with an assessment of the answer confidence (a float value scaled between 0 and 1). Each unique document is annotated with multiple questions of different types, including extractive, abstractive, list, and non-answerable. Annotated QA pairs are not restricted to the answer being explicitly present in the document. Instead, any question on aspect, form, or visual/layout appearance relative to the document under review is allowed.

Additionally, competitors were allowed to submit results for only a specific answer type (provided in annotations) such that, for example for extractive questions, encoder-only architectures could compete in **DUDE**. Another important subtask is to obtain a *calibrated* and *selective* DocVQA system, which lowers answer confidence when unsure about its answers and does not hallucinate in case of non-answerable questions. Regardless of the number of answers (zero in the case of non-answerable or multiple in list-questions), we expect a single confidence estimate for the whole answer to guarantee consistency in calibration evaluation. To promote fair competition, we provided for each document three OCR versions obtained from one open-source (Tesseract) and two commercial engines (Azure, AWS).

5.4.4.2 Evaluation Protocol

The first evaluation phase assumes only independently and identically distributed (i.i.d.) data containing a similar mixture of document and question-answer types for the train-validation-test splits. The same evaluation metrics as the benchmark apply for this phase.

The (implicit) second evaluation phase created a mixture of seen and unseen domain test data. This was launched jointly with the first evaluation phase, as otherwise, one would be able to already detect the novel unseen domain test samples. To score how gracefully a system deals with unseen domain data, the evaluation metric is AUROC [270], which roughly corresponds to the probability that a positive example (in-domain) is assigned a higher detection score than

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

a negative example (out-of-domain). A system is expected to either lower its confidence or abstain from giving an answer.

There is a strict difference between a non-answerable question and an unseen domain question. For the former, the document is from a domain that was included during training, yet the question cannot be solved with the document content, *e.g.*, asking about who signed the document without any signatures present. For the latter, the question is apt for the document content, yet the document is from a domain that was not included during training and validation, which we would expect the system to pick up on.

All metric implementations and evaluation scripts are made available as a standalone repository to allow participants to evaluate close to official blind test evaluations[5].

All submitted predictions are automatically evaluated, and the competition site provides ranking tables and visualization tools newly adapted to PDF inputs to examine the results. After the formal competition period, it will serve as an open archive of results. The main competition winner will be decided based on the aggregate high scores for ANLS, AURC, and AUROC.

5.5 DUDE Benchmark

5.5.1 Baselines

Human performance. To establish the human baseline, we assign test set questions to *Qualified Linguists*, ensuring none of them will face the same documents as reviewed in Phase 4. The procedure results in an estimation of 74.76 ANLS points (Table 5.3). At first glance, this result seems low. Still, when analyzing results case by case, it turns out that it's hard to score much better since the answer format can influence the overall results a lot: *Eagle* vs. *an eagle* (0.625 ANLS), *62%* vs. *62* (0.67 ANLS), *1958-04-29* vs. *4-29-58* (0 ANLS), *Clemson University, Clemson South Carolina* vs. *Clemson University* (0 ANLS). We achieved the lowest performance (67.58) on the extractive question type, which confirms our hypothesis since the abstractive answers are shorter (mostly numbers, yes/no, or colors).

We analyzed the maximum score achieved by the best-performing model for each diagnostic test category and plotted that against the human performance in Figure 5.7.

[5] https://github.com/Jordy-VL/DUDEeval

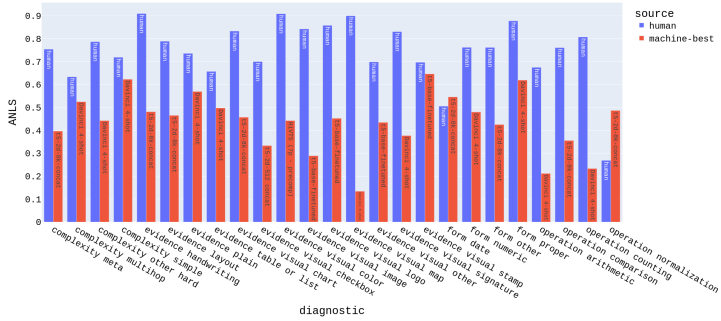


Figure 5.7. We report the average ANLS for the human expert vs. the best-performing model per diagnostic category as a ceiling analysis.

Reference models. We assessed a group of models to determine how their performance is influenced by different factors such as (1) their ability to handle textual, layout, and visual elements, (2) whether they were fine-tuned for the task, (3) their size in (trainable parameters), and (4) the maximum input length they can handle.

To analyze factors (1) and (2), we conducted a zero-shot evaluation of several baseline text-only models. We used three encoder-based models (BERT [94], Longformer [28], and BigBird [521]) that cannot generate text and three that feature a decoder (T5 [383], GPT-3-Davinci [52], and ChatGPT) and have this capability. Next, we extended the T5 architecture with 2D layout embeddings [47, 371] and fine-tuned models with increasing maximum sequence lengths (512 $\rightarrow$ 8192) on **DUDE**. Finally, we evaluated our replication of the hierarchical Hi-VT5 model [451], as this model has the ability to decode text, understand multipage layouts, and comprehend visual page features using DiT [259].

Regarding factors (2) and (3), we evaluated models of various sizes ranging from 131M (BigBird) to 175B (GPT-3-Davinci) and varied the input context from 512 (BERT) to 20480 (Hi-VT5) tokens. Overall, we thoroughly evaluated multiple models in the different testing setups to determine their performance under various conditions, as seen in Table 5.3.

5.5.2 Analysis & Discussion

To summarize, our study reveals that existing advanced language models such as BERT, Longformer, and BigBird struggle with comprehending visual elements and document layouts. To address this issue, we introduced T5, T5-2D, and

Model	Init.	Params	Max Seq. Length	Test Setup	$ANLS_{all}$ ↑	ECE_{all} ↓	$AURC_{all}$ ↓	$ANLS_{do}$	$ANLS_{do}$ Abs	$ANLS_{do}$ Ex	$ANLS_{do}$ NA	$ANLS_{do}$ Li
text-only Encoder-based models												
Big Bird	MPDocVQA	131M	4096	Concat*	26.27	30.14	44.22	30.67	7.11	40.26	12.75	8.46
BERT-Large	MPDocVQA	334M	512	Max Conf.*	25.48	34.06	48.60	32.18	7.28	42.23	5.88	11.13
Longformer	MPDocVQA	148M	4096	Concat*	27.14	27.59	44.59	33.45	8.55	43.58	10.78	10.62
text-only Encoder-Decoder based models												
T5	base	223M	512	Concat-0*	19.65	19.14	48.83	25.62	5.24	33.91	0	7.31
T5	MPDocVQA	223M	512	Max Conf.*	29.48	27.18	43.06	37.56	21.19	44.22	0	10.56
T5	base	223M	512	Concat+FT	37.41	**10.82**	41.09	40.61	42.61	48.20	53.92	16.87
T5	base	223M	8192	Concat+FT	41.80	17.33	49.53	44.95	47.62	50.49	63.72	7.56
text-only Large Language models (LLM)												
ChatGPT	gpt-3.5-turbo	20B	4096	Concat-0	-	-	-	35.07	16.73	42.52	70.59	15.97
				Concat-4	-	-	-	41.89	22.19	49.90	**77.45**	17.74
GPT3	davinci3	175B	4000	Concat-0	-	-	-	43.95	18.16	54.44	73.53	36.32
				Concat-4	-	-	-	47.04	22.37	**57.09**	63.73	**40.01**
text+layout Encoder-Decoder based models												
T5-2D	base	223M	512	Concat+FT	37.10	10.85	41.46	40.50	42.48	48.62	52.94	3.49
T5-2D	base	223M	8192	Concat+FT	42.10	17.00	48.83	45.73	48.37	52.29	63.72	8.02
T5-2D	large	770M	8192	Concat+FT	**46.06**	14.40	**35.70**	**48.14**	**50.81**	55.65	68.62	5.43
text+layout+vision models												
HiVT5		316M	20480	Hierarchical+FT	23.06	11.91	54.35	22.33	33.94	17.60	61.76	6.83
LayoutLMv3	MPDocVQA	125M	512	Max Conf.*	20.31	34.97	47.51	25.27	8.10	32.60	8.82	7.82
Human baseline								74.76	81.95	67.58	83.33	67.74

Table 5.3. 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.

Hi-VT5 models that incorporate layout and visual information. Still, their performance remains unsatisfactory, as evidenced by the comparison with the human baseline, similar to what has been reported for InfographicsVQA. This indicates that there is still scope for enhancing the visual understanding of **DUDE** models. Moreover, our findings indicate that a large LLM capable of processing long inputs alone is insufficient for achieving strong performance in **DUDE**, especially for the extractive type of answer. Finally, the dataset's length significantly affects the models' scores, as seen by the increase in scores by $4.4 - 5.0$ points when the T5 and T5+2D context length is extended from 512 to 8192. Similarly, the model size has a positive correlation with the final score, but it holds only within a particular model-type and is not the main factor influencing the results. State-of-the-art performance of 46.04 $ANLS_{all}$ was achieved on $T5_{large}$ with a 2D layout understanding that consumed 8192 tokens, confirming the observation above.

5.6 Detailed Results Analysis

5.6.1 Within Model Class Analysis

5.6.1.1 Encoder vs. Decoder

A key difference between encoder-only and (encoder-) decoder-based models is the ability to generate answers beyond the explicit document textual content. This is clearly reflected in the results for BigBird, Longformer, BERT, and LayoutLMv3, which score < 10 ANLS% on abstractive questions, whereas they have just average scores for extractive questions. On **DUDE**, we can claim that a generative model is necessary given all considered question types.

Quite remarkably, while the human baseline demonstrates that humans find abstractive questions (ANLS ±82%) easier than extractive questions (ANLS ±68%), the reverse is true for all machine baselines. A potential confounder for these results could be the difference in output formatting for extractive vs. abstractive answers, which is hard to take into account with ANLS evaluation.

5.6.1.2 Incorporating Layout & Vision

When comparing T5 with and without 2D position embeddings on the diagnostic categories, we find the highest improvements on 'evidence table or list', 'complexity simple', and 'evidence plain'.

Our study with the proposed baselines shows that questions requiring visual evidence to be answered are an important future challenge for the vision community. To get further insights into models' performance on these questions, we calculate a weighted average of ANLS over visual categories. This reveals that GPT3 (4-shot) and T5-2d-large-8K obtain a tied score of (ANLS=37%), even though they only have access to the text. The human performance, on the other hand, is close to double (ANLS=72%), thus showing the need for better integration of the visual modality in DU models.

5.6.1.3 Toward Long Document Processing

DUDE clearly requires methods that can process long sequences, as evidenced by its average document length of 1832 ± 2545 tokens. This is particularly evident when comparing standard NLP QA methods like BERT-concat, which underperforms Longformer [28] and BigBird [521], despite being the *large* version.

Experiments with T5 and T5-2D further support this claim, as extending the sequence length from 512 to 8192 leads to a $\sim 5\%$ ANLS improvement.

The exception is HiVT5 [451], which performs worse than the rest of the methods. This is due to the authors of HiVT5 performing a pretraining task of text denoising that helped to better model the [PAGE] tokens. This resulted in a better, compressed representation of the relevant information within a document conditioned by a question. Moreover, the authors also did extensive experimentation and found that 10 [PAGE] tokens per page were the best fit for the MP-DocVQA [451] dataset. We used similar hyperparameters, but **DUDE** might require better fine-tuning of [PAGE] tokens since the images are more visually rich with colored graphics and layouts. The hierarchical processing of documents with a meaningful visual component is a promising avenue for future research.

5.6.1.4 Diagnosis of LLM Results

The reasoning for including these LLMs as baselines stems from our question: "Does advanced text understanding suffice for solving **DUDE**?". Our results for diagnostic categories reveal some strengths and weaknesses of LLMs in the DocVQA task setting.

Strengths GPT3 trumps all other tested models for list-type questions (ANLS=36-40%), which can be explained by the extractive nature of these questions. After 4-shot fine-tuning, ChatGPT (4-shot) is better than all other tested baselines in answering not-answerable questions (ANLS=77.45%). This can partly explain the appeal of this particular GPT checkpoint in recent times. GPT3 (4-shot) outperforms (ANLS=52.51%) other tested baselines on questions from the 'complexity multi-hop' category such as *What city name appears the most often in the timetables?*.

Weaknesses Compared to another (more simple text-only generative baseline, T5-base-512 (ANLS=47%), LLMs perform two times worse on abstractive questions (ANLS=22%). Closer analysis reveals that LLMs (even after 4-shot fine-tuning) predict abstractive questions to be *Not-answerable* in 55% of cases (in reality: 10%). Operations such as arithmetic, counting, and comparisons remain generally elusive skills (<25%ANLS).

Both LLMs we tested scored significantly lower than the human baseline in questions that require visual understanding, with an average ANLS score of 21%. This is understandable because these are text-only models.

While LLMs' zero-shot performance is relatively high, we note that **DUDE** consists of public-license documents from the web, which potentially might have

been included in the LLMs' pretraining corpus.

5.6.2 Assessing Confidence

ECE measures calibration of confidence, whereas AURC assesses both performance and confidence ranking [193] (more detail Section 2.2.3). The latter results in an appropriate metric to select the best model in real-world applications, where wrong predictions can yield undesired scenarios, which could be prevented by manually revising low-confidence answers.

Interestingly, T5-base-512 scores better on calibration (ECE=10.82) than T5-2D-large-8K, the baseline with the highest ANLS, yet worse calibration (ECE=14.4). In general, it seems calibration worsens when extending the maximum sequence length, whereas adding 2D position embeddings only positively affects ANLS. From the baselines tested, T5-2D-large-8K achieves the highest AURC.

Another interesting result comes from analyzing the calibration of models evaluated using the *Concat* strategy vs. *Max Conf.* strategy. In the main paper, we reported results for the model with the relative best ANLS. Thanks to our varied set of evaluation metrics, we discover that *Max Conf.* overall results in poor calibration (see Table 5.4), whereas considering ANLS, there is not always a clear winning strategy. This shows that predicting each page separately and necessarily assuming conditional independence across pages is not a reliable strategy for multipage DocVQA.

5.7 DUDE Competition Results

5.7.1 Submitted Methods

Overall, 6 methods from 3 different participants were submitted for the proposed tasks in the **DUDE** competition. To avoid cherry-picking from considering all submissions of individual participants, we consider only the last submission (accentuated) for the final ranking. All the methods followed an encoder-decoder architecture, which is a standard choice for VQA when abstractive questions are involved. Specifically, the submitted methods are mostly based on T5-base [383] as the decoder. For this reason, we include the *T5-base* baseline to compare how the participant methods improved on it. A short description of each method can be found in Table 5.5.

Two very recent state-of-the-art architectures, UDOP and HiVT5, have been extensively leveraged by participants. The former is geared toward improved

Model	ANLS	ECE	AURC
BertQA MPDocVQA Concat	29.8	**13.83**	**43.28**
BertQA MPDocVQA MaxConf	**32.18**	28.93	48.73
BigBird MPDocVQA Concat	**30.67**	**25.07**	**47.2**
BigBird MPDocVQA MaxConf	29.38	50.79	56.81
LayoutLMv3 MPDocVQA Concat	22.61	**13.19**	57.11
LayoutLMv3 MPDocVQA MaxConf	**25.27**	31.31	58.54
Longformer MPDocVQA Concat	**33.45**	**22.21**	45.83
Longformer MPDocVQA MaxConf	28.67	48.6	58.11
T5 MPDocVQA Concat	34.37	**18.97**	47.31
T5 MPDocVQA MaxConf	**37.56**	23.73	**46.69**
T5-base Concat-0	**25.62**	**20.05**	62.25
T5-base MaxConf-0	22.21	39.47	**58.89**

Table 5.4. Comparison of baselines using Concat or Max Conf strategies.

document page representations, while the latter targets multipage document representations. In their method reports, the UDOP-based models by Lenovo Research mention calculating confidence by multiplying the maximum softmax score of decoded output tokens with two additional post-processing rules: a) predicted not-answerable questions confidence is set to 1, b) when abstaining, confidence is set to 0.

5.7.2 Performance Analysis

Table 5.6 reports the competition results ranking comparing the submitted methods' performance on the test set. Higher ANLS and AUROC values indicate better performance, while lower ECE and AURC values signify improved calibration and confidence ranking. According to the findings, the `UDOP+BLIP2+GPT` approach attains the highest ANLS score (50.02), achieving the best calibration and OOD (out-of-distribution) detection performance. In a direct comparison of the `MMT5` and `HiVT5+modules` methods, the former shows a higher ANLS score, yet did not provide any confidence estimates.

Thus, the overall winner is `UDOP+BLIP2+GPT` by Lenovo Research. Their submitted methods (ranked by highest ANLS) also differentiate themselves by their additional attention to confidence estimation. Based on the numbers in

Method	Description
T5-base (ours)	T5-base [383] fine-tuned on **DUDE** (AWS OCR), with a delimiter combining list answers into a single string, and replacing not-answerable questions with 'none'.
LENOVO RESEARCH	
UDOP(M)	Ensemble (M=10) of UDOP [443] (794M each) models without self-supervised pretraining, only fine-tuned in two stages: 1) SP-DocVQA [450] and MP-DocVQA [451], and 2) **DUDE** (switching between Azure and AWS OCR).
UDOP +BLIP2	UDOP(M=1) with integrated BLIP2 [260] predictions to optimize the image encoder and additional page number features.
UDOP +BLIP2+ GPT	UDOP(M=1) and BLIP2 visual encoder with ChatGPT to generate Python-like modular programs to decompose questions for improved predictions [160, 437].
UPSTAGE AI	
MMT5	Multimodal T5 pretrained in two stages: single-page (ScienceQA [403], VQAonBD2023 [385], HotpotQA [508], SP-DocVQA) with objectives (masked language modeling (MLM) and next sentence prediction (NSP)), multipage (MP-DocVQA and **DUDE**) with three objectives (MLM, NSP, page order matching). Fine-tuning on **DUDE** with answers per page combined for final output.
INFRRD.AI	
HiVT5	Hi-VT5 [451] with 20 <PAGE> tokens pretrained with private document collection (*no information provided*) using span masking objective [204]. Fine-tuned with MP-DocVQA and **DUDE**.
HiVT5 +modules	Hi-VT5 extended with token/object embeddings for a variety of modular document understanding subtasks (detection: table structure, signatures, logo, stamp, checkbox; KIE: generic named entities; classification: font style).

Table 5.5. Short descriptions of the methods participating in the **DUDE** competition, in order of submission. The last submitted method is considered for the final ranking.

the table, several interesting observations can be made to support the suggested future directions and propose additional experiments:

- **ANLS.** The integration of `UDOP`, `BLIP2`, and `ChatGPT` contributes to the method's superior overall performance in answering different question types.

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

Table 5.6. Summary of Method performance on the **DUDE** test set. Average ANLS results per question/answer type are abbreviated as (Abs)tractive, (Ex)tractive, (N)ot-(A)nswerable, (Li)st. (*) All scalars are scaled between 0 and 100 for readability.

- **ECE**, **AURC.** Integrating `UDOP`, `BLIP2` visual encoder, and `ChatGPT` for question decomposition contributes to the method's performance in handling uncertainty across various question types.
- **Abstractive.** The top performance of `UDOP+BLIP2+GPT` in abstractive questions reveals the potential of combining the `UDOP` ensemble, `BLIP2` visual encoder, and `ChatGPT` to enable abstract reasoning and synthesis of information beyond simple extraction.
- **List.** The performance of `UDOP+BLIP2+GPT` in list-based questions suggests that incorporating page number features can enhance the model's capability to process and generate list information, which might be spread across pages.

Figure 5.8 visualizes an overview of the performance of each submitted method respective to diagnostic subset samples matching a certain diagnostic category. The models generally struggle with operations involving *counting*, *arithmetic*, *normalization*, and *comparisons*. As expected, models have higher performance when dealing with simpler questions (*complexity simple*) compared to more complex questions (*complexity multi-hop*, *complexity other hard*, and *complexity meta*). Models tend to perform better when handling evidence in the form of plain text (*evidence plain*) compared to other forms of evidence, such as visual charts, maps, or signatures. Performance across models is notably lower for tasks involving lists compared to other question types. Models show varying performance when dealing with different types of forms (*e.g.*, *date*, *numeric*, *other*, *proper*).

Figure 5.10 studies the ability of the competitors' methods to answer questions respective to increasingly longer documents. We observe a significant drop in ANLS when aggregating scores over gradually longer documents. This is expected as the longer the document is, the more probable that the answer will either be located on a later page or rely on a long-range dependency between the tokens (*e.g.*, a multi-hop question). Strikingly, all methods' scores, except

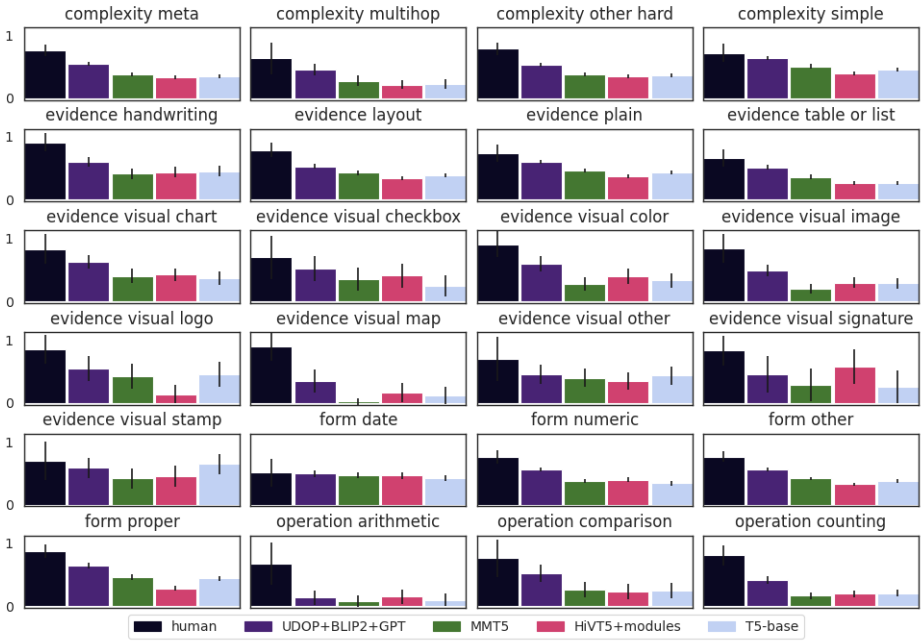


Figure 5.8. We report the average ANLS per diagnostic category for each of the submitted methods vs. **human** and a baseline method T5-base. Since the diagnostic dataset contains a different number of samples per diagnostic category, we added error bars representing 95% confidence intervals. This helps visually determine statistically significant differences.

`Hi-VT5+modules`, drop significantly for questions on 2-page documents. This is likely to have the root cause in the standard input size of T5-based methods equal to 512 tokens, covering roughly 1 page.

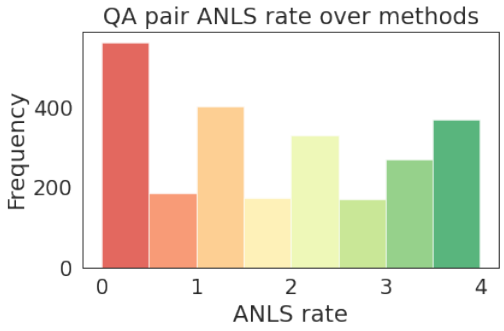


Figure 5.9. A histogram (bins=8, matching ANLS-threshold of 0.5) of the average ANLS rate per QA pair when summing ANLS scores over competitor methods.

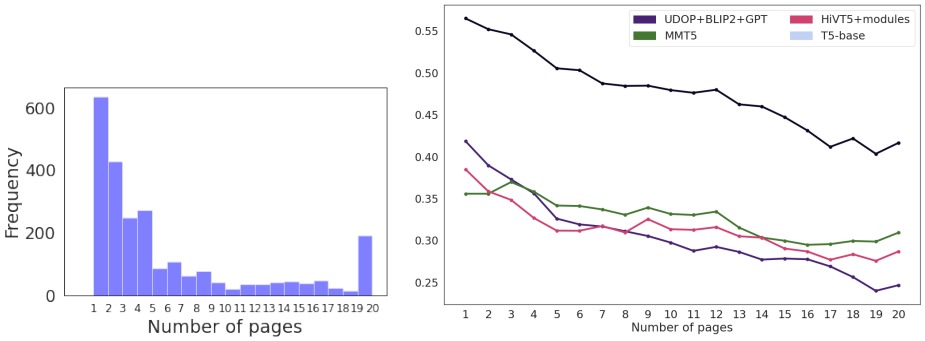


Figure 5.10. Left: A histogram over the number of questions relative to the number of pages in the document (limited to 20 pages). Right: A line plot of the average ANLS score per QA pair: – documents of length *at least* (x-axis) pages.

Figure 5.9 analyzes the correlation of errors over competitor methods. A large portion of QA pairs is predicted completely wrong (ANLS-rate = 0) by all competitor methods. This can have many plausible causes: a) by all sharing a similar decoder (T5), methods suffer from similar deficiencies, b) some QA pairs are too complex for current SOTA competitor methods, particularly questions requiring more complex reasoning or unique document-specific layout processing. To further analyze this phenomenon, we sample qualitative examples with different ANLS rates (Appendix B.1).

5.8 Chapter Conclusion

In conclusion, this chapter introduces a new large-scale multipaged, multi-domain, multi-industry Document Visual Question Answering Benchmark for document understanding. Our dataset is adjusted to the real-world environment where we need to process long documents and understand different types of documents. The benchmark includes visual semantics such as *tables, charts, figures, lists, checkboxes, stamps*, and more, which are essential for real-world document understanding. The performance of SOTA textual and multimodal models still lags behind human performance, indicating the need for further improvement in visual understanding for DU models. Nevertheless, we believe evaluating systems on **DUDE** could inspire new architectures and methods.

Limitations. As our approach is closer to real-world industrial applications, and enables models to recognize and understand new unseen data without the need for re-training, it does come with some limitations and constraining factors, including the use of only English language documents. Future work could address these limitations and expand the benchmark to include other languages. Moreover, although our dataset can be considered large-scale, it still represents a relatively small sample size of the plethora of documents that exist in the real world.

As a core contribution of **DUDE**, we wanted to emphasize the importance of evaluation beyond mere predictive performance. **DUDE** offers an interesting and varied test bed for the evaluation of novel calibration and selective QA approaches (*e.g.*, [96, 273]). While this was not explicitly attempted in this iteration of the competition, we hope that future work will consider testing their methods against **DUDE**.

Future of the Shared Task As the competition evolves, we hope that DUDE will serve as an essential platform for pushing the frontiers of research and driving innovation in the DU field. Currently, our competition focuses on English language documents, which means we miss out on the potential of incorporating *multilingual* data. An ideal extension for future iterations of the shared task would be to introduce multilingualism, which our framework can accommodate, provided that source documents are readily available. However, this would also require specifying language qualifications for annotation experts. Moreover, one could automate part of the data collection process and annotation process by allowing the best-performing competition system to validate the aptitude and complexity of human-proposed QA pairs.

Chapter 6

DistilDoc: Knowledge Distillation for Visually-Rich Document Applications

The contents of this chapter come from a publication under review at CVPR 2024 [471]:

Jordy Van Landeghem, Subhajit Maity, Ayan Banerjee, Matthew B Blaschko, Marie-Francine Moens, Josep Llados, and Sanket Biswas. DistilDoc: Knowledge Distillation for Visually-Rich Document Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (under review)*, 2024

This is an external collaboration with Subhajit Maity, Ayan Bannerjee, Josep Llados, and Sanket Biswas. The work was conceived during a research visit at the Computer Vision Center in Barcelona, Spain.
Disclosing the work done by the authors other than supervisors:

- **Jordy Van Landeghem** created the project's scope, implemented and performed all DIC and downstream DocVQA experiments, including training DLA teacher models, connecting the DLA inference and evaluation, and wrote the manuscript with supplementary.
- **Subhajit Maity** and **Ayan Bannerjee** built the DLA architectures and performed the DLA-KD experiments.
- **Sanket Biswas** brought the team together and helped with related work and the introduction.

This chapter focuses on efficiency via knowledge-distillation (KD) model compression for document understanding (DU) tasks. While DU research is dependent on increasingly sophisticated and cumbersome models, the field has neglected to study efficiency via *model compression*, referring to any technology transforming large and complex models into smaller streamlined models with similar performance [548]. Here, we design a KD experimentation methodology for more lean, performant models on DU tasks that are integral within larger task pipelines, specifically document image classification (DIC) and document layout analysis (DLA).

We carefully selected KD strategies (*response-based, feature-based*) for distilling knowledge to and from backbones with different architectures (*ResNet, ViT, DiT*) and capacities (*base-small-tiny*). We study what affects the teacher-student knowledge gap and find that some methods (tuned *vanilla KD*, *MSE*, *SimKD* with an apt projector) can consistently outperform supervised student training. Furthermore, we design a downstream task setup to evaluate the robustness of distilled DLA models on zero-shot layout-aware document visual question answering (DocVQA).

DLA-KD experiments result in a large mean average precision (mAP) knowledge gap, which unpredictably translates to downstream robustness, accentuating the need to further explore how to efficiently obtain more semantic document layout awareness.

This chapter motivates the need for more efficient DU models, especially for VRD tasks, and provides a benchmarking framework for future research on KD for DU tasks. Additionally, it motivates being smart about when to use which modality when the downstream task has a certain modality-bias (*e.g.*, DocVQA is a text-centric task, whereas DLA is more vision-centric). Finally, it links to efforts in DUDE to use LLMs for DU, with the focus here on incorporating layout information from distilled DLA models into the LLMs.

6.1 Introduction

Visually-rich Document Understanding (DU) has attracted increasing interest over the last few years. It involves multiple tasks such as document image classification (DIC) [165, 195, 210, 284], key information extraction (KIE) [197, 272, 296, 422, 433], document layout analysis (DLA) [35, 36, 80, 362, 544] and document visual question answering (VQA) [100, 309, 310, 450]. Current SOTA DU models [153, 187] solve the task by using modern OCR engines to read the text and then combine them with spatial features to predict the page layout and structure. However, these multimodal architectures come with the following

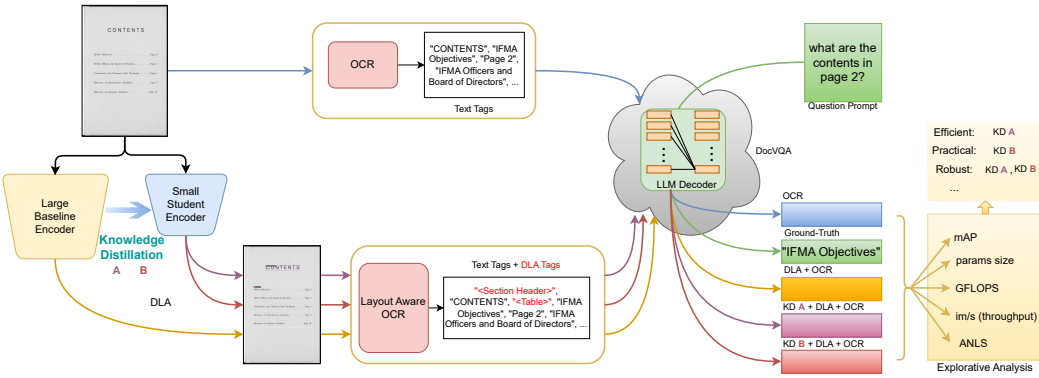


Figure 6.1. `DistilDoc` presents the first framework to investigate the potential of KD-based DLA model compression to enrich LLM prompts with **logical layout structure** to practically and efficiently improve downstream applications such as DocVQA.

drawbacks: 1) They rely primarily on Large Language Models (LLMs) [542] pretrained on millions of samples which depend more on OCR text quality than visual features/document structure; 2) can be computationally heavier due to the need to process and fuse information from different modalities; and 3) may perform poorly in domains with poor OCR results or on low-resource languages.

Therefore, this work focuses on single-modality, vision-only architectures that can be finetuned for handling VRDs in tasks involving understanding visual-layout semantics such as tables, titles, paragraphs, figures, *etc.* DLA is a useful preliminary step in a document processing workflow [35, 80], holding the key to enhancing practical downstream DU tasks such as DIC, KIE, and VQA. DLA can impart *logical layout* structure, beyond *geometric layout* from OCR [164], and structured context to the document, to enable more accurate content extraction and interpretation. A recent DU competition [469] has pleaded to bridge the gap between DLA and DocVQA by introducing layout-navigating or multi-region questions.

To handle the computational demand of modality/task-specific models, knowledge distillation (KD) [21, 150, 178, 394] can prove an effective approach to obtain efficient modules for later re-use in enriching LLM document inputs. Teacher model compression has the potential to make student models that improve over direct finetuning, also making them practical for deployment with resource-constrained devices or for faster real-time inference. The field of Document AI [79] is engaged with representing and understanding VRDs, but thus far has not explored KD-based model compression for improved efficiency

and uncertainty estimation [126].

This work investigates the potential of enriching VRDs with logical layout structure derived from effective DLA model compression using KD methods to practically and efficiently improve downstream DU applications. The nature of the (document) dataset has a major impact on the KD process [434], which required motivated choices (regarding dataset usage [14, 165, 362], architectures, weight initialization [259], KD methods [63, 67, 170, 178, 183, 540], evaluation, downstream procedure [482], *etc.*) in designing our experimental methodology of KD benchmarking for DU tasks (DIC, DLA). This allows us to investigate aspects affecting teacher-student knowledge/capacity/initialization gaps.

The key contributions of the paper are twofold:

I. We are the first to design, apply, and open-source an experimental methodology for comprehensively benchmarking KD-based model compression on DU tasks involving VRDs (DIC and DLA).
II. We design a novel evaluation procedure based on the downstream task of zero-shot layout-aware DocVQA to quantify the robustness of distilled DLA models.

Nevertheless, our contributions go beyond mere KD-based compression benchmarking, promoting **logical layout** analysis over geometric layout to enhance the generalization of DU models toward unseen documents with diverse and complex layouts, as demonstrated in Figure 6.1.

6.2 Related Work

Efficiency and Model Compression Efficiency through model compression is gaining relevance with the increasing parameter size and complexity of models such as LLMs [556]. Although KD is a prominent technique for model compression, several alternative approaches are worth mentioning. *Quantization* has been recently re-discovered in the context of LLMs with LoRA [184] and Q-LoRA [93] that achieves substantial model compression with minimal accuracy degradation. Advances have been made also in vision-and-language [57, 518] and more recently for vision transformer (ViT) training [269]. However, its effectiveness also depends on some key factors, including the model architecture, data type, bit-width, and the training recipes employed. In this direction, *neural architecture search* (NAS) became an important field of study [55, 279, 280, 363]. Popular alternatives include *model weight pruning* [131, 288, 554] that benefits

strongly from joint usage with other efficiency and model compression techniques; *adaptive inference* with multi-exit architectures [501, 547], which are promising yet highly dependent on early exit network design and uncertainty estimation. KD-based training [364] complements the aforementioned techniques, leading to potentially more accurate model exits and pruning. Moreover, KD strategies involve overall simpler design choices, depending mostly on the availability of a large teacher model trained on domain data of interest. Therefore, we prioritize KD-based model compression and efficiency for practical DU applications.

Knowledge Distillation KD strategies can be categorized into three main categories: *response-based* KD [6, 21, 178, 314, 509, 541] seeks to match the final layer predictions of the teacher model; *feature-based* KD [8, 62, 67, 175, 221, 394] aims to mimic features extracted from intermediate hidden layers of the deep network and *relation-based* KD [355, 356, 447, 511] which exploits the relations between different layers or sampled data points. However, the latter approach is more geared toward pixel-based semantic segmentation tasks. While feature-based KD is more versatile, it is more expensive and harder to implement than soft teacher predictions. While offline methods [178, 394] consider an existing frozen teacher model, online methods [61, 538] update both student and teacher networks jointly. Self-distillation [22, 528] represents a special case of online KD, which employs the same network as both the teacher and student, progressively outperforming the network's performance, albeit disregarding the aim of efficiency.

Our work's scope will be offline KD schemes, with a single converged teacher (vs. intermediate checkpoints [479] or ensembles [515]), single modality inputs (vision only), with three different feature extraction backbones (ResNets, ViT and a self-supervised pretrained document foundation model DiT [259]). Our study seeks to extend the empirical utility of KD to popular DU tasks (DIC & DLA) with a versatile benchmarking framework to ensure future compatibility, fostering KD-based DU model compression research.

Practical and Efficient Document Understanding Recent efforts to represent layout and document structure have gained substantial recognition, particularly with the incorporation of structural information into LLMs. The LayoutLM family [187, 502, 503] and GeoLayoutLM [296] laid the foundation of using 2D positional information of text (word blocks) tokens obtained from OCR as a *geometric layout* representation for the input. Recent work [416] has further enhanced this 2D representation by incorporating text lines or text blocks as layout groups inside the OCR text tokens. [482] further experiment with structure-preserving OCR, that uses appropriate spaces and line breaks as an

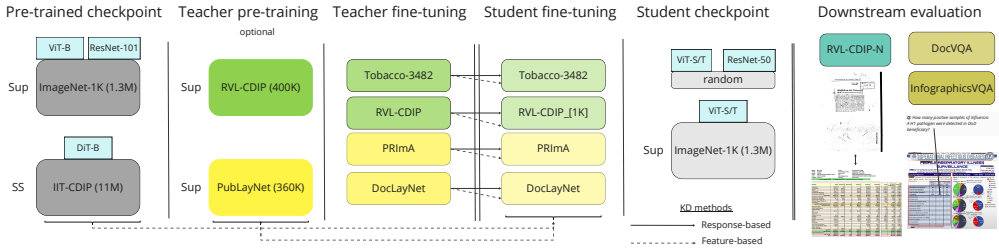


Figure 6.2. **Proposed experimental methodology** to comprehensively study all aspects (left-to-right) that impact *KD methods* (response, feature; projectors) adapted for *VDU task specifics* (architecture, weight initialization, pretraining & finetuning datasets, student capacity). Downstream setups evaluate the robustness of distilled students.

LLM input, thereby improving the ability to capture layout and structural cues for zero-shot DocVQA [309, 310] tasks. [153, 263] seek to represent layout as region-level proposal features, representing *logical layout* elements like title, paragraph, figure, tables, *etc.*) as in the DLA task. To further study the utility of logical layout representations, [498] address asking questions conditioned inside a specific region of a page, improving upon the design of DocVQA that provides too many in-line questions (>80%). More recently, PDFTriage [400] generates a structured metadata representation of born-digital documents, extracting both geometric and logical layout elements like section text, figure captions, headers, and tables for a more precise QA approach. DUDE [468] offers a testing bed for DocVQA on multipage, multi-type documents with varying layouts, including questions conditioned on layout navigation, *e.g.*, '*Which pages have tables?*'.

Our explorations focus on making the most of the logical layout features obtained from the multi-domain DLA benchmark, DocLayNet [362]. We build upon the aforementioned advancements and explore how incorporating document structure can enhance the performance of downstream task models, aligning with the trend of enriching LLMs with rich-text prompting and layout-aware representations.

6.3 Experimental Setup

This Section documents the experimental methodology established in this work as visualized in Figure 6.2, including datasets, architectures and backbones for teacher and student models, KD methods, and evaluation metrics for the

tasks and distillation effectiveness. The goal is to provide a framework for future research on KD for DU tasks and allow pinpoint comparisons on KD aspects such as teacher-student knowledge and capacity gap, teacher-pretraining, student network initialization, *etc.*

Table 6.1. Dataset usage for DIC, DLA, and downstream tasks. Symbols: P = pretraining, DP = document pretraining, T = teacher training, S = student training, * = subsampling, E = teacher/student evaluation, D: downstream evaluation

Dataset	**Task**	**Usage**	**Size**	**# Cls**
ImageNet [90]	DIC	P	1.28M	1000
IIT-CDIP [252]	DIC	DP,T,S	11M	/
Tobacco-3482[232]	DIC	T,S,E	3482	10
RVL-CDIP[165]	DIC	DP,T,E	400K	16
PRImA[14]	DLA	T,S,E	400	6
DocLayNet[362]	DLA	T,S,E	80.8K	11
RVL-CDIP-N [241]	DIC	D	1K	12
SP-DocVQA [450]	VQA	D	12.8K	50K
Infographic [310]	VQA	D	5.5K	30K

6.3.1 Datasets

Tab. 6.1 lists all datasets used (in)directly for the experiments. As there is no existing methodology for KD experimentation on the tasks involved, we motivate the design choices:

DIC We benchmark results on both *Tobacco-3482* (original train-val-test splits 800-200-2482) and *RVL-CDIP*. The originally large training size of *RVL-CDIP* hinders experimentation (long iteration cycles), which is why we create a subsampled student training set, $\textit{RVL-CDIP}_{1k}$, by randomly selecting 1K images per class. By evaluating the full *RVL-CDIP* test set, we provide a fair evaluation of the usefulness of KD methods, while avoiding the cumbersomeness of student finetuning on such a large dataset.

While *RVL-CDIP* is the de facto standard for measuring performance on the task of document classification, the literature [242, 470] has reported several undesirable characteristics such as (near-)duplicates causing substantial overlap between train and test distributions. We complement independently and identically distributed (i.i.d.) test set evaluation with benchmarking on RVL-CDIP-N [241], which is a covariate shift dataset allowing us to evaluate the robustness of KD methods to domain shift, which is a common problem in real-world applications.

DLA We benchmark results on *DocLayNet* (reporting evaluation on validation set following common practice) and *PRImA*. The former is a large-scale human-annotated dataset with 81K images and 11 categories of logical layout elements, while the latter is a smaller dataset with 400 images and 6 classes. *DocLayNet* contains a wide layout variability with six diverse document types (patents, scientific, legal, reports, tenders) in English. They have been hand-annotated by trained experts, making it the gold standard for DLA. Alternatively, Publaynet [544] or MS-COCO [274] benchmarks have been used in pretraining DLA models. However, the former lacks diversity as it only contains documents from the scientific domain while the latter is a more common object detection benchmark for natural scenes.

We consider a mirrored data setup for both tasks, with one larger benchmark dataset (*RVL-CDIP*, *DocLayNet*) and a smaller, easier dataset (*Tobacco-3482*, *PRImA*). This allows us to compare KD efficacy with more or less accurate teachers over tasks.

6.3.2 Architectures and Backbones

We evaluated three backbone architectures, representing different approaches to the tasks of DIC and DLA.

Backbones Residual Network (*ResNet*) [167]: A supervised pretrained CNN-based architecture that is a staple in image recognition.

Vision Transformer (*ViT*) [101]: A supervised pretrained Transformer-based architecture that is effective for a variety of CV tasks.

Document Image Transformer (*DiT*) [259]: A self-supervised pretrained architecture specifically designed for DU tasks, as it was pretrained on 11M document images from IIT-CDIP with a Masked Image Modeling objective, as inspired by BeiT [24].

Specific to DLA, we use the Mask R-CNN [168] meta-architecture for instance segmentation with two different backbones, i) classic ResNets and ii) ViT, with the latter more challenging to integrate [267].

Historically, CNNs have been more popular for DLA due to their accuracy, speed, and multiple optimizations built into the meta-architectures (involving a backbone, neck, and head). However, recent work is pointing to the potential of ViT as plain (non-hierarchical) object detectors [268]. Compared

to Transformers, CNNs have strong inductive biases of translation equivariance and locality, a fundamental difference that is less explored in a KD context [33].

Network Architecture and Initialization Document images are very different from natural images, yet most available vision backbones of different sizes are pretrained on the latter, except for DiT. Nevertheless, ViTs seem to struggle to learn a function when starting from random initialization, both as teachers and student networks. Therefore, we will use ImageNet pretrained checkpoints for all models considered, even for student network initialization.

Teacher Models While there are many model variants with different capacities for each of the backbones (Tab. D.1), we opt for the Base variant for Transformers, which arguably is most common. We consider ResNet-101 as it has the attractive property of having similar hidden layers' output dimensionality as the next smaller variant, ResNet-50.

The comparison of ViT-B and DiT-B allows us to evaluate the effects of different pretraining schemes (supervised, self-supervised) and how this affects knowledge transfer.

Student Models For DIC, we consider ViT-small and ViT-tiny, as well as a CNN-based architecture (ResNet-50), whereas, for DLA, we consider Mask-RCNN with a Resnet-50 backbone and a ViT-tiny backbone. Due to the computational demand of training instance segmentation models, we only consider the ViT-tiny backbone for the student model, therefore not making it possible to analyze KD methods for an increasing teacher-student capacity gap. While it would have made an interesting comparison, DiT has not been released in a smaller variant than DiT-B, and given the computational demand of pretraining DiT on the entire IIT-CDIP dataset containing 42 million document images, we did not consider it for student training. One might regard the knowledge transfer of DiT-B to a smaller ViT-(S/T) as potentially resulting in DiT-(S/T), yet the ImageNet or random initialization of the student network differs substantially from that of the self-supervised DiT weight space.

6.3.3 KD Methods

The basic approach of knowledge distillation consists of transferring 'knowledge' from a cumbersome teacher model f^t to a lightweight student model f^s, where $f : \mathcal{X} \to \Delta^{\mathcal{Y}}$ is a function mapping input data $\mathcal{X}$ and outputting a conditional probability distribution $P(y'|x)$ over output labels $y' \in \mathcal{Y} = [K]$ for K classes [368]. When this model compression approach is done effectively, the student model will be more efficient in terms of memory and computation. The top-1 class prediction is $\hat{y} = \text{argmax}_{y' \in \mathcal{Y}}[f(X)]'_y$, with $\hat{p} = \max_{y'}[f(X)]'_y$ the posterior probability. For convenience, $[\tilde{f}(x)]_k$ denotes the k-th element of the logits vector $\tilde{f}(x) \in \mathbb{R}^K$, which when normalized with softmax $f(x) = \sigma\left(\tilde{f}(x)\right) = \frac{\exp(\tilde{f}(x)/\tau)}{\sum_{k=1}^{K} \exp([\tilde{f}(x)]_k/\tau)}$. Let each function f be parameterized by θ holding all trainable parameters of the function, separable into a variable L layers, where $f_l(x)$ denotes the l-th layer output, *e.g.*, the penultimate layer output $f_{L-1}(x)$.

While there exists a wealth of ever-growing KD methods, we have carefully chosen a combination of simplistic methods mimicking the basic principles of KD (i, iv), more advanced KD methods that target specific improvements such as penalizing the non-target class logits (ii), or distilling the knowledge of intermediate layers (iv), and methods that take a step back on established KD practices by optimizing mean squared error (MSE) between teacher-student logits or reusing the teacher classifier (ii, vi).

Every method will be explained with loss functions, additional hyperparameters, and training parameters. (i) **Vanilla KD** [178] optimizes a linear combination of hard-target student cross-entropy (CE) loss and Kullback Leibler (KL) divergence loss with soft-target teacher predictions, including loss KD hyperparameters $\alpha \in [0, 1]$ and the temperature $\tau > 1$, which gives more weight to student loss and controls the softness of teacher logits, respectively.

$$\mathcal{L}_{\text{KD}} = \alpha \underbrace{\mathcal{L}_{\text{CE}}\left(y, \hat{y}^s\right)}_{\tau=1} + (1-\alpha) \underbrace{\tau^2 \mathcal{L}_{\text{KL}}\left(f^t(x), f^s(x)\right)}_{\tau>1}$$

(ii) **MSE** loss between teacher-student logit vectors enables direct logit-level matching [217]

$$\mathcal{L}_{\text{MSE}} = \left\| \tilde{f}^s\left(x\right) - \tilde{f}^t\left(x\right) \right\|_2^2$$

(iii) **NKD** Normalized KD loss [509] decouples vanilla KD into a normalized (indicated $\mathcal{N}$) combination of the target ($c \in \mathcal{Y}$) loss and the non-target loss in CE form, where $\gamma \in [0, 1]$ is a trade-off hyperparameter and τ the temperature.

$$\mathcal{L}_{\text{NKD}} = \underbrace{[f^t(x)]_c[\tilde{f}^s(x)]_c}_{\text{target}} - \gamma \cdot \tau^2 \cdot \underbrace{\sum_{k \neq c}^{K} \mathcal{N}\left([f^t(x)]_k^\tau\right)\left(\mathcal{N}\left([\tilde{f}^s(x)]_k^\tau\right)\right)}_{\text{non-target}}$$

(iv) **FitNet** [394] enables feature-based KD by minimizing the Euclidean distance between the intermediate feature maps of the teacher and student networks (i.e., MSE loss). A trainable projector $\mathcal{P}(\cdot)$ (*e.g.*, a linear projection layer) is required if the dimensionality of the hint layer(s) $h \in [1, L+1]$ outputs does not correspond to that of the student, There are no hyperparameters, except for projector design and where to place hint layers in the teacher network.

(v) **ReviewKD** [67] uses multi-stage information (multiple layers) of the teacher to supervise one student layer. The knowledge review mechanism is too complex to cover here as it involves multiple modules (residual learning, attention-based fusion projector, and a hierarchical context loss). This work claimed the first exploration of KD for instance segmentation, which is why we include it only for DLA.

(vi) **SimKD** [63] is a hybrid KD method that combines the advantages of response-based and feature-based KD. On the one hand, it reuses the pretrained, frozen teacher classifier for student inference ($f_L^t(\mathcal{P}(f_{L-1}^s(x)))$, and on the other hand, it adopts MSE for feature alignment (following a projector) of the penultimate layer feature-representations. Note that the former classification output is not used for training or loss calculation, only the latter projected feature map alignment.

$$\mathcal{L}_{\text{SimKD}} = \mathcal{L}_{\text{MSE}}\left(\mathcal{P}\left(f_{L-1}^s\left(x\right)\right), f_{L-1}^t\left(x\right)\right)$$

While the projector can safely be discarded for (iv,v) to obtain cost-free student inference, SimKD requires both the trained projector and teacher classifier to be used (and stored) for student inference. SimKD originally proposed a CNN-based projector between teacher and student feature maps (assuming C(hannels) x H(eight) x W(idth) inputs). For compatibility with ViT-based architectures, we contribute a novel variant of SimKD, which uses a linear projection layer on the [CLS] token at the penultimate layer. Alternatively, we draw upon [77, Theorem 1] that a multi-head self-attention layer can simulate a convolutional layer, subsequently reshaping the penultimate hidden layer output (ignoring [CLS] pooling) to (C x W x H), where C is the hidden size (*e.g.*, 197(-1) for ViT-B), and W, H are equal to the number of patches (*e.g.*, 14 for ViT-B with patch size 16 and image sizes 224x224), finally applying the original CNN projector to obtain the projected feature maps.

Task considerations The number of KD methods considered between the tasks differs, as some methods were not designed for use in a meta-architecture like Mask R-CNN. Response-based methods using logits are not capable of providing knowledge for object localization (*e.g.*, region proposal network head), making feature mimicking of vital importance. Moreover, the performance of instance segmentation highly depends on the quality of deep features to locate interested objects [509, 541], which is why we only consider feature-based KD methods for DLA (v, vi). When deciding upon KD methods to include, the literature reported ReviewKD as the feature-based SOTA, NKD as the response-based SOTA, and SimKD as the hybrid SOTA on image classification (CIFAR-100).

6.3.4 Evaluation

Metrics Predictive performance evaluation for DIC follows standard practice with accuracy, whereas we forego the F1 score as the classes are balanced. For DLA, we use the standard metrics of mean average precision (mAP) @ intersection over union (IOU) [0.50:0.95] of bounding boxes.

Efficiency evaluation considers the combination of parameter size and FLOPS (floating point operations) to be representative enough to compare distilled models.

Following calls in the DU literature [468] to establish calibration and confidence ranking as defaults to the evaluation methodology, we include Expected Calibration Error (ECE) [156, 332, 340] to evaluate top-1 prediction miscalibration and Area-Under-Risk-Coverage-Curve (AURC) [138, 193] to measure the error rate over selective (% of test set) accuracy (detailed in Section 2.2.3).

Covariate shift DIC-KD evaluation To evaluate the robustness of distilled models, we consider evaluating the impact of domain shift on the downstream task of DIC. Luckily, there exists a dataset similar to *RVL-CDIP* in terms of document types and classes, yet different in terms of document sources and label distribution. This dataset is called RVL-CDIP-N [241], and we will use it to evaluate the robustness of distilled models.

6.3.5 DLA-enriched LLM prompting

Downstream DLA-KD evaluation An important objective of this work is to demonstrate the usefulness of DLA predictions in downstream VRD tasks. As SOTA DLA models are often as cumbersome (parameter size, GFLOPS) as the downstream models, this motivates the need for KD to obtain more efficient DLA predictors that could be used to enrich document inputs with logical layout information.

While we focus on visual-only document inputs in benchmarking KD, we take the opportunity to benchmark DLA as part of a zero-shot DocVQA task setup with text-only LLMs [482], which can benefit from additional layout information when answering questions that appear in certain logical elements ('what is the first column header of Table 3', 'what is the title of the document?'). Similarly, it could benefit to know what falls within an infographic picture or legend; which is why we benchmark on SP-DocVQA and InfographicVQA, with the latter containing more visually-rich information. As a model of choice, we have opted for LLAMA-2-7B-CHAT [452] with 4-bit quantization to keep GPU memory requirements to a minimum, while still performing sufficiently reliably. Evaluation is done using ANLS [39, 468] on predicted answers vs. ground truths.

The prompt design follows [482] with a task instruction and placeholders for the question and the document input, the latter depending on the prompt parameterization (see Tab. 6.2). Possible values are *plain*, single-spaced OCR tokens, *space*, tokens placed heuristically with whitespaces in their approximate position, or *DLA*, which adds start and end tags such as `<Table>` and `</Title>` to indicate logical layout as predicted by a DLA model. A pseudo-algorithm (Sec. 6.3.5) details the procedure to generate DLA-enriched prompts.

KIE is regarded as an important downstream DU task, yet we believe (as supported by [166]) that it would benefit less from DLA, due to most information being organized as key-value pairs with only local context relevance.

6.4 Results & Discussion

DLA-KD This work investigates different SOTA KD methods and integrates them into the DLA framework with ResNet and ViT feature extraction backbones. KD in DLA poses significant challenges owing to the intricate

Algorithm 1: Construction of DLA-enriched prompts $\boldsymbol{p}_{\text{DLA}}$

Input: A finite set $\mathcal{D}_{test} = \{(\mathbf{x}_{(i)}, y_{(i)})\}_{i=1}^{N}$ of holdout data, consisting of document images $\mathbf{x}_{(i)}$ and corresponding labels $y_{(i)}$
Output: Tokenized DLA-enriched prompts $\boldsymbol{p}_{\text{DLA}}$
Parameters: ζ_{iou}: IoU-threshold for layout-token boxes (default: 0.3)
Parameters: Ignore-labels: DLA labels to ignore for enrichment (default: {'Text'})
Input : A document image $\boldsymbol{v}$

Require: A trained DLA model and an OCR engine
(1) **Feed image to DLA model to obtain labeled layout boxes**
$\{(b_j, c_j, m_j)\}_{j=1}^{J} \leftarrow$ `DLA(`$\boldsymbol{v}$`)` // Boxes, classes, metadata
Feed image to OCR engine to obtain tokens and boxes
$u = \{(w_t)\}_{t=1}^{T}, s = \left\{\left(x_t^1, y_t^1, x_t^2, y_t^2\right)\right\}_{t=1}^{T} \leftarrow$ `OCR(`$\boldsymbol{v}'$`)` // Tokens and token-boxes
Standardize layout boxes to similar xy-format
for $j \leftarrow 1$ ***to*** J **do**
  $b_j \leftarrow$ `StandardizeBbox` (b_j) // Standardize to xy-format
  if `OCR` *image dims* $\neq$ `DLA` *image dims* **then**
    // Precomputed OCR (DUE) results can be reused, yet OCR images can have higher resolution
    Interpolate layout boxes to token-boxes
    $b_j \leftarrow$ `InterpolateBbox` $(b_j, \boldsymbol{v}, \boldsymbol{v}')$
    // Interpolate layout box to OCR image size
(2) **Find closest start and end token-boxes**
Input : a set of DLA predictions DLA($\boldsymbol{v}$), a set of OCR tokens u, a set of OCR token-boxes s
Output : an updated set of OCR tokens $\hat{u}$, a set of OCR token-boxes $\hat{s}$
for $j \leftarrow 1$ ***to*** J **do**
  $S \leftarrow (0, \infty)$; $E \leftarrow (-1, \infty)$ // Initialize start and end with dummy index and distance values
  for $t \leftarrow 1$ ***to*** T **do**
    // Multiple relaxing heuristics to find closest token-box to layout-box
    if $c_j \in$ *Ignore-labels* **then**
      continue
    if not `FullyContains`(b_j, s_t) *or* `IntersectionOverUnion`$(b_j, s_t) > \zeta_{iou}$ **then**
      // Token-box fully contained within layout-box or IoU > threshold
      continue
    // Minimal Laplacian distance to cornerpoint
    $S \leftarrow \min(S, (t, \text{Laplacian}(b_j, s_t)))$ // Laplacian distance to top-left corner
    $E \leftarrow \min(E, (t, \text{Laplacian}(b_j, s_t)))$ // Laplacian distance to bottom-right corner
(3) **Insert DLA labels before and after closest tokens**
Input : The original sets of OCR tokens u, token-boxes s, and start and end indices S and E
Output : Updated sets of OCR tokens $\hat{u}$ and token-boxes $\hat{s}$
$C \leftarrow 0$ // Initialize token insertion counter
$\hat{u}, \hat{s} \leftarrow u, s$ // Initialize to be updated OCR tokens $\hat{u}$ and token-boxes $\hat{s}$
$I \leftarrow$`SortAndLabel(`S,E`)` // sort start and end token together by index and add label type
for $j \leftarrow 1$ ***to*** $|I|$ **do**
  if I_j *is a start token* **then**
    $\hat{u} \leftarrow$ insert <c_j> at $I_j + C$ // Insert label such as <Table> before token
    $\hat{s} \leftarrow$ insert b_j at $I_j + C$
    $C \leftarrow C + 1$
  if I_j *is an end token* **then**
    $\hat{u} \leftarrow$ insert </c_j> at $I_j + C + 1$ // Insert label such as </Table> at next token
    $\hat{s} \leftarrow$ insert b_j at $I_j + C + 1$
    $C \leftarrow C + 1$
return $\hat{u}, \hat{s}$ // Tokens and token-boxes with DLA labels to be used in prompt design of [482]

Table 6.2. Prompt design following [482], with placeholders depending on parameterization of document input (*plain, space, DLA*).

#l	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: {}

Table 6.3. Results for KD methods applied on DocLayNet [362].

Teacher	Student	Method	mAP↑	Flops↓	Params↓	Im/s↑
ViT-B	-	Supervised	65.65	107G	114M	20
R101	-	Supervised	73.56	60G	63M	12
-	ViT-T	Supervised	62.85	68G	**26M**	14
-	R50	Supervised	72.43	33G	44M	12
R101	R50	SimKD	*62.71*	**29G**	44M	21
		ReviewKD	61.17	37G	44M	19
ViT-B	ViT-T	SimKD	57.51	42G	**26M**	**22**
		ReviewKD	57.2	84G	**26M**	17

nature of detection, introducing new obstacles related to regression, region proposals, and sparser label volumes [64]. As motivated in Sec. 6.3.3, we prioritize feature-based KD methods, with results on DocLayNet in Tab. 6.3. The performance comparison in terms of mean average precision mAP and FLOP counts show that Resnet-50 students with SimKD are overall superior in

Table 6.4. Validation ANLS (scaled to %) of LLAMA-2-7B-CHAT [452] on SP-DocVQA [309] (top) and InfographicVQA [310] (bottom), where (if marked) the prompt is enriched with DLA predictions from a ViT-B-based Mask-RCNN.

space	task	DLA	$ANLS_{val}$	Image/Photo	Yes/No	Figure/diagram	Form	Free_text	Handwritten	Layout	Others	Table/list
✓	✓	✓	61.2	44.58	49.13	40.28	68.95	68.39	52.81	61.38	56.44	56.7
✗	✓	✓	58.39	44.43	41.67	34.81	66.38	67.82	52.1	59.19	55.91	52.79
✓	✓	✗	62.46	42.95	49.43	40.93	71.15	70.59	55.87	61.87	61.05	58.31
✗	✓	✗	57.63	45.38	51.52	34.97	67.88	69.71	53.19	55.51	55.78	53.81

space	task	DLA	$ANLS_{val}$	Arithmetic	Comparison	Counting	Figure	Map	Multi-span	Abs	Q span	Single span	Table/list	Text	Visual/layout
✓	✓	✓	28.05	9.92	25.28	7.83	26.28	19.0	21.85	8.82	41.84	33.54	25.57	34.6	29.17
✗	✓	✓	28.36	14.93	29.15	7.64	27.05	19.0	19.41	11.21	46.87	33.35	25.56	34.59	26.69
✓	✓	✗	27.97	9.78	25.13	6.99	25.93	21.04	22.33	8.2	43.36	33.53	25.76	35.06	27.47
✗	✓	✗	29.08	14.15	26.94	11.35	27.52	19.1	19.79	12.79	48.44	33.79	26.17	35.24	26.39

terms of both efficiency and detection, while ViT-Tiny student has the smallest number of parameters with comparable performance in terms of mAP.

)However, one can observe a generally large knowledge gap between the teacher and student model ($\approx 8\%$ for ViT and $\approx 10\%$ for the ResNets) as the crucial details about the document object boundaries, shapes, and sizes can get lost during the compression process. Not only that, KD performance with a ViT backbone is worse compared to Resnets due to (i) the attention overhead, *i.e.*, transferring this attention-based knowledge to a student model requires careful consideration of how to distill these complex attention patterns effectively, and (ii) initialization and hyperparameter sensitivity, *e.g.*, finding an appropriate domain pretrained checkpoint and setting patch sizes, attention heads, can affect the KD process, requiring more delicate tuning. The CNN layers of Resnets are, on another hand, permutation invariant and provide more flexibility towards KD.

KD methods are hard to integrate for object detection frameworks, especially when it comes to ViTs where there is no intermediate multi-scaled FPN module. Our contribution lies in extending the hybrid SimKD [63] method for the DLA task and also showing competitive analysis with the existing SOTA ReviewKD [67].

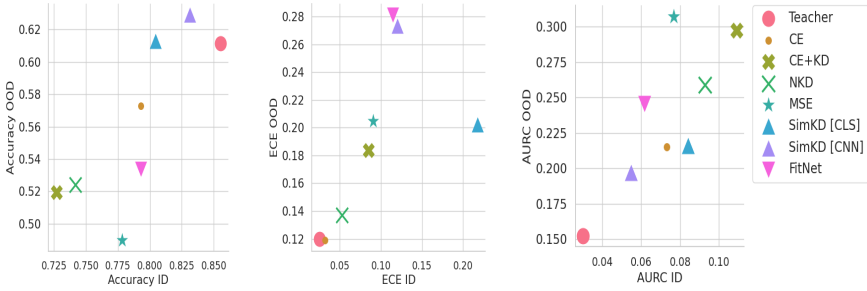
Downstream DLA-KD Tab. 6.4 reports results on the validation sets as these are hyper-annotated with evidence, question and answer types, and operations, allowing for more finegrained analysis. Detail results of distilled DLA-enriched prompts are available in Appendix D.4.

On SP-DocVQA, DLA-enriched prompting (without spacing) improves from 57.63 $\rightarrow$ 58.39, whereas (with spacing) the improvement (27.97 $\rightarrow$ 28.05) is less pronounced on InfographicVQA, yet DLA predictions are still useful in this setting, as also evidenced by questions involving 'Visual/Layout'. This is likely due to the more visual and layout complexity of the dataset, wherefore DLA predictions are less accurate. Strikingly, spacing performs generally worse on Infographics, pointing to the heuristic nature of the structure-preserving OCR algorithm of [482] that fails on structurally complex documents with visually-situated language, charts with axes labels, legends, *etc.*

The objective of these experiments was to make (distilled) DLA output useful in enriching text-only LLMs with more semantic layout information beyond geometric-spatial relations. For every setting tested, the task instruction (Sec. 6.3.5) is vital (else ANLS $< 5\%$) in the zero-shot setting. We hypothesize that for SP-DocVQA line/row/column-level key-value pair recognition suffices for attaining good performance, thus expecting little benefit from DLA-enriched

prompts. However, as these experiments are bound to the layout classes as pre-defined in DocLayNet, we believe that richer layout information, closer to semantic regions (*e.g.*, an address block instead of an OCR block), and including specification of common document objects such as stamps, logos, watermarks, should benefit downstream DU tasks.

Table 6.5. Performance per KD method over metrics averaged over architectures on RVL-CDIP dataset (In-Domain) and RVL-CDIP-N dataset (Out-Of-Distribution).



DIC-KD This task benchmark reports on experiments with 3 backbones, 2 student architectures (except 1 for Resnet), and 6 KD methods each. Tab. 6.6 details the ViT and DiT results, whereas the ResNet results (following similar trends) are available in Appendix D. The same set of experiments was repeated for randomly initialized students (Tabs. D.12 and D.13). Given the comprehensive scope of the DIC experiments, we can make claims regarding the overall most performant KD method, the teacher-student capacity gap, and the architecture-pretraining gap. ViT-Small student distilled with the SimKD [63] method performs best in terms of accuracy and AURC. Note that *the best ViT-Tiny student with only 5.5M parameters reaches 83% accuracy with SimKD, only 2.9% behind the best ViT-Small student with 86M parameters*, showing the potential of advanced KD methods in retaining accuracy at such a large capacity gap. SimKD performs admirably in terms of accuracy, sometimes (depending on the projector type (MLP and CNN)) as well as the supervised teacher. In terms of AURC, NKD and MSE approaches are best-performing, which are both response-based methods. Regarding the pretraining gap, as shown in Tab. 6.6, results indicate that a *self-supervised teacher like DiT does not meet expectations* when distilling the knowledge to a ViT-based student pretrained with ImageNet weights. This could be attributed to the large representation gap in the feature space between the RVL-CDIP pretrained and ImageNet pretrained models. However, evaluation under covariate shift on RVL-CDIP-N (Tab. D.8) demonstrates DiT-based students (distilled with response-based KD strategies)

to outperform ViT→ViT students, pointing to the *potential of self-supervision for robustness to distribution shift.*

Table 6.6. Results of different KD strategies benchmarked for D/ViT-B teachers applied on the *RVL-CDIP* dataset.

ViT-B					DiT-B				
Student	Method	ACC	AURC	ECE	Student	Method	ACC	AURC	ECE
–	ViT-B	0.891	0.017	0.034	–	DiT-B	0.933	0.075	0.010
–	ViT-S	0.853	0.030	0.058	–	ViT-S	0.831	0.042	0.056
–	ViT-T	0.822	0.040	0.043	–	ViT-T	0.801	0.053	0.047
ViT-S	Vanilla [$\tau = 2.5, \alpha = 0.5$]	0.854	**0.028**	**0.049**	**ViT-S**	Vanilla [$\tau = 2.5, \alpha = 0.5$]	0.831	0.060	0.080
	NKD [$\tau = 1, \gamma = 1.5$]	0.840	0.036	0.074		NKD [$\tau = 1, \gamma = 1.5$]	0.790	0.058	**0.040**
	MSE	0.855	**0.028**	0.051		MSE	0.831	0.060	0.082
	SimKD [CLS+MLP]	**0.859**	**0.028**	0.287		SimKD [CLS+MLP]	0.838	0.087	0.438
	SimKD [CNN]	0.847	0.062	0.141		SimKD [CNN]	**0.851**	**0.048**	0.136
	FitNet [middle]	0.843	0.048	0.141		FitNet [middle]	0.775	0.063	0.077
ViT-T	Vanilla [$\tau = 2.5, \alpha =$]	0.825	**0.038**	**0.058**	**ViT-T**	Vanilla [$\tau = 2.5, \alpha =$]	0.801	0.064	0.081
	NKD [$\tau = 1, \gamma = 1.5$]	0.815	0.046	0.094		NKD [$\tau = 1, \gamma = 1.5$]	0.772	0.066	**0.041**
	MSE	0.823	0.040	0.066		MSE	0.795	0.076	0.081
	SimKD [CLS+MLP]	**0.830**	0.095	0.163		SimKD [CLS+MLP]	0.816	0.104	0.439
	SimKD [CNN]	0.829	0.056	0.150		SimKD [CNN]	**0.832**	**0.056**	0.152
	FitNet [middle]	0.812	0.051	0.153		FitNet [middle]	0.753	0.077	0.054

Covariate shift DIC-KD To answer if certain KD methods harm a student model's robustness to covariate shift, we plot results per KD method, averaged over the 3 backbones on the (Tab. 6.5). This re-establishes the superiority of SimKD [CNN] in terms of accuracy, both ID and OOD, yet due to poor calibration, it loses gain on the teacher in terms of AURC. Strikingly, MSE attained the lowest OOD performance, whereas it was a solid ID choice. Tab. D.8 provides more detail on the performance of the different KD methods on RVL-CDIP-N, where we observe that grouped per KD strategy response-based is superior over all metrics.

6.5 Chapter Conclusion

KD-based model compression has been a popular technique in recent years, albeit DU research has not paid much attention to efficiency. Our work explores a limited scope of KD for DU at scale, revealing great potential for creating efficient counterparts of cumbersome DLA models used today. Specifically, we show that SimKD is a particularly strong KD method, always outperforming vanilla KD and even obtaining a 16x smaller model retaining >90% relative accuracy. Moreover, we investigate the potential of DLA for enriching document inputs in downstream DocVQA tasks. Traditionally, DocVQA has relied on plain OCR text. While structure-preserving OCR provides a notion of geometric

layout for downstream use, DLA was never considered before for the same purpose, yet our experiments show promise.

The more comprehensive benchmarking of KD methods in DIC with ID evaluation and a covariate shift protocol reveals interesting observations regarding the feature representation and weight initialization gap between DiT (documents) and ViT (natural images), albeit self-supervision for students is more robust in the OOD setting. Our framework enables informed selection of compressed models and directs several interesting explorations: how pretraining objectives impact the distillation process, if different layout representations (*e.g.*, [15, 187, 263, 443, 555]) allow for a more robust downstream transfer, *etc.*

Limitations While we primarily use DocLayNet, it remains the DLA dataset with the most diversity in layout elements both in terms of categories and shape or size. However, the downstream DocVQA results urge for more diversity in terms of document types, domains, and objects (*e.g.*, layout objects such as logos, watermarks, stamps, signatures). Thus, the community is in dire need of a dataset diverse enough to guarantee a performance improvement downstream. Moreover, multimodal KD was not considered in this work, holding promise for more efficient, all-round DU models. The downstream task was not tested on [468] as multipage documents are more complex to benchmark with limited sequence length LLMs. Also, DLA being a fairly complicated instance segmentation task, makes it difficult to adapt for KD-based model compression, ruling out some KD methods. This calls for a better experimental framework and architectural modeling to boost the exploration of KD in DLA, in turn, incubating downstream advances in processing and understanding VRDs.

Chapter 7

Conclusion

This final chapter summarizes the work done in this thesis. Additionally, we formulate the key contributions and propose some exciting avenues for future research.

7.1 Summary

To summarize, this thesis contains the following contributions (**C**) and key findings (→), respective to the research questions from the introduction:

When tested in realistic language data distributions on various text classification tasks, how well do PUQ methods fare in NLP?
In which settings are PUQ methods most useful, i.e.*, which failure sources/distribution shifts are the most sensitive?*

C 1. We conduct a benchmarking study of established PUQ methods applied to six real-world text classification datasets with a focus on model robustness and uncertainty quality. This large-scale study comes with advanced statistical analysis to validate significant differences between methods and datasets.

C 2. We propose a practical experimental methodology to test relevant distributions shifts —cross-domain classification and novelty detection—, resulting in a better understanding of the individual shortcomings of PUQ methods.

→ General behavior of PUQ methods does not hold over different datasets. We do observe specific correlations between PUQ methods and the problem setting representing task characteristics, for which we formulated practical takeaways. This reconfirms the need for modality to task-specific benchmarking of PUQ methods.
→ In general, PUQ methods are sensitive to distribution shifts and methods that exhibit better in-domain calibration also exhibit better robustness to novel class shifts. The tested setting of cross-domain classification under covariate shift is the most challenging for PUQ methods. This is evident from relatively low AUROC scores due to the presence of comparably similar linguistic patterns across domains.

How can we obtain better PUQ estimates without overrelying on computationally prohibitive methods, e.g.*, Deep Ensemble* [238]?

C 3. We propose novel combinations of PUQ methods, providing both well-motivated intuition and empirical evidence for the complementary benefits of combining different posterior approximation procedures.

→ Our proposed hybrid PUQ methods improve over singular methods, both in in-domain calibration, novelty detection, and out-of-domain detection. In particular, we show that the combination of *Deep Ensemble* with *Concrete Dropout* demonstrates higher diversity in posterior samples and superior performance, even at a smaller ensemble size compared to a Deep Ensemble.

How important are certain prior, neural architecture or hyperparameter influences on the quality of PUQ estimation?

C 4. We conduct a range of ablation experiments to investigate the influence of prior, neural architecture and hyperparameter choices on the quality of PUQ estimation. In particular, the number of stochastic posterior samples, the dropout rate, and the architecture are shown to have a significant impact on the quality of PUQ estimation.

→ The combination of posterior geometry and weight-based priors proves to be a powerful combination for PUQ estimation, with the *Deep Ensemble* and *Concrete Dropout* methods as the best-performing methods in our benchmark. Nevertheless, it is important to consider adapting the dropout rate to the text classification task at hand, which individually and in an ensemble impacts model robustness and uncertainty quality.
→ Contrary to previous work, we find that pretrained transformers in NLP severely underperform in novelty detection compared to 1D CNNs, limiting the applicability of transfer learning when distribution shift from novel classes can be expected.

How severe is the problem of hallucination and control in LLMs when evaluated in a selective, free-form DocVQA task setting?

C 5. We design the DUDE dataset with this task setting in mind, incorporating a large set of unanswerable questions that are realistic and relevant to the document's content.

→ Hallucination and control remain severe problems in LLMs, with a large fraction of unanswerable questions being answered with high confidence. When trained on a large set of unanswerable questions, LLMs improve on identifying unanswerable questions, yet at the expense of abstractive, harder questions to which they become overcautious (*e.g.*, ChatGPT predicting more than 1/2 of abstractive questions as unanswerable). With longer context, LLMs are also more likely to hallucinate answers. Overall, results are lagging behind the human baseline performance on DUDE, indicating that LLMs are still far from being able to reason about documents in their entirety without control measures.

How can we iteratively close the gap between research and practice in DU?

C 6. We take stock of the balance between research and applications in document classification, a prototypical DU task, and we identify the main challenges that are stalling progress in the field, with a focus on data construction and evaluation methodology.

C 7. We propose a novel formalization of multipage document classification scenarios, which we use to construct two novel datasets, RVL-CDIP_MP and RVL-CDIP-N_MP, that are more realistic and more challenging than their single-page counterparts.

C 8. We conduct an insightful experimental analysis of the novel datasets.

→ The experimental analysis reveals that current SOTA models are not able to leverage the additional context provided by multipage documents and that the performance gap between single-page and multipage document classification is still large. Ablation experiments show the promise of advancing multipage document representation learning and inference.

→ Major dataset construction efforts are required to bridge the currently existing gap and be able to rely on benchmarks for transfer to real-world applications. In particular, we identify the need for more realistic and more challenging datasets, about *e.g.*, the type and diversity of document data, and the variety and quality of label sets, as well as the need for more comprehensive evaluation methodologies.

How can we design a resource that comprehensively challenges the state-of-the-art? Which DU aspects are most challenging for current state-of-the-art LLMs? How can these be incorporated in a benchmark to allow proper measurements of future improvements?

C 9. We have designed a completely novel benchmark from the ground up, DUDE, collecting 40K QA pairs for 5K documents, constructing a multi-faceted (multipage ($\mu = 6$), multi-domain (± 15), multi-type (± 200), multi-QA (extractive, abstractive, list, unanswerable), multi-task (DIC, KIE, DLA, DOD, *etc.*), multi-OCR (Tesseract, Azure, AWS), multi-source, multi-stage (<5) annotations) dataset to foster research on *generic* DU, bypassing long context restrictions and evaluating the reliability and robustness of DU technology, as close as possible to real-world requirements.

C 10. The dataset construction approach of DUDE is based on a set of principles that we have formulated, which we believe are essential for a comprehensive benchmark for generic DU. More specifically, leveraging the *DocVQA* task paradigm and learning paradigm of *Multi-Domain Long-Tailed Recognition* allowed us to both incentivize harder questions on visual/layout semantics, layout navigation, or multi-step reasoning, while organically obtaining questions relevant to the document type and instance.

C 11. We have conducted our own baseline experiments of DUDE, evaluating the performance of SOTA DU models on the different facets of DUDE, as well as the reliability and robustness of LLMs in the context of DU. Next, we have organized a competition to challenge the community's best, additionally incorporating OOD detection and selective generation to evaluate CSFs on two common failure sources.

→ The best results attain ANLS <= 50% with our baseline T5-2D (8K context) scoring 46%, the competition winner improves 4% absolute by leveraging multimodal LLMs (BLIP2 and ChatGPT). Generally, stronger performance is expected from models that incorporate layout understanding and reasoning over multiple pages. Nevertheless, diagnostic results prove that the current SOTA still suffers on questions with visual evidence (only half of the human performance) or any reasoning operations (counting, comparison, *etc.*). With the rise of multimodal LLMs, better solutions are coming, yet due to its designed complexity, DUDE might remain "the benchmark to beat" for a long time.

→ Even while DUDE presents a great test bed for the challenge of long-context processing (Section 2.3.4.1), the evaluated models have not yet reached the point where they can fully leverage the additional context. This is a clear indication that more research is needed in the direction of efficient processing of long, structured documents.

→ We find that the quality of confidence estimation worsens with longer context, potentially from having to consider more possible answers. We also find that models using a maximum confidence strategy over answers generated per page results in substantially worse calibration. These interactions between multiple DU challenges prove the usefulness of incorporating and evaluating these jointly in a benchmark.

How can we efficiently infuse LLMs with semantic layout awareness for more focused information extraction?
To what degree can model compression resolve the problem of efficiency in processing documents?

C 12. We propose a novel experimental methodology to investigate enrichment of VRDs with semantic layout structure derived from effective distillation of DLA models to practically and efficiently improve downstream DU applications. This includes evaluation under covariate shift of KD methods in DIC and a downstream evaluation setup to evaluate the robustness of distilled DLA models on zero-shot layout-aware DocVQA.

C 13. We present the first application of KD to visual document tasks (DIC, DLA), investigating the teacher-student knowledge gap in KD-based model compression methods (response and feature-based) with task architectures involving different inductive biases (CNN vs. ViT), pretraining (self-supervised), student initialization, and capacities (base-small-tiny).

→ While we have promoted the use of semantic layout over geometric layout for enriching LLM prompts, this only results in limited improvements in performance, which we attribute to either the zero-shot evaluation setup or the limited subset of layout classes and domain shift from the DLA training data (DocLayNet). In some cases, *e.g.*, questions involving visual/layout evidence, DLA-enriched prompting proves more useful.

→ KD-based model compression is very effective in reducing model size, while maintaining accuracy at large capacity gaps, *e.g.*, a strong student is SimKD ViT-tiny, which retains relatively 93% of teacher accuracy, while being 16x smaller. Ablations show how the teacher-student knowledge gap is affected by the inductive biases of the task architecture, the pretraining of the student, the student initialization, and the student capacity. For example, a self-supervised teacher provides more robust students when evaluated under covariate shift. Nevertheless, model compression is but one tool in a larger toolbox for efficient processing of documents, which we believe is a key challenge going hand-to-hand with efficient longer-context modeling, for future research.

As this thesis was conducted in an applied research environment and keeping in mind that nowadays DL research is primarily empirical, the contributions of our work have been very focused on datasets and the experimental methodology, rather than on novel algorithms, which more often than not present mere incremental improvements on the state-of-the-art. Nevertheless, we believe that the proposed datasets and experimental methodologies are of great value to the community, as they provide a more realistic and more challenging test bed for future DU research. We are happy to see the proposed datasets and experimental methodologies increasingly being adopted by the community and hopefully this will foster research on more efficient and closer to real-world document processing, which will ultimately lead to more reliable and robust DU technology.

7.2 Perspectives For Future Research

This Section discusses some exciting research opportunities left for future research. First, we present a curated set of research questions particular to PUQ, calibration, and failure prediction, which when relevant are linked to DU applications. Next, we take a futuristic look at the design of a fully-fledged IA-DU solution, dreaming up the ultimate dataset and system design for DU.

7.2.1 Open Problems In Reliability & Robustness

Recent advancements in LLMs have brought a lot of groundbreaking improvements to the field of DU, yet the reliability of LLMs is still far from being solved. This is further increased by API-based services or closed-source LLMs [344], which are to be treated as black-boxes without access to model internals or token-level output logits, making it hard to apply most PUQ methods. Popular white-box approaches include verbalized probabilities [273] or semantic entropy [226] for taking into account semantic equivalence or specificity (*e.g.*, Where was the 2023 International Conference on Computer Vision held? → *In Paris* vs. *In the capital of France* vs. *In Europe*). Specific to selective generation, when knowledge on a topic is limited, it can be hard to censor LLM outputs (even when finetuning further with human feedback) or evaluate abstention reliably (*e.g.*, *I don't know* vs. *I don't care* vs. ‘’).

[111] implement a framework bundling a battery of white-box and black-box methods for LLM confidence estimation in text generation, yet it still requires human inspection of generated text together with the confidence score, which is not very scalable for large-scale document processing. This ties into the evaluation crisis of LLMs, which is a topic of active research [137]. In the short term, it might suffice to reward models that predict the full distribution of human judgments or learn human preferences for generated text. However, how can we expect models “to do what humans do” when even humans disagree or are not consistent in their judgments? Alternative approaches can be to rationalize judgments, attribute or ground evidence used for the judgment, or ask for clarifications when needed. In the long term, we should move beyond human evaluation, which is expensive, time-consuming, and not scalable. Important explorations include prompt chaining (*Please give a confidence between 0 and 1 about how certain you are this is the correct answer*) or self-evaluation [207, 391] to induce reflections on the quality of LLM outputs.

Beyond the potentially infinite, though countable output spaces of generative tasks, there exists an opportunity to study calibration for specific output spaces, *e.g.*, sequence-structured in the context of sequence tagging or restricted sequence-to-sequence tasks. Moreover, calibration metrics and methods can be adapted to the specific task or output space such as structured prediction [227], named entity recognition [222], object detection and segmentation [85, 234, 350] *etc.* With most works (if at all) reporting top-1 miscalibration, efficient estimation of “stronger" calibration notions is a crucial area of study to inform the derivation of calibrated regularized loss functions [370]. On the more theoretical side, it remains vital to investigate the link between non-convex optimization (*e.g.*, flat minima) and calibration, as well as when optimizing a proper loss yields calibration [42, 549].

Selective prediction has been garnering increased attention thanks to intensively comprehensive benchmarks [127, 193], yet these have (again) been focused on vision problems and architectures, inviting the same level of benchmarking on alternative modalities and tasks. To the extent of our knowledge there exists no work on extending selective prediction methodology to multi-task settings (*e.g.*, consider the typical combination of document classification and KIE) requiring a more complex learned CSF (for different output spaces) or a combination of multiple CSFs with multiple thresholding. Similar to calibration, differentiable loss functions for failure prediction are an open problem. More theoretical questions include the relationship between stronger notions of calibration and confidence ranking, as well as the link between feature space disentanglement and CSF ranking [552]. In the low-data regime, sample-efficient failure prediction is an open problem, which could leverage connections to semi-supervised and active learning [112].

7.2.2 A Future-Proof Design Of IA-DU

Downstream datasets are a key component of any practical, supervised ML solution, yet they are often overlooked in expectation of decent zero-shot performance with LLMs, which are trained on large-scale, generic language datasets, such as Common Crawl or the Pile [130]. While these datasets are very useful for pretraining general language understanding, they are not sufficient for all possible downstream tasks. This is especially true for DU, where text is but one of the modalities to be considered. As part of the conclusion to this thesis, we first discuss how to obtain the ultimate dataset for generic DU, and next we detail the design of a fully-fledged IA-DU solution.

7.2.2.1 The 'Ultimate' DU Dataset?

Arguably, a core contribution of this thesis is the design of the DUDE dataset, which we believe is a step in the right direction toward the ultimate dataset for generic DU. Top-of-mind extensions of DUDE include: multilingual or cross-lingual documents and questions; answer and evidence grounding to improve evaluation and interpretability; and question decomposition and simplification.

Finding a complete answer to the question of the ultimate DU dataset would be transformative to DU technology, yet here we can only provide some pointers, discussed in the structure of goal, starting points, and aspects to target.

Goal DU requires reasoning over documents in their entirety, which is a very complex task with the aforementioned challenges. With the current technology, this involves learning document representations that are both rich and compact, and that can be used to answer any question about the document. Consider how challenging this is when most relevant questions are either about the intentionality of the document's author or the way a user interacts with it, hinting at a potential observer's paradox in future data collection. For example, on a car invoice, an accountant would ask *What is the total amount due?*, or *Is this a valid invoice with correct taxation?*, while a customer would ask *How much do I finally have to pay?*, or the insurance broker *What is the chassis identifier to link the omnium coverage to?*. A model should be able to capture all these nuances about the complexity of a document which could be seen as the expectation of all possible relevant questions that can be asked on it, while also being able to generalize to unseen documents and questions. Therefore, the goal of the ultimate DU dataset is to provide a test bed for evaluating the progress in **commonsense reasoning on documents from real-world interactions**, to which we hypothesize that the scale and depth of supervision are vital.

Starting points The ultimate DU dataset should be designed with the aforementioned goal in mind, yet some seminal ML datasets could be inspiring. While the 'ImageNet moment' is etched in everyone's memory, MS COCO [274] was arguably a more impactful dataset thanks to its large-scale, diverse, and high-quality nature combining multiple tasks (image captioning, object detection, semantic segmentation, *etc.*). To build the equivalent of MS COCO in document understanding, DUDE offers a good starting point, under some conditions and necessary extensions. An important aspect concerns *ground truth collection* for DocVQA and the *complexity and specificity of questions and answers*, which has been approached differently by recent works: DUDE uses a multi-stage approach to collect a large set of minimally constrained, human-generated questions under the MDLT paradigm, which were afterward annotated with diagnostic categories; PDFTriage [400] pre-defines question types and collects a small set of human-generated questions; DocEdit [311] establishes a pre-defined taxonomy and tests language as a universal UI to interact with the hierarchical, discrete structure of documents. The extent to which the collected QA pairs constitute a representative sample of the space of all possible and relevant questions that can be asked on a document instance is an open problem, which can be approached by (**A**) *extending and scaling up existing practices* or (**B**) *deepening supervision for models to generalize better from limited inputs.*

A. Scale We identify three targets to scale up: (I) document collection, (II) question collection and validation, and (III) question-answer generation.

(I) Throughout the document dataset construction, the goal is to collect a large set of diverse document types and instances, differing on all modalities: language, layout, visual, *etc.*, and additional meta-criteria: industry, language, type, *etc.* The document collection approach taken in DUDE was a fairly artisanal process: based on experience, we designed an industry-document taxonomy, which we used to collect a large set of document types and instances, also taking into account the presence of different visual semantics or document objects *e.g.*, handwriting, stamps, watermarks, address blocks, *etc.* We leveraged a semi-automatically created keyword-style search (*'Please list 30 common retail document types with their synonyms like Credit memos - {"credit notes", "credit slips", "refund slips"}'*) on public document collections, and validated diversity post-hoc in terms of modality-specific features (TF-IDF or ResNet features) vs. other datasets.

A more scalable approach would be to leverage a cluster-based diverse sampling from larger document collections, such as Common Crawl [460]. While this approach would be more scalable, it would be challenging to ensure that the collected documents are diverse in terms of all modalities, which is a topic to be investigated. Relevant caveats are the presence of duplicates, sensitive information, and the need to balance language priors to not create Clever Hans effects for models to later exploit [405]. An active topic of research is document generation [169] or augmentation [304], which could fill the gap in document diversity, yet it would be challenging to ensure that the generated documents are both realistic and diverse. Seeing that business documents are hard to obtain, one could backtrack to visually-situated language.

(II) To ensure that questions are specific to a document, and not testing language understanding, cross-lingual questions could help counter reliance on language priors. However, both multilingual documents and cross-lingual questions are challenging to collect, as they require annotators capable of reading multiple languages. How people interact (*i.e.*, the questions asked) with documents without being systematically observed is what makes for interesting data, yet it is also the more challenging to collect. This is certainly true for subject-matter experts from different industries (government, finance, legal, *etc.*) who are not readily available for annotating documents. Naturally, as more documents are being collected, one should define a strategy to scale up the number of questions per document in a balanced way. Ideally, the number of questions per document should be a function of the document complexity, which is another open problem. Some basic strategies would be to (i) split questions evenly over pages by chunked annotation, yet this would constrain multi-hop and naturally complex questions, or (ii) to exploit the Gestalt principle [294], which states

that the number of questions should be higher on heterogeneous elements in a document. Finally, an untapped approach would be to generate questions automatically, which is an open research challenge.

(III) QA generation holds promise to grow a large-scale dataset. A possible approach would be to teach the current SOTA model on DUDE to generate questions (given possible answers, predict questions) similar to those in the training set. A harder problem is the generation of unanswerable questions, which we found hard to even elicit from humans. Potential caveats are the quality and factuality [303] of the generated questions. This might be improved upon by first generating rich and compositional captions for a document relative to the content and visual appearance, and then generating different questions based on the descriptions, with both paraphrasing and backtranslation for question variations and augmentations.

B. Supervision Depth The reasoning behind increasing the depth of supervision is that we might be expecting too much, *i.e.*, answering complex questions involving multiple manipulations of document-instance and/or domain-specific concepts based on a single set of reference answers, with a poor stimulus [476], *i.e.*, not providing enough, complex enough and diverse enough examples for models to generalize well.

Accounting for every possible question will be impossible. A possible approach inspired by MDLT and diagnostic categories in DUDE is to (i) decompose questions in terms of the skills and concepts (Definitions 15 and 16) required to answer it and pass this together as instructions; and (ii) hyperannotate more explicit answers, with answer and evidence grounding for attribution, better explaining the relations between primitives (skill-concept compositions). Figure 7.1 illustrates an example of (ii), where the answer is decomposed into a skill-concept composition, and the evidence is grounded to the relevant document objects. Such rich supervision should help models to both *discriminate* known skills and concepts and *generalize* better to new skill-concept compositions. Although it would be expensive to obtain such supervision in large quantities, the use of human-in-the-loop or active learning could reduce the annotation burden.

Definition 15 [*concept*]. An abstract term to denote *document visual objects* (atomic [cell, barcode] and molecular [table, chart, form]), and *entities* (generic [document identifier, person, date] and domain-specific [invoice number, insured, payment date]).

Definition 16 [*skill*]. Any manipulation [existence, counting, relation,

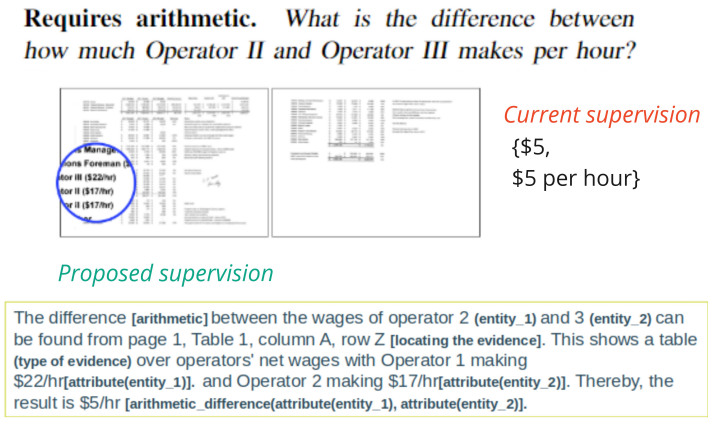


Figure 7.1. Example of ground truth formatting for a question-answer pair in DUDE.

hasattribute, *etc.*] of a concept, or a combination of involved concepts (evidence) involved.

Our overall idea is similar to how [243] alludes to **intelligence**: "the ability to decompose a problem into a set of skills and concepts, to reuse those skills and concepts in new situations, or acquire new ones quickly". The proposed format would be a full-featured instruction tuning dataset, which has proven very useful in other settings [404, 486] and which could be a valuable resource for future research on instruction-based learning of already existing and future DU tasks.

Naturally, all of this relies on the assumption that each question-answer pair can be decomposed into skill-concept compositions, and that there exists an exhaustive taxonomy of skills and concepts for DU, which thus far has not been created. A possible approach would be to leverage existing resources such as VerbNet [410] to define skills, or build an API for DocVQA similar to [437, 535] to decompose questions into programs with subroutines, *e.g.*, *How many of the contract's pages have signatures? Counting([Navigation(document), Existence(signature, page)])*; and construct a complete taxonomy of document concepts in both a bottom-up (human prior) and top-down (data-driven) fashion to extend it over time with domain-specific concepts. Ideally, this taxonomy should not be static at inference time, hinting at more research needed into neuro-symbolic learning for dynamic knowledge graphs to assist in recognizing and adding new concepts [32].

There are several fundamental questions that can be asked here "Is it needed to

collect thousands of QA pair examples to learn a specific document skill-concept composition, *e.g.*, address block detection?" Recent works seem to suggest not, indicating an emergent ability of the current best LLMs to find zero-shot solutions to a broad range of analogy problems [486]. Finally, building ground truth more amenable to advanced prompting and instruction-based learning [248] will likely prove as useful as question decomposition has in semantic parsing [189, 358, 525].

7.2.2.2 A Feature-complete IA-DU Solution?

The main takeaway of this thesis is that while more compute, more data and more powerful algorithmic tools have allowed significant progress in DU, there is still a long way to go toward the objective of reliable, robust, realistic, and efficient DU. For now, a major component would be a general-purpose Transformer-based stack for interfacing with a document through natural language. Most likely, this would be a multimodal LLM pretrained with a variety of pretraining objectives on the richest and largest possible corpus of documents and related data. When zero-shot performance is not sufficient, it would be instruction finetuned on new QA pairs, *e.g.*, in the rich format proposed in Section 7.2.2.1, resulting in efficient adapters that can be served concurrently on the same prediction model [417]. However, this is not a complete solution as generative modeling brings additional challenges (*e.g.*, expensive pretraining, decoding-based inference, confidence estimation, dependence on human evaluation, scalability).

Instead, we will focus here on another component of a complete solution, namely a **failure forecaster**, which we believe to be equally important for bringing LLMs closer to real-world applications. We envision this to be a lightweight module separate from the prediction model, that could be easily fully retrained and updated with new data, bypassing the risk of catastrophic forgetting and the need for retraining the more cumbersome LLM. The failure forecaster should predict the performance of the LLM on a given input (document, question, metadata *etc.*) and output (answer). It can be a very simple (*e.g.*, logistic regression) or complex model (*e.g.*, a large DNN), yet most of its complexity resides in the *feature modeling* and subsequent learning of *sources of uncertainty*. Our failure forecaster design is informed by [114]. We non-exhaustively identify sources of failure or uncertainty that can be modeled by the failure forecaster: (i) input uncertainty, (ii) output uncertainty, and (iii) distributional metrics. We discuss each of these in turn.

(i) Before answering any question, the document instance should be analyzed for inherent uncertainty or quality issues: *e.g.*, is it born-digital or OCR, the quality of OCR, readability metrics to capture how easy the document text

is to read, the complexity of the layout graph, visual richness. Next, follows the question analysis: *e.g.*, specificity, complexity, ambiguity, relevance, and novelty. Each of these can be measured by heuristic approximations such as the number of tokens or entities, how many of the entities literally appear in the document, the number of possible answers, the context size required to answer the question, the semantic overlap between the question and the document, how similar is the question to training data questions, the grammatical correctness, and syntactic complexity. Finally, the metadata analysis: *e.g.*, the number of documents in the same domain, the number of documents in the same type, the number of documents in the same language.

(ii) The output uncertainty can be modeled by the confidence of the LLM in its predicted answer, which can be estimated by PUQ methods and a variety of CSFs [111], which are hypothesized to capture complementary sources of uncertainty. Specific to the answer, the same question-document aspects return here, with additionally how extractive the answer is, the answer structure, and paraphrasing diversity.

(iii) Feature representations of new documents, questions, and answers can be assessed relative to their individual and joint distance to the training distribution [477]. This will be quintessential for distributional shift detection.

A failure forecaster trained to predict the performance of the LLM on all this information can be used to decide whether to abstain from answering, ask for clarifications from the model or human, ask for additional context, demand question rephrasing or a more clear document input, or even additional metadata. Ultimately, this will be useful to improve reliability and robustness for real-world IA-DU applications, where the risk of failure demands substantial control.

Bibliography

[1] IEEE Guide for Terms and Concepts in Intelligent Process Automation. Technical Report IEEE Std 2755-2017, IEEE, 2017.

[2] Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, 2020.

[3] Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Technical report, European Commission, Office for Official Publications of the European Communities Luxembourg, 2021.

[4] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, 2021.

[5] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L Hamilton, and Jimmy Lin. Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification With DocBERT. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 72–77, 2020.

[6] Somak Aditya, Rudra Saha, Yezhou Yang, and Chitta Baral. Spatial Knowledge Distillation to Aid Visual Reasoning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 227–235, 2019.

[7] Rajas Agashe, Srinivasan Iyer, and Luke Zettlemoyer. JuICe: A Large Scale Distantly Supervised Dataset for Open Domain Context-Based Code Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5436–5446, Hong Kong, China, 2019. Association for Computational Linguistics.

[8] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational Information Distillation for Knowledge Transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

[9] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models From Multi-Head Checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[10] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards Interpretable Math Word Problem Solving With Operation-Based Formalisms, 2019.

[11] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *CoRR*, abs/1606.06565, 2016.

[12] Anonymous. Calibration Regularized Training of Deep Neural Networks Using Kernel Density Estimation. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. under review.

[13] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering, 2015.

[14] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A Realistic Dataset for Performance Evaluation of Document Layout Analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300. Ieee, 2009.

[15] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-End Transformer for Document Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021.

[16] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. DocFormerv2: Local Features for Document Understanding. *arXiv preprint arXiv:2306.01733*, 2023.

[17] Chidanand Apté, Fred Damerau, and Sholom M. Weiss. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 1994.

[18] Martin Arjovsky. *Out of Distribution Generalization in Machine Learning*. PhD thesis, New York University, 2020.

[19] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-Domain Uncertainty Estimation and Ensembling in Deep Learning. In *International Conference on Learning Representations*, page 11, 2019.

[20] SpaCy authors. SpaCy `en_core_web_lg` Label Scheme, 2021.

[21] Jimmy Ba and Rich Caruana. Do Deep Nets Really Need to Be Deep? *Advances in neural information processing systems*, 2014.

[22] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label Refinery: Improving Imagenet Classification Through Label Progression. *arXiv preprint arXiv:1805.02641*, 2018.

[23] Souhail Bakkali, Zuheng Ming, Mickael Coustaty, Marçal Rusiñol, and Oriol Ramos Terrades. VLCDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification. *Pattern Recognition*, 139:109419, 2023.

[24] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*, 2022.

[25] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[26] Thomas Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[27] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[28] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[29] Emily M. Bender and Alexander Koller. Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, 2020. Association for Computational Linguistics.

[30] Oliver Bensch, Mirela Popa, and Constantin Spille. Key Information Extraction From Documents: Evaluation and Generator. In *European Semantic Web Conference (ESWC 2021) and 2nd International Workshop, in conjunction with ESWC 2021: Workshop: Deep Learning meets Ontologies and Natural Language Processing*, 2021.

[31] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are We Done With Imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[32] Sarthak Bhagat, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Sample-Efficient Learning of Novel Visual Concepts. *arXiv preprint arXiv:2306.09482*, 2023.

[33] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding Robustness of Transformers for Image Classification. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 10231–10241, 2021.

[34] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative Learning Under Covariate Shift. *Journal of Machine Learning Research*, 10(9), 2009.

[35] Galal M Binmakhashen and Sabri A Mahmoud. Document Layout Analysis: A Comprehensive Survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36, 2019.

[36] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. Beyond Document Object Detection: Instance-Level Segmentation of Complex Layouts. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):269–281, 2021.

[37] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. Docsynth: A Layout Guided Approach for Controllable Document Image Synthesis. In *International Conference on Document Analysis and Recognition*, pages 555–568. Springer, 2021.

[38] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 Competition on Scene Text Visual Question Answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. Ieee, 2019.

[39] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene Text Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on computer vision*, 2019.

[40] Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-Idl: Ocr Annotations for Industry Document Library Dataset. *arXiv preprint arXiv:2202.12985*, 2022.

[41] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online, 2020. Association for Computational Linguistics.

[42] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When Does Optimizing a Proper Loss Yield Calibration? *arXiv preprint arXiv:2305.18764*, 2023.

[43] 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, 2023.

[44] Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. Gram: Global reasoning for multi-page vqa. *arXiv preprint arXiv:2401.03411*, 2024.

[45] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

[46] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *arXiv preprint arXiv:1505.05424*, 2015.

[47] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. DUE: End-to-End Document Understanding Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[48] Pascal Bornet, Ian Barkin, and Jochen Wirtz. *Intelligent Automation: Welcome to the World of Hyperautomation: Learn How to Harness Artificial Intelligence to Boost Business & Make Our World More Human.* World Scientific, 2021.

[49] Benjamin Brazowski and Elad Schneidman. Collective Learning by Ensembles of Altruistic Diversifying Neural Networks. *arXiv:2006.11671 [cs, stat]*, 2020. arXiv: 2006.11671.

[50] Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[51] 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.

[52] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models Are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[53] Michael Buckland. Document Theory. *KO Knowledge Organization*, 45(5):425–436, 2018.

[54] 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.

[55] Han Cai, Tianyao Chen, Weinan Zhang, Yong Yu, and Jun Wang. Efficient Architecture Search by Network Transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[56] João Caldeira and Brian Nord. Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. *Machine Learning: Science and Technology*, 2(1):015002, 2020.

[57] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Deep Visual-Semantic Quantization for Efficient Image Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1328–1337, 2017.

[58] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. LifeQA: A Real-Life Dataset for Video Question Answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France, 2020. European Language Resources Association.

[59] Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. In-the-Wild Video Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.

[60] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A Dataset for Question Answering on Choropleth Maps, 2022.

[61] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online Knowledge Distillation With Diverse Peers. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3430–3437, 2020.

[62] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-Layer Distillation With Semantic Calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[63] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge Distillation With the Reused Teacher Classifier. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2022.

[64] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning Efficient Object Detection Models With Knowledge Distillation. *Advances in neural information processing systems*, 30, 2017.

[65] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online, 2021. Association for Computational Linguistics.

[66] Jingye Chen, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. XDoc: Unified Pre-Training for Cross-Format Document Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1006–1016, 2022.

[67] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling Knowledge via Knowledge Review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[68] Wenhu Chen, Yu Su, Yilin Shen, Zhiyu Chen, Xifeng Yan, and William Yang Wang. How Large a Vocabulary Does Text Classification Need? A Variational Approach to Vocabulary Selection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3487–3497, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[69] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-Learning via Language Model in-Context Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, 2022.

[70] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal Image-Text Representation Learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[71] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15138–15147, 2023.

[72] Arnaud Chevallier. *Strategic Thinking in Complex Problem Solving*. Oxford University Press, 2016.

[73] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, 2019.

[74] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. TutorialVQA: Question Answering Dataset for Tutorial Videos. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5450–5455, Marseille, France, 2020. European Language Resources Association.

[75] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated Input-Dependent Label Noise in Large-Scale Image Classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1551–1560, 2021.

[76] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) From Scratch. *Journal of machine learning research*, 12(Article):2493–2537, 2011.

[77] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the Relationship Between Self-Attention and Convolutional Layers. *arXiv preprint arXiv:1911.03584*, 2019.

[78] Anthony Corso, David Karamadian, Romeo Valentin, Mary Cooper, and Mykel J Kochenderfer. A Holistic Assessment of the Reliability of Machine Learning Systems. *arXiv preprint arXiv:2307.10586*, 2023.

[79] Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document Ai: Benchmarks, Models and Applications. *arXiv preprint arXiv:2111.08609*, 2021.

[80] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision Grid Transformer for Document Layout Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19462–19472, 2023.

[81] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving Selective Visual Question Answering by Learning From Your Peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24049–24059, 2023.

[82] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and Memory-Efficient Exact Attention With Io-Awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[83] Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. Quoref: A Reading Comprehension Dataset With Questions Requiring Coreferential Reasoning. *arXiv preprint arXiv:1908.05803*, 2019.

[84] Hal Daumé III. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[85] 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.

[86] A Philip Dawid. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[87] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[88] 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.

[89] Janez Demsar. *Statistical Comparisons of Classifiers Over Multiple Data Sets*. 2006.

[90] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[91] John Denker, Daniel Schwartz, Ben Wittner, Sara Solla, Richard Howard, Lawrence Jackel, and John Hopfield. Large Automatic Learning, Rule Extraction, and Generalization. *Complex Systems*, 1(5):877–922, 1987.

[92] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-Sensitive Learning. In *International Conference on Machine Learning*, pages 1184–1193. Pmlr, 2018.

[93] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient Finetuning of Quantized Llms. *arXiv preprint arXiv:2305.14314*, 2023.

[94] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, pages 4171–4186, 2018.

[95] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Naacl-hlt (1)*, 2019.

[96] Shehzaad Dhuliawala, Leonard Adolphs, Rajarshi Das, and Mrinmaya Sachan. Calibration of Machine Reading Systems at Scale. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1682–1693, Dublin, Ireland, 2022. Association for Computational Linguistics.

[97] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 193–202, 2014.

[98] DL Dimmick, MD Garris, and CL Wilson. Nist Special Database 6. Structured Forms Database 2. Technical report, Technical report, National Institute od Standards and Technology. Advanced~. . . , 1992.

[99] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling Transformers to 1,000,000,000 Tokens. *arXiv preprint arXiv:2307.02486*, 2023.

[100] Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. V-Doc: Visual Questions Answers With Documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21492–21498, 2022.

[101] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020.

[102] Thibault Douzon, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. Long-Range Transformer Architectures for Document Understanding. *ArXiv*, abs/2309.05503, 2023.

[103] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, 2020.

[104] John Duchi and Hongseok Namkoong. Learning Models With Uniform Performance via Distributionally Robust Optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[105] Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

[106] Ritam Dutt, Kasturi Bhattacharjee, Rashmi Gangadharaiah, Dan Roth, and Carolyn Rose. PerKGQA: Question Answering Over Personalized Knowledge Graphs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 253–268, Seattle, United States, 2022. Association for Computational Linguistics.

[107] David Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding Pathologies in Very Deep Networks. In *Artificial Intelligence and Statistics*, pages 202–210. Pmlr, 2014.

[108] A. Emmott, S. Das, Thomas G. Dietterich, A. Fern, and W. Wong. A Meta-Analysis of the Anomaly Detection Problem. *arXiv: Artificial Intelligence*, 2015.

[109] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[110] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

[111] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. LM-Polygraph: Uncertainty Estimation for Language Models. *arXiv preprint arXiv:2311.07383*, 2023.

[112] Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H. Abdi. Towards Better Selective Classification. In *International Conference on Learning Representations*, 2023.

[113] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim G J Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking Bayesian Deep Learning With Diabetic Retinopathy Diagnosis. *arXiv preprint arXiv:1912.10481*, page 11, 2019.

[114] Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated Selective Classification. *arXiv preprint arXiv:2208.12084*, 2022.

[115] Peter A. Flach. *Classifier Calibration*, pages 1–8. Springer US, Boston, MA, 2016.

[116] Andrew YK Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. In-Between Uncertainty in Bayesian Neural Networks. *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2019.

[117] Stanislav Fort and Stanislaw Jastrzebski. Large Scale Structure of Neural Network Loss Landscapes. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[118] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[119] Vincent Fortuin. Priors in Bayesian Deep Learning: A Review. *arXiv preprint arXiv:2105.06868*, 2021.

[120] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. A practical survey on faster and lighter transformers. *ACM Computing Surveys*, 55(14s):1–40, 2023.

[121] Sumam Francis, Jordy Van Landeghem, and Marie-Francine Moens. Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents. *Information*, 10(8):248, 2019.

[122] Giulio Franzese, Rosa Candela, Dimitrios Milios, Maurizio Filippone, and Pietro Michiardi. Isotropic SGD: A Practical Approach to Bayesian Posterior Sampling. *arXiv preprint arXiv:2006.05087*, 2020.

[123] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[124] 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.

[125] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete Dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017.

[126] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What Can We Learn From the Selective Prediction and Uncertainty Estimation Performance of 523 Imagenet Classifiers. *arXiv preprint arXiv:2302.11874*, 2023.

[127] 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.

[128] Ignazio Gallo, Lucia Noce, Alessandro Zamberletti, and Alessandro Calefati. Deep Neural Networks for Page Stream Segmentation and Classification. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. Ieee, 2016.

[129] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research*, 17(1): 2096–2030, 2016.

[130] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800gb Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[131] Shangqian Gao, Feihu Huang, Weidong Cai, and Heng Huang. Network Pruning via Performance Maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9270–9280, 2021.

[132] Siddharth Garimella. Identification of Receipts in a Multi-Receipt Image Using Spectral Clustering. *International Journal of Computer Applications*, 155(2), 2016.

[133] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *arXiv:1802.10026 [cs, stat]*, pages 8789–8798, 2018. arXiv: 1802.10026.

[134] Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: Layout-Aware Language Modeling for Information Extraction. In *International Conference on Document Analysis and Recognition*, pages 532–547. Springer, 2021.

[135] Inc Gartner. Gartner Peer Community, Hyperautomation: Are You Automating Your Decision Making? Survey. https://www.gartner.com/peercommunity/oneminuteinsights/hyperautomation-automating-processes-q2t, 2023. [Online; accessed Dec. 30, 2023].

[136] 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.

[137] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

[138] Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. *Advances in neural information processing systems*, 30, 2017.

[139] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. In *International Conference on Learning Representations*, 2018.

[140] Zoubin Ghahramani. A History of Bayesian Neural Networks. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

[141] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Taylor & Francis, 1995.

[142] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the Rules of the Game for Adversarial Example Research. *arXiv preprint arXiv:1807.06732*, 2018.

[143] Ingo Glaser, Tom Schamberger, and Florian Matthes. Anonymization of German Legal Court Rulings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 205–209, 2021.

[144] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[145] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence Calibration for Domain Generalization Under Covariate Shift. *arXiv preprint arXiv:2104.00742*, 2021.

[146] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

[147] Albert Gordo and Florent Perronnin. A Bag-of-Pages Approach to Unordered Multi-Page Document Classification. In *2010 20th International Conference on Pattern Recognition*, pages 1920–1923. Ieee, 2010.

[148] Albert Gordo, Marçal Rusinol, Dimosthenis Karatzas, and Andrew D Bagdanov. Document Classification and Page Stream Segmentation for Digital Mailroom Applications. In *2013 12th International Conference on Document Analysis and Recognition*, pages 621–625. Ieee, 2013.

[149] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Using Mode Connectivity for Loss Landscape Analysis. *ICML Workshop on Modern Trends in Nonconvex Optimization for Machine Learning*, 2018.

[150] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[151] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of Neural Networks by Enforcing Lipschitz Continuity. *Machine Learning*, 110(2):393–416, 2018.

[152] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling With Selective State Spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[153] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified Pretraining Framework for Document Understanding. *Advances in Neural Information Processing Systems*, 34: 39–50, 2021.

[154] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. XYLayoutLM: Towards Layout-Aware Multimodal Networks for Visually-Rich Document Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4583–4592, 2022.

[155] Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. A Question Answering Approach for Emotion Cause Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[156] 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*, page 1321–1330, 2017.

[157] Chirag Gupta and Aaditya K Ramdas. Top-Label Calibration and Multiclass-to-Binary Reductions. *arXiv preprint arXiv:2107.08353*, 2021.

[158] Deepak Gupta and Dina Demner-Fushman. Overview of the MedVidQA 2022 Shared Task on Medical Video Question-Answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, Dublin, Ireland, 2022. Association for Computational Linguistics.

[159] 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.

[160] Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional Visual Reasoning Without Training. *arXiv preprint arXiv:2211.11559*, 2022.

[161] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering Visual Questions From Blind People, 2018.

[162] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics,*, 2018.

[163] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[164] Haralick. Document Image Understanding: Geometric and Logical Layout. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 385–390. Ieee, 1994.

[165] 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.

[166] Jiabang HE, Yi HU, Lei WANG, Xing XU, Ning LIU, and Hui LIU. Do-Good: Towards Distribution Shift Evaluation for Pre-Trained Visual Document Understanding Models.(2023). In *Sigir*, pages 23–27.

[167] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[168] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-Cnn. In *Proceedings of the IEEE International Conference on computer vision*, pages 2961–2969, 2017.

[169] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-Based Document Layout Generation. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part I*, page 361–378, Berlin, Heidelberg, 2023. Springer-Verlag.

[170] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling Virtual Examples for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021.

[171] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine Learning With a Reject Option: A Survey. *arXiv preprint arXiv:2107.11277*, 2021.

[172] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint arXiv:1610.02136*, 2016.

[173] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[174] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained Transformers Improve Out-of-Distribution Robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, 2020. Association for Computational Linguistics.

[175] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2019.

[176] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[177] 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.

[178] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

[179] Geoffrey E Hinton and Drew Van Camp. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.

[180] Matt Hoffman. Langevin Dynamics as Nonparametric Variational Inference. *2nd Symposium on Advances in Approximate Bayesian Inference*, 2019.

[181] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A Pre-Trained Language Model Focusing on Text and Layout for Better Key Information Extraction From Documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10767–10775, 2022.

[182] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 Task 10: Math Question Answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.

[183] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models With Less Training Data and Smaller Model Sizes. *arXiv preprint arXiv:2305.02301*, 2023.

[184] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021.

[185] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States, 2022. Association for Computational Linguistics.

[186] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot Ensembles: Train 1, Get M for Free. In *International Conference on Learning Representations*, 2017.

[187] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-Training for Document AI With Unified Text and Image Masking. *ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[188] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 Competition on Scanned Receipt Ocr and Information Extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. Ieee, 2019.

[189] Drew A Hudson and Christopher D Manning. Gqa: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[190] Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: A Tutorial Introduction. *arXiv preprint arXiv:1910.09457*, 2019.

[191] Alekseĭ Grigorevich Ivakhnenko and Valentin Grigorevich Lapa. Cybernetic Predicting Devices. 1966.

[192] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *arXiv preprint arXiv:2104.14421*, 2021.

[193] 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.

[194] 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. NIH Public Access, 2020.

[195] Rajiv Jain and Curtis Wigington. Multimodal Document Image Classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–77. Ieee, 2019.

[196] Siddhartha Jain, Ge Liu, Jonas Mueller, and David Gifford. Maximizing Overall Diversity for Improved Uncertainty Estimates in Deep Ensembles. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 4264–4271, 2020.

[197] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A Dataset for Form Understanding in Noisy Scanned Documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6. Ieee, 2019.

[198] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[199] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[200] 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.

[201] Antonio Jimeno Yepes, Peter Zhong, and Douglas Burdick. ICDAR 2021 Competition on Scientific Literature Parsing. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 605–617. Springer, 2021.

[202] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics.

[203] Mahesh Joshi, Mark Dredze, William Cohen, and Carolyn Rose. Multi-Domain Learning: When Do Domains Matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312, 2012.

[204] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[205] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on Bayesian Neural Networks–a Tutorial for Deep Learning Users. *arXiv preprint arXiv:2007.06823*, 2020.

[206] Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, Jens Lehmann, and Maria Maleshkova. Conversational Question Answering Over Knowledge Graphs With Transformer and Graph Attention Networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 850–862, Online, 2021. Association for Computational Linguistics.

[207] 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.

[208] 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.

[209] Sanjay Kamath, Brigitte Grau, and Yue Ma. Verification of the Expected Answer Type for Biomedical Question Answering. In *First International Workshop on Hybrid Question Answering with Structured and Unstructured Knowledge (HQA'18)*, pages 1093–1097, Lyon, France, 2018. ACM Press.

[210] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. Convolutional Neural Networks for Document Image Classification. In *2014 22nd International Conference on pattern recognition*, pages 3168–3172. Ieee, 2014.

[211] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Rebecca Roelofs. Soft Calibration Objectives for Neural Networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[212] Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards Understanding 2d Documents. *arXiv preprint arXiv:1809.08799*, 2018.

[213] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, pages 5574–5584, 2017.

[214] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian Segnet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[215] Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian Learning Rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

[216] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-Free Document Understanding Transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.

[217] Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation. *arXiv preprint arXiv:2105.08919*, 2021.

[218] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar, 2014. Association for Computational Linguistics.

[219] Andreas Kirsch. Player of Jeopardy: ChatGPT Evaluation, 2023.

[220] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*, pages 5637–5664. Pmlr, 2021.

[221] Nikos Komodakis and Sergey Zagoruyko. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *Iclr*, 2017.

[222] 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, 2020. Association for Computational Linguistics.

[223] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification With Deep Convolutional Neural Networks. *Advances in neural information processing systems*, 25, 2012.

[224] Anders Krogh and John A. Hertz. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems 4*, pages 950–957. Morgan-Kaufmann, 1992.

[225] Anders Krogh and Jesper Vedelsby. Neural Network Ensembles, Cross Validation, and Active Learning. In *Advances in Neural Information Processing Systems*. MIT Press, 1995.

[226] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *arXiv preprint arXiv:2302.09664*, 2023.

[227] Volodymyr Kuleshov and Percy S Liang. Calibrated Structured Prediction. *Advances in Neural Information Processing Systems*, 28:3474–3482, 2015.

[228] 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.

[229] 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.

[230] 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.

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

[232] Jayant Kumar and David Doermann. Unsupervised Classification of Structurally Similar Document Images. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1225–1229. Ieee, 2013.

[233] Jayant Kumar, Peng Ye, and David Doermann. Structural Similarity for Document Image Classification and Retrieval. *Pattern Recognition Letters*, 43:119–126, 2014.

[234] 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.

[235] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 2019.

[236] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty Quantification Using Bayesian Neural Networks in Classification. *Computational Statistics and Data Analysis*, 142:13, 2018.

[237] Egor Lakomkin, Sven Magg, Cornelius Weber, and Stefan Wermter. KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition From YouTube Videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 90–95, Brussels, Belgium, 2018. Association for Computational Linguistics.

[238] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles, 2016.

[239] Ken Lang. Newsweeder: Learning to Filter Netnews. Version 20news-18828. *Machine Learning Proceedings 1995*, pages 331–339, 1995.

[240] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China, 2019. Association for Computational Linguistics.

[241] Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. Evaluating Out-of-Distribution Performance on Document Image Classifiers. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[242] 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*, pages 2665–2678, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.

[243] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 62(1), 2022. version 0.9. 2, 2022-06-27.

[244] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[245] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. FormNet: Structural Encoding Beyond Sequential Modeling in Form Document Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3735–3754, 2022.

[246] Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolay Glushnev, Renshen Wang, Joshua Ainslie, Shangbang Long, Siyang Qin, Yasuhisa Fujii, Nan Hua, and Tomas Pfister. FormNetV2: Multimodal graph contrastive learning for form document information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9011–9026, Toronto, Canada, 2023. Association for Computational Linguistics.

[247] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. *arXiv preprint arXiv:2210.03347*, pages 18893–18912, 2022.

[248] Bin Lei, Chunhua Liao, Caiwen Ding, et al. Boosting Logical Reasoning in Large Language Models Through a New Framework: The Graph of Thought. *arXiv preprint arXiv:2308.08614*, 2023.

[249] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics.

[250] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging Uncertainty Information From Deep Neural Networks for Disease Detection. *Scientific reports*, 7(1):1–14, 2017.

[251] Vladimir I Levenshtein et al. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union, 1966.

[252] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a Test Collection for Complex Document Information Processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666, 2006.

[253] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[254] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6309–6318, 2021.

[255] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. *Advances in neural information processing systems*, 31, 2018.

[256] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. MultiSpanQA: A Dataset for Multi-Span Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, 2022.

[257] Jing Li, Shangping Zhong, and Kaizhi Chen. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[258] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*, pages 12888–12900. Pmlr, 2022.

[259] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-Supervised Pre-Training for Document Image Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3530–3539, 2022.

[260] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023.

[261] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A Benchmark Dataset for Document Layout Analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[262] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13094–13102, 2023.

[263] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-Supervised Document Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021.

[264] Qiwei Li, Zuchao Li, Xiantao Cai, Bo Du, and Hai Zhao. Enhancing Visually-Rich Document Understanding via Layout Structure Modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4513–4523, 2023.

[265] Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi R Fung, Charles Yu, Joel R Tetreault, Eduard H Hovy, et al. Defining a New NLP Playground. *arXiv preprint arXiv:2310.20633*, 2023.

[266] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920, 2021.

[267] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking Detection Transfer Learning With Vision Transformers. *arXiv preprint arXiv:2111.11429*, 2021.

[268] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring Plain Vision Transformer Backbones for Object Detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[269] Zhikai Li and Qingyi Gu. I-Vit: Integer-Only Quantization for Efficient Vision Transformer Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17065–17075, 2023.

[270] 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.

[271] 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.

[272] Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. DocTr: Document Transformer for Structured Information Extraction in Documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19584–19594, 2023.

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

[274] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft Coco: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[275] Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. ViBERTgrid: A Jointly Trained Multi-Modal 2D Document Representation for Key Information Extraction From Documents. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 548–563. Springer, 2021.

[276] Zachary C Lipton and Jacob Steinhardt. Troubling Trends in Machine Learning Scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

[277] 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.

[278] Chenxiao Liu and Xiaojun Wan. CodeQA: A Question Answering Dataset for Source Code Comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2618–2632, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[279] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[280] Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical Representations for Efficient Architecture Search. *arXiv preprint arXiv:1711.00436*, 2017.

[281] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. XQA: A Cross-Lingual Open-Domain Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy, 2019. Association for Computational Linguistics.

[282] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A Challenge Dataset for Machine Reading Comprehension With Logical Reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 2020. Main track.

[283] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and Principled Uncertainty Estimation With Deterministic Deep Learning via Distance Awareness. *Neural Information Processing Systems 2020*, 2020.

[284] Li Liu, Zhiyu Wang, Taorong Qiu, Qiu Chen, Yue Lu, and Ching Y Suen. Document Image Classification: Progress Over Two Decades. *Neurocomputing*, 453:223–240, 2021.

[285] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-Based Out-of-Distribution Detection. *Advances in Neural Information Processing Systems*, 33: 21464–21475, 2020.

[286] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph Convolution for Multimodal Information Extraction From Visually Rich Documents. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 32–39, 2019.

[287] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2020.

[288] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the Value of Network Pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[289] Josep Lladós, Daniel Lopresti, and Seiichi Uchida. *Document Analysis and Recognition–Icdar 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings*. Springer Nature, 2021.

[290] L Rodney Long, Sameer K Antani, and George R Thoma. Image Informatics at a National Research Center. *Computerized Medical Imaging and Graphics*, 29(2-3): 171–193, 2005.

[291] Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering, 2020.

[292] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent With Warm Restarts. *International Conference on Learning Representations*, 2017.

[293] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[294] Tan Lu and Ann Dooms. Probabilistic Homogeneity for Document Image Segmentation. *Pattern Recognition*, 109:107591, 2021.

[295] Zhiyun Lu, Eugene Ie, and Fei Sha. Uncertainty Estimation With Infinitesimal Jackknife, Its Distribution and Mean-Field Approximation. *CoRR*, abs/2006.07584, 2020.

[296] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. GeoLayoutLM: Geometric Pre-Training for Visual Information Extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7092–7101, 2023.

[297] Rui Luo, Jianhong Wang, Yaodong Yang, Jun WANG, and Zhanxing Zhu. Thermostat-Assisted Continuously-Tempered Hamiltonian Monte Carlo for Bayesian Learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

[298] Xingchen Ma and Matthew B. Blaschko. Meta-Cal: Well-Controlled Post-Hoc Calibration by Ranking. *Proceedings of machine learning research (PMLR)*, 2021.

[299] David JC MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[300] David JC MacKay. Probable Networks and Plausible Predictions—a Review of Practical Bayesian Methods for Supervised Neural Networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[301] 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]*, 32:13153–13164, 2019. arXiv: 1902.02476 version: 2.

[302] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. QAlayout: Question Answering Layout Based on Multimodal Attention for Visual Question Answering on Corporate Document. In *Document Analysis Systems*, pages 659–673, Cham, 2022. Springer International Publishing.

[303] Himanshu Maheshwari, Sumit Shekhar, Apoorv Saxena, and Niyati Chhaya. Open-World Factually Consistent Question Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2390–2404, 2023.

[304] Samay Maini, Alexander Groleau, Kok Wei Chee, Stefan Larson, and Jonathan Boarman. Augraphy: A Data Augmentation Library for Document Images. *arXiv preprint arXiv:2208.14558*, 2022.

[305] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble Distribution Distillation. In *International Conference on Learning Representations*, 2019.

[306] Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, et al. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[307] Markos Markou and Sameer Singh. Novelty Detection: A Review—part 1: Statistical Approaches. *Signal processing*, 83(12):2481–2497, 2003.

[308] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document Visual Question Answering Challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020.

[309] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A Dataset for Vqa on Document Images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[310] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[311] Puneet Mathur, Rajiv Jain, Jiuxiang Gu, Franck Dernoncourt, Dinesh Manocha, and Vlad I Morariu. DocEdit: Language-Guided Document Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1914–1922, 2023.

[312] Andrew Kachites McCallum. Multi-Label Text Classification With a Mixture Model Trained by EM. In *AAAI 1999 Workshop on Text Learning*. Citeseer, 1999.

[313] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering Ambiguous Open-Domain Questions, 2020.

[314] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020.

[315] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-Vqa: Visual Question Answering by Reading Text in Images. In *2019 International Conference on document analysis and recognition (ICDAR)*, pages 947–952. Ieee, 2019.

[316] Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. NumGLUE: A Suite of Fundamental Yet Challenging Mathematical Reasoning Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland, 2022. Association for Computational Linguistics.

[317] Tom M Mitchell. Artificial Neural Networks. *Machine learning*, 45(81):127, 1997.

[318] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A Question Answering Dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, 2020. Association for Computational Linguistics.

[319] Hans Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

[320] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A Unifying View on Dataset Shift in Classification. *Pattern Recognition*, 45(1):521–530, 2012.

[321] Lili Mou, Hao Zhou, and Lei Li. Discreteness in Neural Natural Language Processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019. Association for Computational Linguistics.

[322] Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-Aware Self-Training for Few-Shot Text Classification. In *Advances in Neural Information Processing Systems*, Online, 2020.

[323] Jishnu Mukhoti, Pontus Stenetorp, and Yarin Gal. On the Importance of Strong Baselines in Bayesian Deep Learning. *arXiv preprint arXiv:1811.09385*, 2018.

[324] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating Deep Neural Networks Using Focal Loss. 2020.

[325] 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.

[326] 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.

[327] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When Does Label Smoothing Help? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[328] Thisanaporn Mungmeeprued, Yuxin Ma, Nisarg Mehta, and Aldo Lipani. Tab This Folder of Documents: Page Stream Segmentation of Business Documents. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, pages 1–10, 2022.

[329] 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.

[330] Allan H. Murphy and Robert L. Winkler. Scoring Rules in Probability Assessment and Evaluation. *Acta Psychologica*, 34:273–286, 1970.

[331] 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.

[332] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

[333] Eric Nalisnick, José Miguel Hernández-Lobato, and Padhraic Smyth. Dropout as a Structured Shrinkage Prior, 2018.

[334] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? In *International Conference on Learning Representations*, 2018.

[335] Hongseok Namkoong. *Reliable Machine Learning via Distributional Robustness*. Stanford University, 2019.

[336] RM Neal. Bayesian Learning for Neural Networks. *Springer New York*, 1996.

[337] Radford M Neal. Bayesian Mixture Modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

[338] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview Of~BioASQ 2022: The Tenth BioASQ Challenge On~Large-Scale Biomedical Semantic Indexing And~Question Answering. In *Lecture Notes in Computer Science*, pages 337–361. Springer International Publishing, 2022.

[339] 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.

[340] 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.

[341] Giannis Nikolentzos, Antoine Tixier, and Michalis Vazirgiannis. Message Passing Attention Networks for Document Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8544–8551, 2020.

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

[343] Andrew Nobel. Histogram Regression Estimation Using Data-Dependent Partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

[344] R OpenAI. GPT-4 Technical Report. *arXiv*, pages 2303–08774, 2023.

[345] Ian Osband. Risk Versus Uncertainty in Deep Learning : Bayes , Bootstrap and the Dangers of Dropout. In *NIPS Workshop on Bayesian Deep Learning*, 2016.

[346] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic Neural Networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[347] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*, 2023.

[348] 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.

[349] Anusri Pampari and Stefano Ermon. Unsupervised Calibration Under Covariate Shift. *arXiv preprint arXiv:2006.16405*, 2020.

[350] 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.

[351] Nicolas Papernot and Patrick McDaniel. Deep K-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv preprint arXiv:1803.04765*, 2018.

[352] Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. BioMRC: A Dataset for Biomedical Machine Reading Comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online, 2020. Association for Computational Linguistics.

[353] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning About the Dynamic Context of a Still Image, 2020.

[354] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated Prediction With Covariate Shift via Unsupervised Domain Adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. Pmlr, 2020.

[355] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[356] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous Knowledge Distillation Using Information Flow Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2020.

[357] Panupong Pasupat and Percy Liang. Compositional Semantic Parsing on Semi-Structured Tables. *arXiv preprint arXiv:1508.00305*, 2015.

[358] Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. Is a Question Decomposition Unit All We Need? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, 2022.

[359] Tzuf Paz-Argaman and Reut Tsarfaty. RUN Through the Streets: A New Dataset and Baseline Models for Realistic Urban Navigation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6449–6455, Hong Kong, China, 2019. Association for Computational Linguistics.

[360] Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in Neural Networks: Approximately Bayesian Ensembling. In *International Conference on Artificial Intelligence and Statistics*, pages 234–244. Pmlr, 2020.

[361] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing Neural Networks by Penalizing Confident Output Distributions. *ICLR Workshops*, 2017.

[362] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3743–3751, 2022.

[363] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient Neural Architecture Search via Parameters Sharing. In *International conference on machine learning*, pages 4095–4104. Pmlr, 2018.

[364] Mary Phuong and Christoph H Lampert. Distillation-Based Training for Multi-Exit Architectures. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 1355–1364, 2019.

[365] Michał Pietruszka, Michał Turski, Łukasz Borchmann, Tomasz Dwojak, Gabriela Pałka, Karolina Szyndler, Dawid Jurkiewicz, and Łukasz Garncarek. STable: Table Generation Framework for Encoder-Decoder Models, 2022.

[366] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The Text Anonymization Benchmark (Tab): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4): 1053–1101, 2022.

[367] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A Review of Novelty Detection. *Signal Processing*, 99:215–249, 2014.

[368] 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.

[369] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational Vision and Regularization Theory. *Readings in Computer Vision*, pages 638–643, 1987.

[370] 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.

[371] Rafal Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Pałka. Going Full-Tilt Boogie on Document Understanding With Text-Image-Layout Transformer. In *Icdar*, 2021.

[372] Subhojeet Pramanik, Shashank Mujumdar, and Hima Patel. Towards a Multi-Modal, Multi-Task Learning Based Pre-Training Framework for Document Representation Learning. *arXiv preprint arXiv:2009.14457*, 2020.

[373] Adarsh Prasad. *Towards Robust and Resilient Machine Learning*. PhD thesis, 2022.

[374] Ofir Press, Noah Smith, and Mike Lewis. Train Short, Test Long: Attention With Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*, 2021.

[375] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. DuReader_vis: A Chinese Dataset for Open-Domain Document Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1338–1351, Dublin, Ireland, 2022. Association for Computational Linguistics.

[376] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. LGPMA: Complicated Table Structure Recognition With Local and Global Pyramid Mask Alignment. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I*, pages 99–114. Springer, 2021.

[377] Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. T5Score: Discriminative Fine-Tuning of Generative Evaluation Metrics. *arXiv preprint arXiv:2212.05726*, 2022.

[378] 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.

[379] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[380] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models Are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

[381] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021.

[382] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning With a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[383] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the Limits of Transfer Learning With a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(140): 1–67, 2020.

[384] Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. emrKBQA: A Clinical Knowledge-Base Question Answering Dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online, 2021. Association for Computational Linguistics.

[385] Sachin Raja, Ajoy Mondal, and C. Jawahar. ICDAR 2023 Competition on Visual Question Answering on Business Document Images, 2023.

[386] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250*, 2016.

[387] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics.

[388] Alan Ramponi and Barbara Plank. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, 2020.

[389] Sebastian Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv preprint arXiv:1811.12808*, 2018.

[390] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-Ood Detection. *arXiv preprint arXiv:2106.09022*, 2021.

[391] Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-Evaluation Improves Selective Generation in Large Language Models. *arXiv preprint arXiv:2312.09300*, 2023.

[392] M. Rimol. Gartner Forecasts Worldwide Hyperautomation-Enabling Software Market to Reach Nearly $600 Billion by 2022. https://www.gartner.com/en/newsroom/press-releases/2021-04-28-gartner-forecasts-worldwide-hyperautomation-enabling-softwaremarket-to-reach-nearly-600-billion-by-2022, 2021. [Online; accessed Feb. 19, 2022].

[393] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating Bias in Calibration Error Estimation. *arXiv preprint arXiv:2012.08668*, pages 4036–4054, 2020.

[394] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for Thin Deep Nets. *arXiv preprint arXiv:1412.6550*, 2014.

[395] John Peter Rooney. IEC61508: An Opportunity for Reliability. In *Annual Reliability and Maintainability Symposium. 2001 Proceedings. International Symposium on Product Quality and Integrity (Cat. No. 01CH37179)*, pages 272–277. Ieee, 2001.

[396] Frank Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological review*, 65(6):386, 1958.

[397] Dan Roth. Learning to Resolve Natural Language Ambiguities: A Unified Approach. In *Aaai/iaai*, pages 806–813, 1998.

[398] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning Representations by Back-Propagating Errors. *nature*, 323(6088):533–536, 1986.

[399] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International journal of computer vision*, 115:211–252, 2015.

[400] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A Rossi, and Franck Dernoncourt. PDFTriage: Question Answering Over Long, Structured Documents. *arXiv preprint arXiv:2309.08872*, 2023.

[401] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the WILDS Benchmark for Unsupervised Adaptation. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[402] Clément Sage, Thibault Douzon, Alex Aussem, Véronique Eglin, Haytham Elghazel, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. Data-Efficient Information Extraction From Documents With Pre-Trained Language Models. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 455–469. Springer, 2021.

[403] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. ScienceQA: A Novel Resource for Question Answering on Scholarly Articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[404] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation Guidelines Improve Zero-Shot Information-Extraction. *arXiv preprint arXiv:2310.03668*, 2023.

[405] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM*, 64(9):99–106, 2021.

[406] Arthur L Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[407] Abdellatif Sassioui, Rachid Benouini, Yasser El Ouargui, Mohamed El Kamili, Meriyem Chergui, and Mohammed Ouzzif. Visually-Rich Document Understanding: Concepts, Taxonomy and Challenges. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–7. Ieee, 2023.

[408] Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question Answering Over Temporal Knowledge Graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676, Online, 2021. Association for Computational Linguistics.

[409] Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *Journal of Chemical Information and Modeling*, 2020.

[410] Karin Kipper Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. University of Pennsylvania, 2005.

[411] Nabeel Seedat and Christopher Kanan. Towards Calibrated and Scalable Uncertainty Representations for Neural Networks, 2019.

[412] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words With Subword Units. *arXiv preprint arXiv:1508.07909*, 2015.

[413] Burr Settles. Active Learning Literature Survey. *Science*, 10(3):237–304, 1995.

[414] Alireza Shafaei, Mark Schmidt, and James Little. A Less Biased Evaluation of Out-of-Distribution Sample Detectors. In *Bmvc*, 2019.

[415] Claude E Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[416] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. VILA: Improving Structured Content Extraction From Scientific PDFs Using Visual Layout Groups. *Transactions of the Association for Computational Linguistics*, 10: 376–392, 2022.

[417] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-Lora: Serving Thousands of Concurrent Lora Adapters. *arXiv preprint arXiv:2311.03285*, 2023.

[418] Hidetoshi Shimodaira. Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[419] Avanti Shrikumar and Anshul Kundaje. Maximum Likelihood With Bias-Corrected Calibration Is Hard-to-Beat at Label Shift Adaptation. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[420] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. Prompting GPT-3 to Be Reliable. In *The Eleventh International Conference on Learning Representations*, 2022.

[421] E. H. Simpson. Measurement of Diversity. *Nature*, 163(4148):688–688, 1949.

[422] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. DocILE Benchmark for Document Information Localization and Extraction. *arXiv preprint arXiv:2302.05658*, 2023.

[423] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A Foundational Language and Vision Alignment Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[424] Matyas Skalicky, Stepan Simsa, Michal Uricar, and Milan Sulc. Business Document Information Extraction: Towards Practical Benchmarks, 2022.

[425] 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.

[426] Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. In *Proceedings of the Conference on Uncertainty on Artificial Intelligence (UAI)*, 2018.

[427] Brandon Smock, Rohith Pesala, and Robin Abraham. PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.

[428] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution Calibration for Regression. In *International Conference on Machine Learning*, pages 5897–5906. Pmlr, 2019.

[429] Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. Event-Qa: A Dataset for Event-Centric Question Answering Over Knowledge Graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, page 3157–3164, New York, NY, USA, 2020. Association for Computing Machinery.

[430] Sargur Srihari, Stephen Lam, Venu Govindaraju, Rohini Srihari, Jonathan Hull, and E Yair. Document Understanding: Research Directions. In *DARPA Document Understanding Workshop, Xerox PARC, Palo Alto, CA*. Citeseer, 1992.

[431] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[432] Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. Kleister: Key Information Extraction Datasets Involving Long Documents With Complex Layouts. In *Icdar*, pages 564–579. Springer, 2021.

[433] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: Key Information Extraction Datasets Involving Long Documents With Complex Layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021.

[434] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does Knowledge Distillation Really Work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.

[435] J Stray and S Svetlichnaya. DeepForm: Extract Information From Documents (2020).

[436] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[437] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023.

[438] Nassim Nicholas Taleb and Raphael Douady. Mathematical Definition, Mapping, and Detection of (Anti) Fragility. *Quantitative Finance*, 13(11):1677–1689, 2013.

[439] Mingxing Tan and Quoc Le. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*, pages 6105–6114. Pmlr, 2019.

[440] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. VisualMRC: Machine Reading Comprehension on Document Images. In *Aaai*, 2021.

[441] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images, 2023.

[442] Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. *arXiv preprint arXiv:2401.13313*, 2024.

[443] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying Vision, Text, and Layout for Universal Document Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.

[444] Ajay Kumar Tanwani and Muddassar Farooq. Classification Potential vs. Classification Accuracy: A Comprehensive Study of Evolutionary Algorithms With Biomedical Datasets. In *Learning Classifier Systems*, pages 127–144. Springer, 2009.

[445] Choon Hui Teo, SVN Vishwanathan, Alex Smola, and Quoc V Le. Bundle Methods for Regularized Risk Minimization. *Journal of Machine Learning Research*, 11(1), 2010.

[446] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A Large-Scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[447] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Representation Distillation. In *International Conference on Learning Representations (ICLR)*, 2019.

[448] Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

[449] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document Collection Visual Question Answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021.

[450] Rubèn Tito, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2021 Competition on Document Visual Question Answering. In *International Conference on Document Analysis and Recognition*, pages 635–649. Springer, 2021.

[451] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical Multimodal Transformers for Multipage DocVQA. *Pattern Recognition*, 144:109834, 2023.

[452] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

[453] Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All You Need Is a Good Functional Prior for Bayesian Deep Learning. *arXiv preprint arXiv:2011.12829*, 2020.

[454] Dustin Tran, Michael W. Dusenberry, Danijar Hafner, and Mark van der Wilk. Bayesian Layers: A Module for Neural Network Uncertainty. In *Neural Information Processing Systems*, 2019.

[455] Dustin Tran, Jeremiah Zhe Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda E Mariet, Huiyi Hu, et al. Plex: Towards Reliability Using Pretrained Large Model Extensions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[456] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A Machine Comprehension Dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[457] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. Lc-Quad: A Corpus for Complex Question Answering Over Knowledge Graphs. In *International Semantic Web Conference*, pages 210–218. Springer, 2017.

[458] Yi Tu, Ya Guo, Huan Chen, and Jinyang Tang. LayoutMask: Enhance text-layout interaction in multi-modal pre-training for document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15200–15212, Toronto, Canada, 2023. Association for Computational Linguistics.

[459] R. E. Turner and M. Sahani. Two Problems With Variational Expectation Maximisation for Time-Series Models. In *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

[460] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. CCpdf: Building a High Quality Corpus for Visually Rich Documents From Web Crawl Data. *arXiv preprint arXiv:2304.14953*, 2023.

[461] Meet Vadera, Brian Jalaian, and Benjamin Marlin. Generalized Bayesian Posterior Expectation Distillation for Deep Neural Networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 719–728. Pmlr, 2020.

[462] Meet P Vadera, Adam D Cobb, Brian Jalaian, and Benjamin M Marlin. URSABench: Comprehensive Benchmarking of Approximate Bayesian Inference Methods for Deep Neural Networks. *arXiv preprint arXiv:2007.04466*, 2020.

[463] 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.

[464] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9690–9700. Pmlr, 2020.

[465] Jordy Van Landeghem, Matthew B Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[466] Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Benchmarking Scalable Predictive Uncertainty in Text Classification. *IEEE Access*, 2022.

[467] Jordy Van Landeghem, Lukasz Borchmann, Rubèn Tito, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *Proceedings of the ICDAR 2023*, pages 420–434. Springer, 2023.

[468] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Marie-Francine Moens, and Tomasz Stanisławek. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023.

[469] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *International Conference on Document Analysis and Recognition*, pages 420–434. Springer, 2023.

[470] Jordy Van Landeghem, Sanket Biswas, Matthew Blaschko, and Marie-Francine Moens. Beyond Document Page Classification: Design, Datasets, and Challenges. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2962–2972, 2024.

[471] Jordy Van Landeghem, Subhajit Maity, Ayan Banerjee, Matthew B Blaschko, Marie-Francine Moens, Josep Llados, and Sanket Biswas. DistilDoc: Knowledge Distillation for Visually-Rich Document Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (under review)*, 2024.

[472] Vladimir Vapnik. Principles of Risk Minimization for Learning Theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[473] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[474] Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, 2023.

[475] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Computer Law Review International*, 22(4):97–112, 2021.

[476] Csaba Veres and Jennifer Sampson. Self Supervised Learning and the Poverty of the Stimulus. *Data & Knowledge Engineering*, 147:102208, 2023.

[477] Tomás Vojír, Jan Sochman, Rahaf Aljundi, and Jirí Matas. Calibrated Out-of-Distribution Detection With a Generic Representation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4509–4518. Ieee, 2023.

[478] Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. Long-Length Legal Document Classification. *arXiv preprint arXiv:1912.06905*, 2019.

[479] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient Knowledge Distillation From Model Checkpoints. *Advances in Neural Information Processing Systems*, 35:607–619, 2022.

[480] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding, 2023.

[481] Jiapeng Wang, Lianwen Jin, and Kai Ding. LiLT: A Simple Yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757, 2022.

[482] Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. Layout and Task Aware Instruction Prompt for Zero-Shot Document Image Question Answering. *arXiv preprint arXiv:2306.00526*, 2023.

[483] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable Calibration With Lower Bias and Variance in Domain Adaptation. In *Advances in Neural Information Processing Systems*, pages 19212–19223. Curran Associates, Inc., 2020.

[484] Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. *arXiv preprint arXiv:2010.11685*, 2020.

[485] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. Vrdu: A Benchmark for Visually-Rich Document Understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193, 2023.

[486] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent Analogical Reasoning in Large Language Models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[487] 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.

[488] Jun Wen, Nenggan Zheng, Junsong Yuan, Zhefeng Gong, and Changyou Chen. Bayesian Uncertainty Matching for Unsupervised Domain Adaptation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3849–3855, 2019.

[489] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning. In *International Conference on Learning Representations*, 2019.

[490] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-Parametric Calibration for Classification. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 178–190. Pmlr, 2020.

[491] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good Is the Bayes Posterior in Deep Neural Networks Really? In *International Conference on Machine Learning*, pages 10248–10259. Pmlr, 2020.

[492] 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.

[493] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration Tests Beyond Classification. In *International Conference on Learning Representations*, 2021.

[494] Gregor Wiedemann and Gerhard Heyer. Multi-Modal Page Stream Segmentation With Convolutional Neural Networks. *Language Resources and Evaluation*, 55:127–150, 2021.

[495] Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite Multiclass Losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

[496] Andrew Gordon Wilson. The Case for Bayesian Deep Learning. *arXiv preprint arXiv:2001.10995*, 2020.

[497] Fangzhao Wu and Yongfeng Huang. Sentiment Domain Adaptation With Multiple Sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 301–310, 2016.

[498] Xinya Wu, Duo Zheng, Ruonan Wang, Jiashen Sun, Minzhen Hu, Fangxiang Feng, Xiaojie Wang, Huixing Jiang, and Fan Yang. A Region-Based Document VQA. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4909–4920, 2022.

[499] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[500] Yijun Xiao and William Yang Wang. Quantifying Uncertainties in Natural Language Processing Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7322–7329, 2019.

[501] Qunliang Xing, Mai Xu, Tianyi Li, and Zhenyu Guan. Early Exit or Not: Resource-Efficient Blind Quality Enhancement for Compressed Images. In *European Conference on Computer Vision*, pages 275–292. Springer, 2020.

[502] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-Training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

[503] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-Modal Pre-Training for Visually-Rich Document Understanding. *arXiv preprint arXiv:2012.14740*, 2020.

[504] Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. Towards Fine-Grained Causal Reasoning and QA, 2022.

[505] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. SGM: Sequence Generation Model for Multi-Label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

[506] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[507] Yuzhe Yang, Hao Wang, and Dina Katabi. On Multi-Domain Long-Tailed Recognition, Generalization and Beyond. *arXiv preprint arXiv:2203.09513*, page 57–75, 2022.

[508] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics.

[509] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach With Normalized Loss and Customized Soft Labels. *arXiv preprint arXiv:2303.13005*, 2023.

[510] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. UReader: Universal OCR-free Visually-Situated Language Understanding With Multimodal Large Language Model, 2023.

[511] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017.

[512] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*, 2023.

[513] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, 2017. Association for Computational Linguistics.

[514] Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. End-to-End Spoken Conversational Question Answering: Task, Dataset and Model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States, 2022. Association for Computational Linguistics.

[515] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning From Multiple Teacher Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017.

[516] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[517] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. Pick: Processing Key Information Extraction From Documents Using Improved Graph Learning-Convolutional Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. Ieee, 2021.

[518] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central Similarity Quantization for Efficient Image and Video Retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3083–3092, 2020.

[519] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-Labeling Imagenet: From Single to Multi-Labels, From Global to Localized Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021.

[520] 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.

[521] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

[522] Hugo Zaragoza and Florence d'Alché Buc. Confidence Measures for Neural Network Classifiers. In *Proceedings of the Seventh International Conference Information Processing and Management of Uncertainty in Knowledge Based Systems*, 1998.

[523] Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. ParsFEVER: A Dataset for Farsi Fact Extraction and Verification. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 99–104, Online, 2021. Association for Computational Linguistics.

[524] Dell Zhang, Murat Sensoy, Masoud Makrehchi, Bilyana Taneva-Popova, Lin Gui, and Yulan He. Uncertainty Quantification for Text Classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3426–3429, 2023.

[525] Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. Complex Question Decomposition for Semantic Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4477–4486, 2019.

[526] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-N-Match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. In *International Conference on Machine Learning*, pages 11117–11128. Pmlr, 2020.

[527] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD V1.5: Enhanced Benchmark for Out-of-Distribution Detection. *arXiv preprint arXiv:2306.09301*, 2023.

[528] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[529] Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. NOAHQA: Numerical Reasoning With Interpretable Graph Question Answering Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[530] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical Stochastic Gradient McMc for Bayesian Deep Learning. *arXiv:1902.03932 [cs, stat]*, 2019. arXiv: 1902.03932.

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

[532] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction Tuning for Large Language Models: A Survey. *arXiv preprint arXiv:2308.10792*, 2023.

[533] Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. Mitigating Uncertainty in Document Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3126–3136, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[534] Xinbo Zhang, Changzhi Sun, Yue Zhang, Lei Li, and Hao Zhou. NAIL: A Challenging Benchmark for Na\"ive logical reasoning, 2022.

[535] Xi Zhang, Feifei Zhang, and Changsheng Xu. VQACL: A Novel Visual Question Answering Continual Learning Setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19102–19112, 2023.

[536] Xu-Yao Zhang, Guo-Sen Xie, Xiuli Li, Tao Mei, and Cheng-Lin Liu. A Survey on Learning to Reject. *Proceedings of the IEEE*, 111(2):185–215, 2023.

[537] Ye Zhang and Byron Wallace. A Sensitivity Analysis of (And Practitioners' Guide To) Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1510.03820*, 2015.

[538] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep Mutual Learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[539] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3388–3397, 2023.

[540] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling Effective Supervision From Severe Label Noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303, 2020.

[541] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[542] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.

[543] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global Table Extractor (Gte): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021.

[544] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest Dataset Ever for Document Layout Analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. Ieee, 2019.

[545] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-Based Table Recognition: Data, Model, and Evaluation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 564–580. Springer, 2020.

[546] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[547] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert Loses Patience: Fast and Robust Inference With Early Exit. *Advances in Neural Information Processing Systems*, 33:18330–18341, 2020.

[548] Xichuan Zhou, Haijun Liu, Cong Shi, and Ji Liu. *Deep Learning on Edge Computing Devices: Design Challenges of Algorithm and Architecture*. Elsevier, 2022.

[549] 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.

[550] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards Complex Document Understanding by Discrete Reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*. Acm, 2022.

[551] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards Complex Document Understanding by Discrete Reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022.

[552] 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.

[553] Guangyu Zhu and David Doermann. Automatic Document Logo Detection. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 864–868. Ieee, 2007.

[554] Michael Zhu and Suyog Gupta. To Prune, or Not to Prune: Exploring the Efficacy of Pruning for Model Compression. *arXiv preprint arXiv:1710.01878*, 2017.

[555] Xi Zhu, Xue Han, Shuyuan Peng, Shuo Lei, Chao Deng, and Junlan Feng. Beyond Layout Embedding: Layout Attention With Gaussian Biases for Structured Document Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7773–7784, Singapore, 2023. Association for Computational Linguistics.

[556] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A Survey on Model Compression for Large Language Models. *arXiv preprint arXiv:2308.07633*, 2023.

[557] Yftah Ziser and Roi Reichart. Neural Structural Correspondence Learning for Domain Adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 400–410, Vancouver, Canada, 2017. Association for Computational Linguistics.

Appendices

Appendix A

Appendix - PUQ

A Implementation Details

In this Section, we describe the implementation details for the different datasets, architectures and inference methods used in our benchmark.

A.1 Software and Data

We have published our benchmarking software at `https://github.com/Jordy-VL/uncertainty-bench` so that the community can continue to build on our work. We have added detailed instructions for reproducibility and extensibility. This allows anyone to test on a new dateset of interest, implement a new uncertainty estimation method, or evaluate on our benchmark datasets.

A.2 Hyperparameter Defaults

For each baseline architecture and uncertainty method combination, we describe hyperparameter values in detail for facilitating future replication.

Our choice of hyperparameter values for TextCNN is heavily based on [537], for fine-tuning BERT on [94, 322] and we draw inspiration from [500] for uncertainty estimation method parameters. We seek to restrict hyperparameters as much as empirically plausible to 1 static setting over datasets per architecture.
We constrain the input vocabulary to the 20,000 most frequent words (30K for

`20news` and `AAPD`), retain the original document lengths, remapping tokens with a frequency lower than 3 to *UNK* and *PAD* tokens are masked throughout. For TextCNN 300-D embeddings are uniformly initialized upon which three different kernels (3,4,5) operate with 100 feature maps per kernel followed by a max pooling operation. For BERT we tokenize and encode using the standard BERT tokenizer with maximum sequence length determined per dataset [`20news`: 250, `CLINC`: 50, `IMDB`: 350 and `Reuters/AAPD`: 200].

Following the MC Dropout procedure we apply dropout [431] with a rate of 0.5 after each non-linear weights layer. We found a global weight decay rate of 1e-4 [224, 293] to work well for TextCNN, whereas we disabled weight decay for BERT since it overpenalized model complexity, resulting in vanishing gradients. During training TextCNN, Adam optimizes cross-entropy or heteroscedastic loss (see Section 3.3.2.4) with a learning rate of 1e-3 for 45 epochs on batches of size 32. For fine-tuning BERT, we schedule the learning rate starting from 1e-5 to 1e-6 with batch size 16 and train for 20 epochs (longer than the original recommendation, following [436]). We use early stopping conditioned on the validation loss with sufficient epochs to ensure all models are trained until convergence. Else the models might have learned to approximate well the mean of the predictive posterior distribution, but not the variance. At evaluation time, we estimate predictive mean and uncertainties by drawing T samples from the approximated predictive posterior distribution or by averaging over M models. We have empirically set T to 10 and for ensembles the number of models M to 5.

B Practical Considerations

B.1 Take-home Summary

Concretely, for a multi-class problem with a large number of classes, incorporating input-dependent data uncertainty improves accuracy and novelty detection. With high label cardinality in multi-label classification, we recommend ensembling for more reliable epistemic uncertainty estimation. More generally, we advise against using *MC Dropout* if the dropout rate and weight regularization are not fine-tuned for the problem at hand, drawing parallels to dropout probability rates adaptively learned with *Concrete Dropout*.

Hyperparameter considerations We reiterate important hyperparameters and reasonable defaults for text classification tasks similar to our benchmark setup and applications of the above.

- Dropout rate p: the original work suggested a fixed binary rate (p=0.5), whereas our experiments indicate different rates are more applicable per dataset. It is best to cross-validate layer-wise dropout probabilities for any real-world application, where impossible it warrants the low effort of incorporating Concrete Dropout, consequently reducing experimentation time.
- Weight decay $L2$: best to start with small values [1e-6 - 1e-4] and fine-tune accordingly. Take note to not apply global weight decay in case of pretrained weights, which already have high weight magnitudes, possibly impeding learning.
- MC Dropout T: a small number (T=10) of stochastic samples suffices, if large number of classes, scale sub-linearly with K. T also applies to the number of samples drawn to calculate heteroscedastic loss, so beware increasing to too large values since it affects training compute.
- Ensemble size M: a total of (M=5) ensemble models is plenty, certainly when combining with fine-tuned dropout rate at the individual model level.

B.2 Compute vs. Performance Trade-off

Next to performance, practitioners are generally concerned with computational and memory costs. [462] present similar concerns in the benchmarking of uncertainty methods. Considering the cost of compute vs. storage, each uncertainty method impacts both differently. Following [348], we present computational and memory costs for evaluated methods symbolically (Big-O), with m flops or storage for a trained model, l represents flops or storage for the last layer, T denotes sampling or replications, and ι GP inducing points.

Table A.1. Compute and storage costs in Big-O notation [348] for uncertainty methods.

Method	Compute/N	Storage
Baseline	m	m
MC (Concrete) Dropout	mT	m
Heteroscedastic	$m + l(T-1)$	$m(+l)$
Deep Ensemble	mT	mT
cSGMCMC	m	mT
SNGP	$m + \iota^2$	m

Our experiments were carried out on a system with a Intel Core i7-10750H 2.6 GHz CPU and NVIDIA GeForce RTX 2070 Max-Q GPU.
Additionally, we provide an informative table with training (*Table* A.2) and test (*Table* A.3) timings provided over all single models on `CLINC-OOS`.

Table A.2. CLINC-OOS models with training timings (in seconds) per epoch and total running time.

methods	**architecture**	**train time/epoch**	**epoch finished**	**train runtime**
Unregularized	TextCNN	32	8	256
Regularized	TextCNN	32	28	896
Heteroscedastic	TextCNN	59	17	1003
Concrete Dropout	TextCNN	35	12	420
Heteroscedastic Concrete Dropout	TextCNN	58	10	580
Unregularized	BERT	420	5	2100
Regularized	BERT	691	11	7601
Heteroscedastic	BERT	710	16	11360
Concrete Dropout	BERT	679	9	6111
Heteroscedastic Concrete Dropout	BERT	707	16	11312

Table A.3. CLINC-OOS models with inference timings presented in unit time for how many batches or samples can be processed in 1 second wall-clock time over CPU and GPU. For the short sequences of CLINC, both models allow a batch size of 32.

architecture	method	# batch (gpu)	# sample (gpu)	# batch (cpu)	# sample (cpu)
TextCNN	Unregularized	59.0	1891	63.0	2043
TextCNN	Regularized	66.0	2134	60.0	1922
TextCNN	MC Dropout	53.0	1708	32.0	1050
TextCNN	Heteroscedastic	693.0	22176	482.0	15444
TextCNN	MC Heteroscedastic	47.0	1525	38.0	1216
TextCNN	Concrete Dropout	66.0	2130	40.0	1293
TextCNN	MC Concrete Dropout	48.0	1541	25.0	827
TextCNN	Heteroscedastic Concrete Dropout	756.0	24205	318.0	10197
TextCNN	MC Heteroscedastic Concrete Dropout	48.0	1561	27.0	874
BERT	Unregularized	6.0	223	0.8	25
BERT	Regularized	9.0	306	0.8	26
BERT	MC Dropout	0.9	28	0.1	2
BERT	Heteroscedastic	10.0	325	0.8	26
BERT	MC Heteroscedastic	1.0	31	0.1	2
BERT	Concrete Dropout	7.0	245	0.9	27
BERT	MC Concrete Dropout	1.0	30	0.1	2
BERT	Heteroscedastic Concrete Dropout	6.0	218	0.9	27
BERT	MC Heteroscedastic Concrete Dropout	0.9	30	0.1	2

C Detailed Experiment Results

C.1 Zoom-in Benchmark Evidence

In this Subsection we report additional evidence in support of our results, which did not suit the main manuscript.

C.2 Absolute Benchmark Results

Next to reporting critical differences to analyze the relative performance of uncertainty methods, we also report results as summary statistics, following the

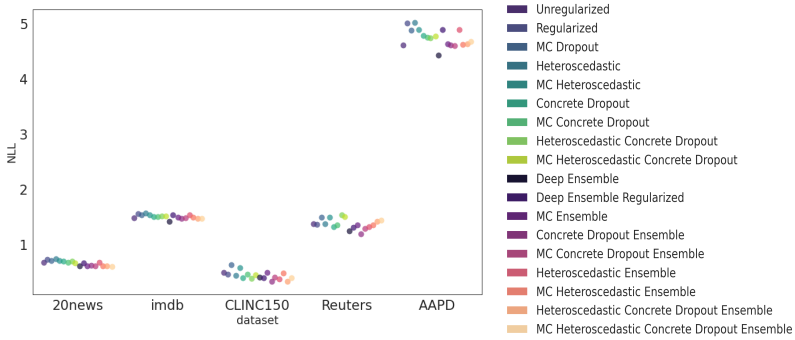


Figure A.1. Comparison with NLL(↓) for dataset-specific differences in method performance.

methodology of [462]. Firstly, we report performance averaged over both runs and datasets, with the standard deviation over datasets. We indicate the best mean performance in bold. For various metrics the standard deviation is very large, which shows that the average over datasets for our benchmark would be a poor measure of central tendency. Since we benchmark on three multiclass and two multilabel datasets, any aggregate would be biased towards multiclass performance, hence why we specifically opted for rank and critical difference to analyze relative performance of each method.

Additionally, we compute the performance averaged over datasets, with the standard deviation over multiple runs for all individual models. All raw model results are available at `https://github.com/Jordy-VL/uncertainty-bench/tree/main/experiments/raw_results`. We refer to the original paper for the larger detail tables with results averaged over datasets and runs.

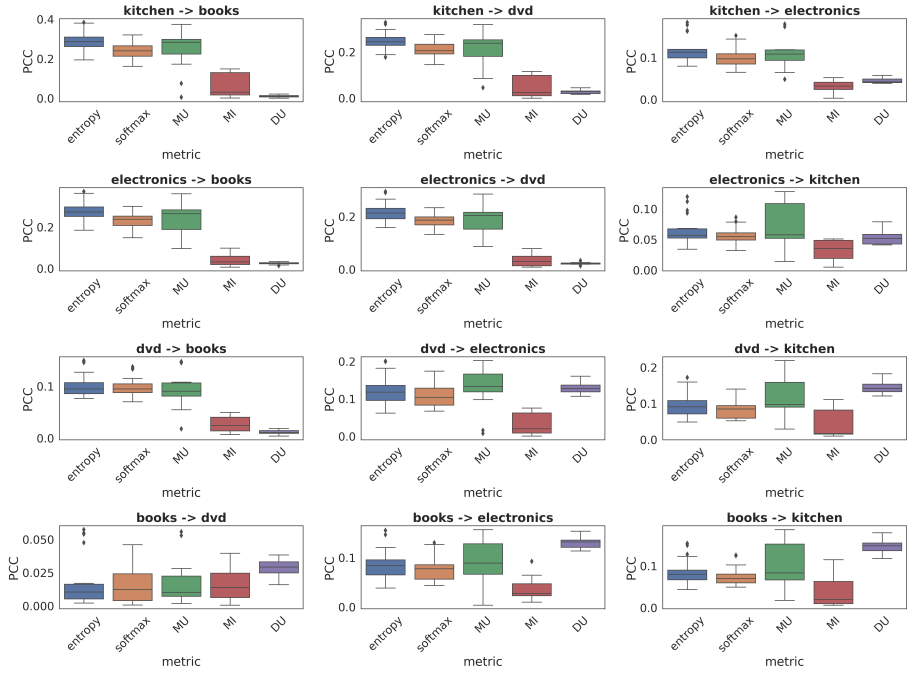
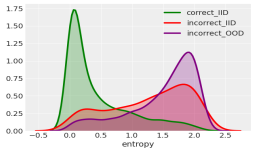
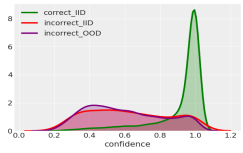


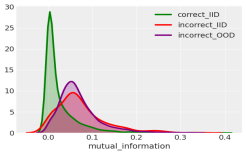
Figure A.2. We report the Pearson Correlation Coefficient (PCC) between uncertainty values and binary variable ID-OOD for Amazon product review datasets. A higher absolute correlation score points to stronger association of uncertainty and out-of-domain detection. *Model Uncertainty (MU), Data Uncertainty (DU), Mutual Information (MI).



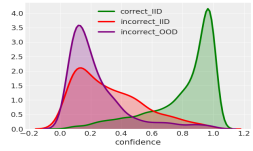
(a) Heteroscedastic Ensemble - H



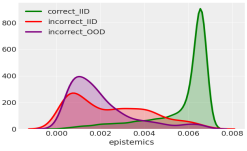
(b) MC Ensemble - S



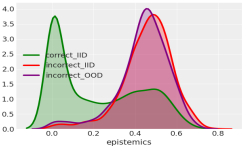
(c) Deep Ensemble Regularized - MI



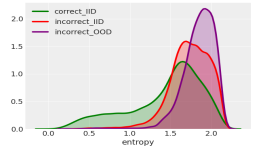
(d) Heteroscedastic CD Ensemble - S



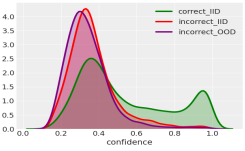
(e) Deep Ensemble - MU



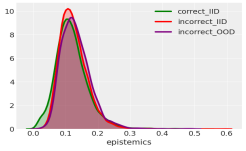
(f) MC Dropout - MU



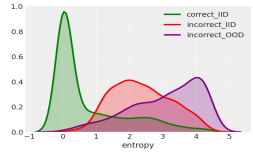
(g) Heteroscedastic Ensemble - H



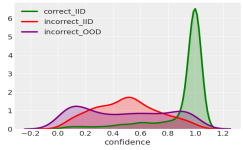
(h) Concrete Dropout - S



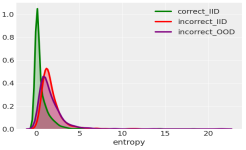
(i) MC Concrete Dropout - MU



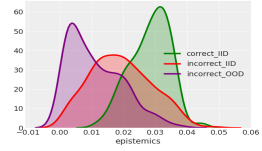
(j) MC CD Ensemble - H



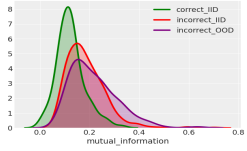
(k) Concrete Dropout Ensemble - S



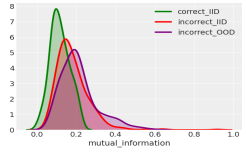
(l) Deep Ensemble - H



(m) Deep Ensemble - MU



(n) Heteroscedastic Ensemble - MI



(o) MC Dropout - MI

Figure A.3. A selection of most interesting Gaussian kernel density plots over (abbreviated) model setup metrics evaluated on all datasets in row order `20news` (a-c), `CLINC150` (d-f), `imdb` (g-i), `Reuters` (j-l), `AAPD` (m-o). Each plot captures probabilistic density over correct ID (green), incorrect ID (red) and OOD (purple). From left to right, we have selected a high rank, middle rank, and low-rank method and uncertainty quantity combination. The density estimates demonstrate clear empirical difference over all datasets for various uncertainty quantities.

Appendix B

Appendix - BDPC

A Existing DC Datasets

As the datasets from Table 2 did not satisfy large-scale benchmarking multipage DC benchmarking requirements, we discuss them in supplementary for interested readers.

Tobacco-3482 [232] is another subset of IIT-CDIP with fewer samples and a smaller label set than RVL-CDIP.

Tobacco-800 [553] has been used for page stream segmentation ([494], similarly defined as in [328]) as it contains consecutively numbered multipage business documents.

NIST The NIST Structured Forms Database [98] consists of 5,590 binary synthesized documents from 20 different classes of tax forms.

MARG The MARG (Medical Article Records Groundtruth) database [290] is a layout-based classification benchmark containing 1553 documents which are mainly the first pages of medical journals.

TAB [328] is a recently introduced page stream segmentation dataset targeting binary classification to detect document boundaries on multipage streams. It consists of a sample of 44,769 PDF documents from the Truth Tobacco Industry Documents (TTID) archives.

B Visualization of Proposed DC Datasets

As we have contributed two novel datasets consisting of multipage documents in PDF format, adding visualizations is non-trivial. The datasets are hosted at the HuggingFace Hub (https://huggingface.co/datasets/bdpc), for which at the time of submission, the dataset viewer does not support PDF data. Rather than adding examples in the manuscript, which is tedious for PDF documents with multiple pages, we have built an interactive app (https://huggingface.co/spaces/jordyvl/viz_bdpc). This allows for the visualization of samples from the proposed datasets, with an additional filter on the labels, whereas both datasets follow the original RVL-CDIP label taxonomy.

Appendix C

Appendix - DUDE

A Baseline Experiments Setup

In this Section, we describe the implementation details[1] for the architectures and inference methods used in our benchmark.

A.1 Hyperparameter Defaults

Refer to Table C.1.

A.2 Generative LLM Prompt Fine-tuning

The performance of GPT3.5 models was assessed in two settings: 0-shot and 4-shot. In the 0-shot setting, the prompt included instructions similar to those provided to annotators to teach them how to annotate. In the 4-shot setting, the prompt was enhanced with the content of a single document from the training set along with four questions of different types (extractive, abstractive, list, and not answerable) to better gauge the models' abilities.

The 0-shot prompt is analogous to the 4-shot prompt, but the key distinction is that it lacks the first document and the example question-and-answer pairs.

[1] Main framework used: https://github.com/rubenpt91/MP-DocVQA-Framework

Hyper-Parameter	T5	T5+2D	HiVT5
Epochs	10	10	10
Warm-up (iterations)	1000	250	1000
Optimizer	Adam, AdamW	Adafactor	Adam
Gradient acc.	False	8	False
Lower case	True	True	True
Max. Seq. Length	512, 8192	512, 8192	20480
Generation (Max. Tokens)	100	100	50
Batch size	3	8	1
Learning rate	1E-04, 2E-04	2E-04	2E-04
Training time (per epoch)	1h, 10h	1.5h, 5h	10h
GPU Hardware	TITAN RTX, A100	A100 (80GB)	TITAN RTX (24GB)

Table C.1. Hyperparameters used for fine-tuning `T5`, `T5-2D` and `HiVT5` on **DUDE**. When two values are placed in a single column, they refer to the model's versions with 512 and 8192 input sequence length, respectively.

For the inference process, we utilized the prompt completion default settings outlined in the OpenAI documentation, with the exception of the temperature parameter, which was lowered to a value of 0.0. This adjustment was made to ensure that the output would be more deterministic and focused, with less emphasis on generating creative variations.

Only after our prompting experiments had been completed, we realized the opportunity to assess confidence estimation using chained prompts (*Please give a confidence between 0 and 1 about how certain you are this is the answer.*) as in [219]. Since we did not save our dialogue states and considered the expenses, we leave this for future work.

A.3 Confidence Estimation

This Subsection details confidence scoring functions for the baselines, as this is not reported in standard practice.

We define *confidence* as the predicted probability of the top-1 prediction, often arising as the largest value from softmax normalization of logits from a final model layer (head).

Encoder-based models will output logits for all possible start and end positions of the answer within the provided context. While the predicted answer of such a span prediction architecture will come from the highest valid (no negative span) combination of the sum of a start and end logit, the predicted answer confidence can be obtained by the following procedure (BS: batch size and S: sequence length):

```
% # 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]
```

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). The following confidence estimation procedure is applied for decoder outputs:

```
# 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)
```

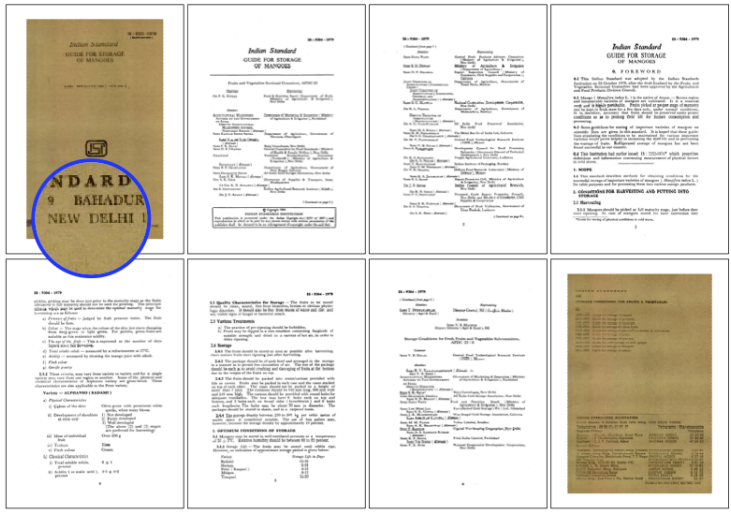
A.4 Evaluation

All metric implementations (ANLS, ECE, AURC) are made available as a standalone repository. Additionally, we provide an online service where researchers can evaluate their methods against a blind (questions-only) test dataset. General metric descriptions are provided in Section 2.2.3 with additional implementation details and motivated design choices. While ANLS can account for shortcomings of OCR and formatting issues, evaluation of generated text is notoriously complex [377] and requires more research.

B Qualitative Examples

As is customary, we provide some interesting, handpicked test set examples with predictions from some of the baselines in our study.

Low complexity. *Where the document has been printed?*
Simple, extractive question, plain-text evidence.



Source	Answer	ANLS	Conf.
Ground truth	*New Delhi, India*		
Human	*India*	0.0	—
T5	*IS : 9304 - 1979*	0.0	0.56
ChatGPT	*The document does not mention where it has been printed.*	0.0	—
GPT3	*Bela Pack n Print. New Delhi, India*	0.0	—
T5-2D	*New Delhi, India*	1.0	0.09
HiVT5	*Page 1*	0.0	0.18
Longformer	new delhi, india	1.0	0.72

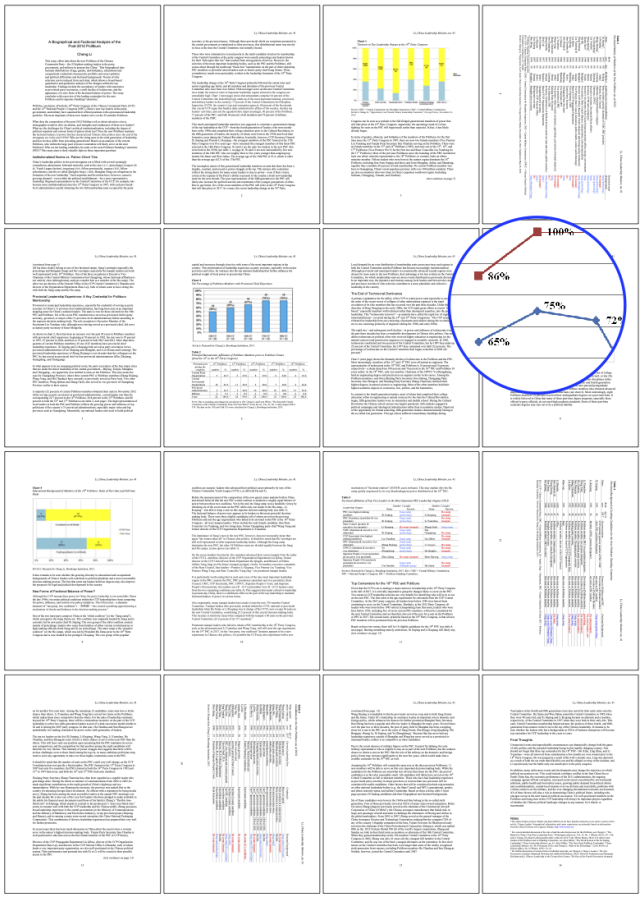
High complexity. *Is there any redacted section on the document?*
Abstractive question that requires knowledge about possible document elements.

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

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.

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

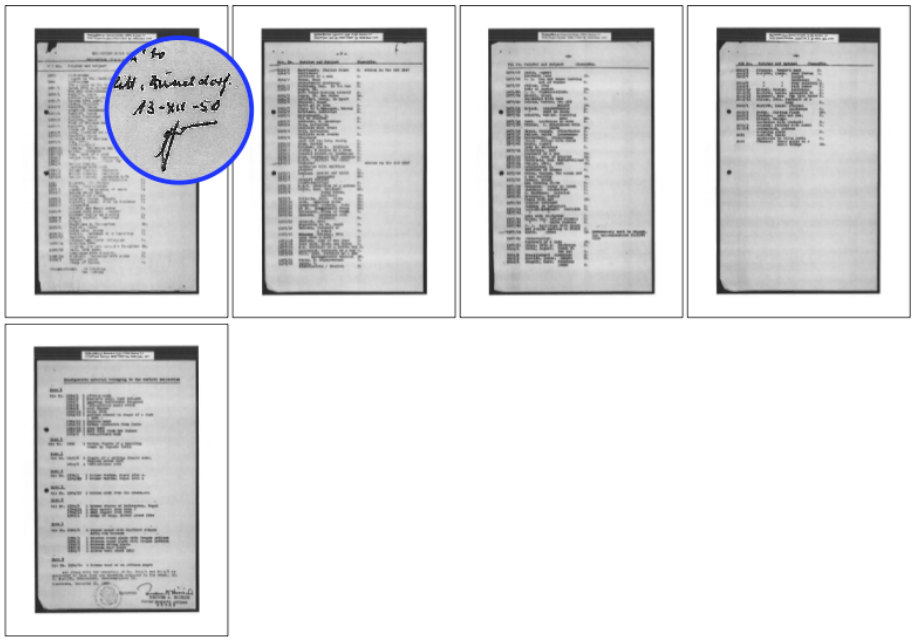
Visual evidence (chart). *What is the maximum percentage of the blue graph line on page 8?*
A highly demanding question that requires simultaneous competency of visual comprehension (locating chart and line color), navigating through layout (determining adequate page), and numerical comparison (deciding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



Source	Answer	ANLS	Conf.
Ground truth	*75%*		
Human	*75*	0.7	—
T5	*76*	0.0	0.25
ChatGPT	*[Not-answerable]*	0.0	—
GPT3	*76%*	0.7	—
T5-2D	*32.0*	0.0	0.00
HiVT5	*45%*	0.7	0.05
BigBird	*32*	0.0	0.47
LayoutLMv3	*80%*	0.0	0.15

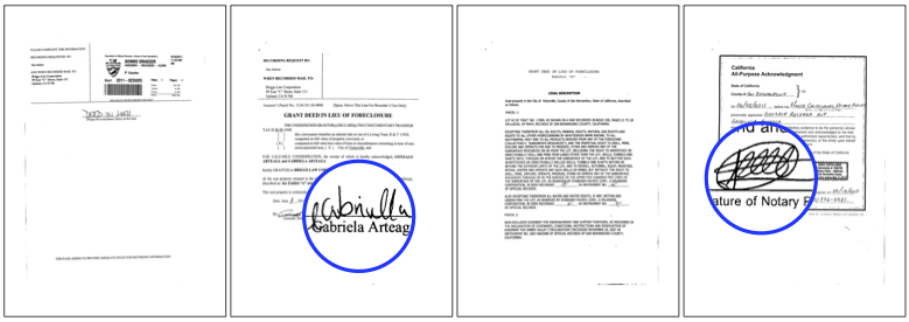
Visual evidence (handwriting). *What is the handwritten date on page 1?*

The question requires visual comprehension (recognition of handwriting) and layout navigation (determining the adequate page).



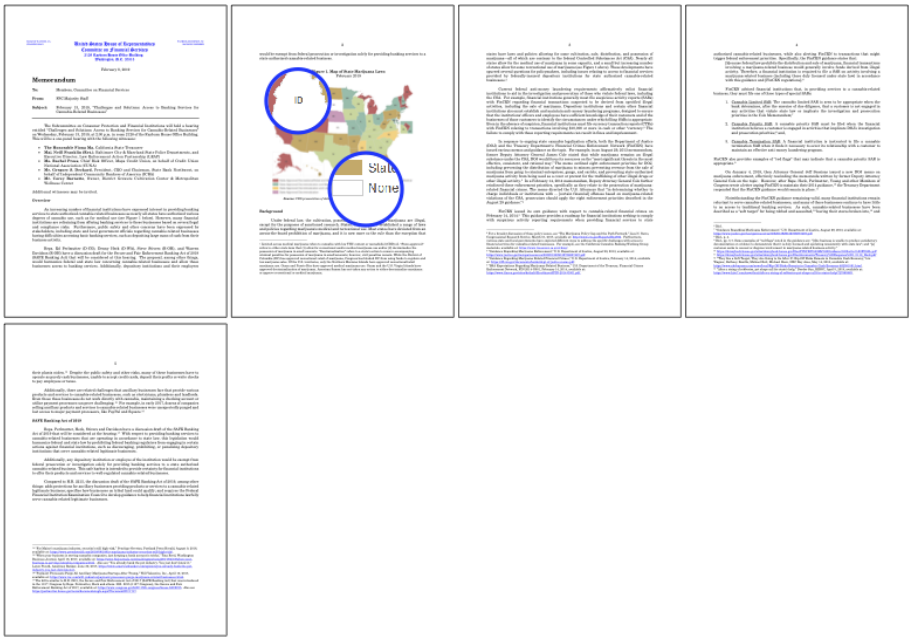
Source	Answer	ANLS	Conf.
Ground truth	*13-XII-50*		
Human	*13-XII-50*	1.0	—
T5	*1977-01-01*	0.0	0.24
ChatGPT	*[Not-answerable]*	0.0	—
GPT3	*15 December 1950*	0.0	—
T5-2D	*1950-12-15*	0.0	0.24
HiVT5	*1977-07-01*	0.0	0.11
BERTQA	*2006 / 1*	0.0	0.5

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



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

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



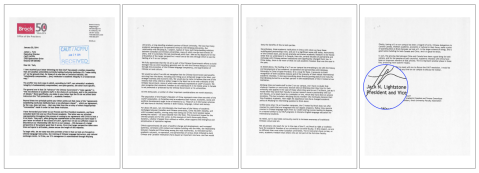
Source	Answer	ANLS	Conf.
Ground truth	*ID / SD / KS*		
Human	*ID / SD / KS*	1.0	—
T5	*WA ME MT ND MN OR VT ID NH SD WI NY MA MI*	0.0	0.28
ChatGPT	*[Not-answerable]*	0.0	—-
GPT3	*American Samoa*	0.0	—-
T5-2D	*i*	0.0	0.03
HiVT5	-	0.0	0.02

B.1 Qualitative Examples - Competition

We provide some interesting, hand-picked test set examples with predictions from the submitted competition methods.

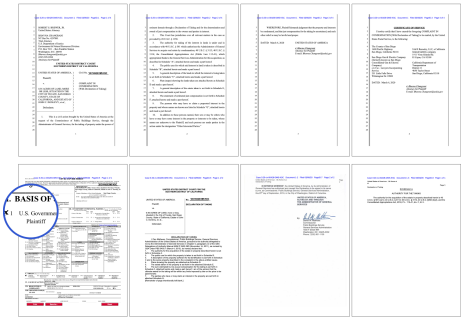
Low complexity. *Who is the president and vice-chancellor?* Despite the question's relatively straightforward nature, some systems struggle with providing the appropriate answer. One can hypothesize it is the result of limited

context (the answer is located at the end of the document), i.e., models either hallucinate a value or provide a name found earlier within the document.



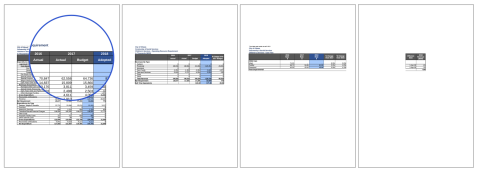
Source	Answer	ANLS	Conf.
Ground truth	Jack N. Lightstone		
Human	Jack N. Lightstone	1.0	—
T5-base	James L. Turk	0.0	0.0
MMT5	james l. turk	0.0	1.0
UDOP+BLIP2+GPT	jack n. lightstone	1.0	0.9
HiVT5+modules	Jack N. Whiteside	0.6	0.6

Requires graphical comprehension. *Which is the basis for jurisdiction?* To provide a valid answer, the model needs to comprehend the meaning of the form field and recognize the selected checkbox. None of the participating systems was able to spot the answer correctly.



Source	Answer	ANLS	Conf.
Ground truth	U.S. Goverment Plaintiff		
Human	U.S. Goverment Plaintiff	1.0	—
T5-base	Declaration of taking	0.0	0.1
MMT5	united states district court	0.0	1.0
HiVT5+modules		0.0	1.0
UDOP+BLIP2+GPT	public purpose	0.0	0.4

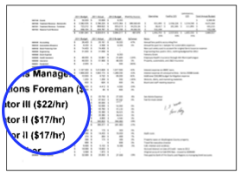
Requires comparison. *In which year does the Net Requirement exceed 25,000?* The question requires comprehending a multipage table and spotting if any values fulfill the posed condition. Some of the models resort to plausible answers (one of the three dates that the document covers), whereas others correctly decide there is no value exceeding the provided amount.



Source	Answer	ANLS	Conf.
Ground truth	[*Unanswerable*]		
Human	[*Unanswerable*]	1.0	—
T5-base	[*Unanswerable*]	1.0	0.2
MMT5	2018	0.0	1.0
UDOP+BLIP2+GPT	[*Unanswerable*]	1.0	1.0
HiVT5+modules	2017	0.0	0.8

Requires arithmetic. *What is the difference between how much Operator II*

and Operator III make per hour? The question requires table comprehension, determining relevant values, and dividing extracted integers. None of the participating models was able to fulfill this requirement.



Source	Answer	ANLS	Conf.
Ground truth	$5		
Human	$5	1.0	—
T5-base	$0.00	0.0	0.0
MMT5	65%	0.0	1.0
UDOP+BLIP2+GPT	-1.5 mile	0.0	0.0
HiVT5+modules	$5,700.00	0.0	0.4

Requires counting and list output. *What are the first two behavioral and intellectual disabilities of people with FASDs?* It seems most of the models correctly recognized that this type of question requires a list answer but either failed to comprehend the question or provided a list with incorrect length (incomplete or with too many values).

Source	Answer	ANLS	Conf.
Ground truth	Learning disabilities | Hyperactivity		
Human	learning disabilities	0.5	—
T5-base	Early embryo brain development | External Genitals	0.0	0.0
MMT5	heart beats | difficulty with attention | lung function | hyperactivity | problem with judgment | speech and language delays	0.2	1.0
UDOP+BLIP2+GPT	hyperactivity | speech and language delays	0.5	0.2
HiVT5+modules	HIV/AIDS	0.0	0.6

Appendix D

Appendix - KDD

A Code and Datasets

The proposed KD-VDU experimentation framework is available at https://github.com/Jordy-VL/DistilDoc_ICDAR24/tree/main/src. This includes the DIC benchmarking that is made fully compatible with HuggingFace *transformers*, even allowing arbitrary image classification models and (document) image datasets from HuggingFace *hub*.
The DLA benchmark is built around the *Detectron2* framework, with additional scripts for efficiency evaluation, visualization, and document data preparation for downstream tasks. Downstream task experiments are made available as a fork of the original LATIN-prompt [482] implementations with additional modifications (4-bit quantization, question type ANLS evaluation, InfographicsVQA dataloader, structure-preserving OCR respecting DLA tokens).

B Implementation Details

DIC All runs are documented with hyperparameter configuration and commandline arguments in a *wandb* project for complete transparency in experiment results and reproducibility.

For *RVL-CDIP*, both teacher and student training is carried out for 10 epochs with a batch size of (32 ViT, 64 ResNet) and AdamW with weight decay 5e-4 and a learning rate of 1e-4 with a linear warmup of 10%. For *Tobacco-3482*,

the default recipe is similarly trained for 100 epochs. All experiments were performed on a single NVIDIA GeForce RTX 3090 GPU (24GB GPU vRAM). For some feature-based KD methods, the batch size was necessarily lowered to 16 due to memory constraints. KD method hyperparameters were cross-validated to find the best performing configuration for each method, and are listed in the main manuscript result tables.

DLA In this paper, MaskRCNN detection architecture is considered with two different backbones (1) CNNs: ResNet50 and ResNet101 (2) Transformers: ViT base and ViT tiny. All the detection models are trained with Detectron2 [499] which uses the PyTorch deep learning library. The hyperparameters used are the following: (a) learning rate of 1e-4 (b) iterations 300k (c) optimizer: Adam (d) batch size: 16 (e) ROI heads predictions: 128 (f) NMS threshold: 0.4 (g) confidence threshold: 0.6 For reproducibility, we share the exact config files used for each experiment as part of the Supplementary,

Teacher and student model variants Tables D.1 and D.2 indicate the differences between used teacher and student models in terms of parameterization and efficiency.

Table D.1. Details of Vision Transformer model variants [101].

Variants	Settings of D/ViT				
	Layers	Width	FFN	Heads	#Param
Tiny (T)	12	192	768	3	5.5M
Small (S)	12	384	1536	6	21.7M
Base (B)	12	768	3072	12	85.8M

Table D.2. Details of the efficiency of model checkpoints considered in this work.

Model	**GFLOPs**	**GMACs**	**Params (M)**
microsoft/resnet-101	15.65	7.8	42.5
microsoft/resnet-50	8.21	4.09	23.51
google/vit-base-patch16-224	35.15	17.56	86.39
microsoft/dit-base	35.15	17.56	85.81
WinKawaks/vit-small-patch16-224	9.21	4.6	21.81
WinKawaks/vit-tiny-patch16-224	2.51	1.25	5.56

Downstream We extended the implementation of [482] to incorporate Llama-2 [452] and build a similar dataloader for InfographicsVQA [310]. To enable strict compatibility, we used the same unified OCR format, DUE [47], for all datasets.

This facilitated easy incorporation of DLA tokens into the OCR tokens without disrupting the logic behind the original layout-aware representation of document text. As it involved zero-shot evaluation, no finetuning was attempted for this task, and while it could be left for future work, we want to iterate that we sought to explore the innate ability of LLMs to ingest DLA-enriched prompts, and not the downstream task performance itself.

C Task Definitions

The definitions have been incorporated as part of the fundamentals. Here we will only point to details that are not included in the main manuscript.

To place each task in the context of document inputs, we define the following tasks and their respective inputs with common notation. We follow notation established in [470] for document page inputs.

A **page** p consists of an image $\boldsymbol{v} \in \mathbb{R}^{C \times H \times W}$ (number of channels, height, and width, respectively) with T word tokens $u = \{w_t\}_{t=1}^{T}$ organized according to a layout structure $s = \left\{\left(x_t^1, y_t^1, x_t^2, y_t^2\right)\right\}_{t=1}^{T}$, typically referred to as token bounding boxes, coming from OCR or available from a born-digital document.

DIC As a prototypical instance of classification [472] the goal is to learn an estimator $f : \mathcal{X} \to \mathcal{Y}$ using N supervised input-output pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from an unknown joint distribution $P(X, Y)$. In the context of DIC, the input space $\mathcal{X}$ is the set of all document images, and the output space $\mathcal{Y}$ is the set of all document classes (*e.g.*, *invoice, email, form, advertisement*, *etc.*). The goal is to learn a function f that maps a document image $x \in \mathcal{X}$ to a document class $y \in \mathcal{Y}$, such that $f(x) = y$. *Covariate shift* [418] occurs when the input distribution $P(X)$ changes between the training and evaluation sets, but the conditional distribution $P(Y|X)$ remains the same. Put plainly, both sets share the same document classes, yet the visual appearance, layout and content of the document images can be different. For example, RVL-CDIP [241] contains more modern documents with color, whereas all *RVL-CDIP* documents are greyscale.

DLA The task of DLA can be formulated as a function that processes a document image input and outputs structured information about its logical layout elements (*e.g.*, text blocks, headers, figures, charts, plots, tables). Let DLA(x) represent the output predictions of the DLA process as a set of tuples,

where each tuple (b_j, c_j, p_j) represents one of J detected logical layout element.

$$\text{DLA}(x) = \{(b_j, c_j, m_j)\}_{j=1}^{J} \tag{D.1}$$

For each, b_j denotes the bounding box for the j-th detected element, defined as (x_j, y_j, w_j, h_j) (in the popular COCO format). c_j is the class label for the j-th element, indicating its object category. m_j is a set of additional properties or information (metadata attributes, predicted scores, *considered optional*) associated with the j-th element, which can vary depending on the type and context of the layout components.

Zero-shot Document Visual Question Answering Given a document d and a question q, the goal of zero-shot DocVQA is to predict the answer a to the question q from the document, assuming a single document image for simplicity. Following the text-only LLM approach in [482], each document image requires to be translated to text, either from OCR or from a born-digital document, and the question is translated to a prompt p. The prompt $\boldsymbol{p}$ is a sequence of tokens that is fed to the LLM model, together with a potential task instruction, and the document image text D, which is structured following a heuristic procedure operating on the text tokens (T) and respective bounding boxes (see Table 6.2).

D Additional Experiment Results

Table D.3. Results of different KD strategies benchmarked for ResNets applied on the *RVL-CDIP* dataset.

Dataset	Teacher	Student	Method	ACC	AURC	ECE
RVL-CDIP	ResNet-101	–	Baseline	0.819	0.043	0.017
	–	ResNet-50	Baseline	0.783	0.059	0.039
RVL-CDIP$_{1k}$	ResNet-101	*ResNet-50*	Vanilla [$\tau = 2.5, \alpha = 0.5$]	0.783	0.059	0.039
RVL-CDIP$_{1k}$	ResNet-101		NKD [$\tau = 1, \gamma = 1.5$]	0.785	0.063	0.073
RVL-CDIP$_{1k}$	ResNet-101		MSE	0.786	0.058	0.032
RVL-CDIP$_{1k}$	ResNet-101		SimKD [$\varnothing$ projector]	0.769	0.067	0.025
RVL-CDIP$_{1k}$	ResNet-101		SimKD [CNN]	**0.797**	**0.053**	**0.023**
RVL-CDIP$_{1k}$	ResNet-101		FitNet [middle]	0.758	0.087	0.178

Table D.4. Results of different KD strategies benchmarked for ResNets applied on the *Tobacco-3482* dataset.

Student	Method	ACC	ECE	AURC
–	Teacher	0.445	0.102	0.360
ResNet-50	CE	0.552	0.096	0.256
	CE+KD	0.667	0.127	0.149
	NKD	0.436	0.076	0.330
	MSE	0.399	0.083	0.379
	SimKD [CLS+MLP]	0.176	0.250	0.768
	SimKD [CNN]	0.314	0.103	0.429
	FitNet	0.577	0.085	0.219

Table D.5. Results of different KD strategies benchmarked for ViT-B applied on the *Tobacco-3482* datasets.

Student	Method	ACC	ECE	AURC
	Teacher	0.876	0.082	0.040
ViT-S	CE	0.783	0.096	0.071
	CE+KD	0.814	0.072	0.063
	NKD	0.803	0.094	0.066
	MSE	0.807	0.161	0.062
	SimKD [CNN]	0.836	0.125	0.072
	FitNet	0.821	0.151	0.059
ViT-T	NKD	0.792	0.064	0.069
	MSE	0.798	0.198	0.074
	SimKD [CLS+MLP]	0.811	0.599	0.065
	SimKD [CNN]	0.810	0.135	0.081
	FitNet	0.805	0.160	0.070

Table D.6. Results of different KD strategies benchmarked for DiT-B applied on the *Tobacco-3482* dataset.

Student	Method	ACC	ECE	AURC
ViT-S	Teacher	0.916	0.109	0.020
	CE	0.820	0.081	0.059
	CE+KD	0.825	0.086	0.064
	NKD	0.813	0.101	0.055
	MSE	0.818	0.090	0.063
	SimKD [CLS+MLP]	0.829	0.153	0.056
	SimKD [CNN]	0.810	0.144	0.062
	FitNet	0.827	0.152	0.067
ViT-T	CE	0.810	0.066	0.065
	CE+KD	0.816	0.078	0.065
	NKD	0.807	0.087	0.063
	MSE	0.811	0.072	0.061
	SimKD [CLS+MLP]	0.778	0.162	0.093
	SimKD [CNN]	0.783	0.187	0.079
	FitNet	0.793	0.168	0.077

Table D.7. Results for DLA-KD experiments on *PRImA* dataset.

Teacher	**Student**	**Method**	**mAP**
Vit-B	-	Teacher	36.01
Resnet-101	-	Teacher	38.34
-	ViT-T	Baseline	32.64
-	Resnet-50	Baseline	35.61
Resnet-101	Resnet-50	SimKD	35.00
		ReviewKD	34.31
Vit-B	ViT-T	SimKD	32.05
		ReviewKD	31.94

D.1 *Tobacco-3482* Results

D.2 *PRImA* Results

D.3 RVL-CDIP-N Results

D.4 Downstream DocVQA Results

D.5 Ablation Experiments

The experiments with random student weight initialization (Tables D.12 and D.13) show that ViTs suffer more from student weight initialization, which is

Table D.8. Evaluation including relative runtime of KD methods on *RVL-CDIP-N*, where from left-to-right results are grouped per KD strategy, per backbone, per student size.

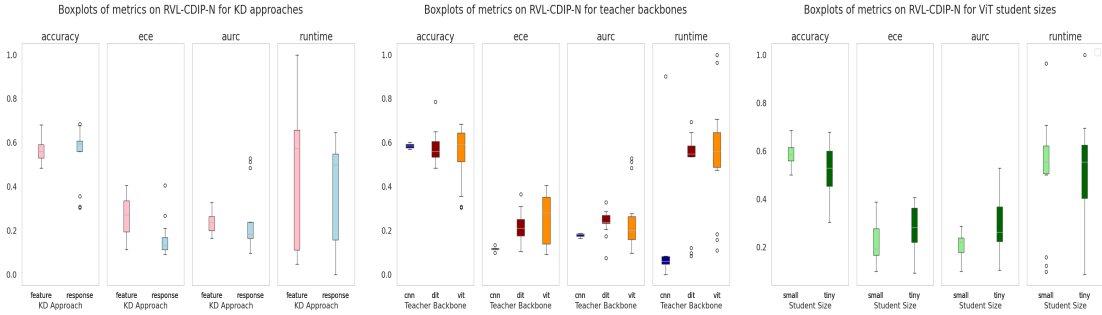


Table D.9. Results for KD methods when averaged over architectures and student sizes on *RVL-CDIP-N*.

KD method	ACC	ECE	AURC
Teacher	0.611	0.120	0.152
CE	0.573	0.119	0.215
CE+KD	0.519	0.184	0.298
NKD	0.524	**0.137**	0.259
MSE	0.490	0.205	0.308
SimKD [CLS+MLP]	0.613	0.202	0.216
SimKD [CNN]	**0.629**	0.273	**0.197**
FitNet	0.534	0.281	0.246

evidenced by an average accuracy of 0.5962 for ViT-S/T_{rand} compared to 0.7675 for $R50_{rand}$. When the student initialization is not dependent on pretraining, NKD pops up as a performant method, showing the versatility of response-based methods when transfer of feature representations is harder.

Table D.10. Validation ANLS (scaled to %) of LLAMA-2-7B-CHAT [452] on SP-DocVQA [309], with a KD-DLA model enriching the prompt.

prompt	DLA	ANLS	Image/Photo	Yes/No	Figure/diagram	Form	Free_text	Handwritten	Layout	Others	Table/list
plain		4.3	4.25	5.36	1.46	2.69	8.99	1.74	6.1	7.72	1.87
space		4.61	2.97	0.0	1.25	3.31	7.55	2.14	6.48	8.45	2.59
task		57.63	45.38	51.52	34.97	67.88	69.71	53.19	55.51	55.78	53.81
+DLA	Resnet-101	57.76	43.31	47.02	35.01	66.84	70.03	52.27	57.16	58.77	52.22
	Resnet-101	57.55	44.44	49.4	34.0	66.99	68.64	51.97	56.52	58.23	52.64
	Resnet-50 ReviewKD	57.76	43.31	47.02	35.01	66.84	70.03	52.27	57.16	58.77	52.22
	Resnet-50 SimKD	57.53	45.45	51.52	35.28	67.39	68.73	52.23	56.71	56.5	52.2
	Vit-B	58.39	44.43	41.67	34.81	66.38	67.82	52.1	59.19	55.91	52.79
	Vit-T	58.65	44.7	50.3	36.19	67.65	68.0	52.49	59.29	57.03	52.72
	Vit-T ReviewKD	57.96	45.9	47.32	33.49	66.68	68.92	51.15	58.46	56.32	51.89
	Vit-T SimKD	58.58	45.09	49.43	34.92	67.28	70.64	52.19	58.44	57.68	52.82
task_space		62.46	42.95	49.43	40.93	71.15	70.59	55.87	61.87	61.05	58.31
+DLA	Resnet-101	61.86	41.51	48.24	40.63	71.12	69.39	54.56	61.38	58.62	57.48
	Resnet-50	62.08	39.62	49.13	42.4	71.27	70.37	54.43	61.54	59.86	57.59
	Resnet-50 ReviewKD	62.14	44.09	42.26	40.39	70.6	69.69	53.07	61.8	60.14	58.29
	Resnet-50 SimKD	61.95	43.93	44.97	40.57	71.02	70.12	54.95	61.43	60.74	57.69
	Vit-B	61.2	44.58	49.13	40.28	68.95	68.39	52.81	61.38	56.44	56.7
	Vit-T	58.65	44.7	50.3	36.19	67.65	68.0	52.49	59.29	57.03	52.72
	Vit-T ReviewKD	61.58	46.25	46.75	37.84	69.37	69.27	53.86	61.5	58.44	57.63
	Vit-T SimKD	61.46	44.79	48.24	40.25	69.55	69.95	53.15	61.0	58.18	57.05

Table D.11. Validation ANLS (scaled to %) of LLAMA-2-7B-CHAT [452] on InfographicsVQA [310], with a KD-DLA model enriching the prompt.

prompt	DLA	ANLS	Arithmetic	Comparison	Counting	Figure	Map	Multi-span	Non-extractive	Question span	Single span	Table/list	Text	Visual/layout
plain		0.81	0.0	0.0	0.23	0.42	0.0	0.93	0.12	0.64	0.98	1.0	1.93	0.47
space		0.69	0.0	0.0	0.0	0.32	0.0	0.9	0.0	0.53	0.86	1.08	1.55	0.0
task		29.08	14.15	26.94	11.35	27.52	19.1	19.79	12.79	48.44	33.79	26.17	35.24	26.39
+DLA	Resnet-50	27.94	14.1	26.21	10.28	26.19	20.25	17.7	12.28	45.14	32.7	24.79	34.3	26.96
	Resnet-101	27.86	12.12	24.96	11.35	26.32	18.82	18.32	11.93	44.81	32.62	24.51	33.89	25.94
	Resnet-50 ReviewKD	28.16	13.33	25.81	12.05	26.39	22.11	21.06	12.93	46.95	32.42	25.02	34.18	26.86
	Resnet-50 SimKD	27.65	13.79	25.78	9.95	26.16	19.53	18.78	11.97	45.95	32.17	24.31	33.8	26.31
	Vit-B	28.36	14.93	29.15	7.64	27.05	19.0	19.41	11.21	46.87	33.35	25.56	34.59	26.69
	Vit-T	28.32	15.06	28.02	9.58	27.25	19.01	17.0	11.82	45.67	33.48	25.02	34.81	28.33
	Vit-T ReviewKD	28.23	13.35	27.7	10.78	26.39	20.03	20.4	11.92	45.95	32.95	25.9	35.28	27.46
	Vit-T SimKD	28.18	14.82	26.31	9.6	26.19	18.96	18.09	12.51	45.36	32.87	24.93	34.71	30.98
task+space		27.97	9.78	25.13	6.99	25.93	21.04	22.33	8.2	43.36	33.53	25.76	35.06	27.47
+DLA	Resnet-50	27.14	8.12	23.78	6.27	24.68	18.67	19.26	7.0	41.95	33.03	25.93	34.07	28.48
	Resnet-101	28.08	9.49	24.31	8.04	25.88	19.72	21.01	8.63	41.23	33.77	25.87	35.24	28.44
	Resnet-50 ReviewKD	28.07	9.59	24.18	8.41	25.88	18.67	21.37	9.01	42.86	33.53	26.2	35.49	27.8
	Resnet-50 SimKD	27.68	9.98	24.45	7.11	25.71	20.65	20.87	8.4	43.36	33.19	25.51	34.56	27.81
	Vit-B	28.05	9.92	25.28	7.83	26.28	19.0	21.85	8.82	41.84	33.54	25.57	34.6	29.17
	Vit-T	27.0	9.06	23.19	7.34	25.81	21.9	18.9	8.04	39.82	32.65	23.69	33.93	28.33
	Vit-T ReviewKD	28.47	10.89	25.9	5.42	26.8	22.23	20.59	8.28	45.67	34.24	26.44	35.81	29.14
	Vit-T SimKD	27.97	10.56	25.54	8.35	26.23	20.65	20.34	9.19	44.08	33.43	25.04	33.89	30.49

Table D.12. Results of different KD strategies benchmarked for ViT-B teacher with **randomly** initialized (rand) ViT students applied on the *RVL-CDIP* dataset.

Teacher	Student	Method	ACC	AURC	ECE
ViT-B_rand	–	Baseline	0.540	0.235	0.078
–	ViT-S_{rand}	Vanilla [$\tau = 2.5, \alpha = 0.5$]	0.613	0.175	0.220
ViT-B		NKD [$\tau = 1, \gamma = 1.5$]	0.579	0.193	**0.046**
ViT-B		MSE	0.626	0.159	0.203
ViT-B		SimKD [CLS+MLP]	0.609	0.181	0.120
ViT-B		SimKD [CNN]	**0.681**	0.181	0.297
ViT-B		FitNet [middle]	0.628	**0.161**	0.155
ViT-B	ViT-T_{rand}	Vanilla [$\tau = 2.5, \alpha =$]	0.560	0.212	0.141
ViT-B		NKD [$\tau = 1, \gamma = 1.5$]	0.552	0.215	**0.025**
ViT-B		MSE	0.579	**0.198**	0.232
ViT-B		SimKD [CLS+MLP]	0.582	0.199	0.196
ViT-B		SimKD [CNN]	**0.663**	0.205	0.316
ViT-B		FitNet [middle]	0.570	0.207	0.143

Table D.13. Results of different KD strategies benchmarked for ResNet-101 teacher with **randomly** initialized (rand) ResNet-50 students applied on the *RVL-CDIP* dataset.

Teacher	Student	Method	ACC	AURC	ECE
R101_rand	–	Baseline			
–	R50	Baseline	0.769	0.015	0.066
R101	**R50**$_{\text{rand}}$	Vanilla [$\tau = 2.5, \alpha = 0.5$]	0.760	**0.017**	0.071
R101		NKD [$\tau = 1, \gamma = 1.5$]	0.770	0.051	0.072
R101		MSE	0.765	0.022	0.068
R101		SimKD [CLS+MLP]	0.766	0.037	0.068
R101		SimKD [$\varnothing$ projector]	**0.774**	0.025	**0.063**
R101		FitNet [middle]	0.760	0.177	0.078

Curriculum

JORDY VAN LANDEGHEM received an M.A. degree in Linguistics in 2015 and an M.Sc. degree in artificial intelligence in 2017, both from KU Leuven, where he is currently pursuing a Ph.D. degree in computer science. He completed research internships at Oracle and Nuance Communications, and is currently the lead AI Researcher at Contract.fit, a European SaaS start-up building intelligent document processing solutions.

His industrial Ph.D. project entitled "Intelligent Automation for AI-Driven Document Understanding" focuses on the fundamentals of probabilistic deep learning, emphasizing calibration, uncertainty quantification, and out-of-distribution robustness to obtain more reliable document intelligence systems. Recently, he spearheaded the Document UnderstanDing of Everything (DUDE) project and the ensuing ICDAR 2023 competition, with more research published on reliable and scalable document understanding.

Publications

Journal Articles

Sumam Francis, Jordy Van Landeghem, and Marie-Francine Moens. Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents. *Information*, 10(8):248, 2019

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Benchmarking Scalable Predictive Uncertainty in Text Classification. *IEEE Access*, 2022

Peer-reviewed International Conference and Workshop Articles

Jordy Van Landeghem, Matthew B Blaschko, Bertrand Anckaert, and Marie-Francine Moens. Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *International Conference on Document Analysis and Recognition*, pages 420–434. Springer, 2023 ***Oral Presentation**

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Marie-Francine Moens, and Tomasz Stanisławek. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023

Jordy Van Landeghem, Sanket Biswas, Matthew Blaschko, and Marie-Francine Moens. Beyond Document Page Classification: Design, Datasets, and Challenges. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2962–2972, 2024 ***Oral Presentation**

Jordy Van Landeghem, Subhajit Maity, Ayan Banerjee, Matthew B Blaschko, Marie-Francine Moens, Josep Llados, and Sanket Biswas. DistilDoc: Knowledge Distillation for Visually-Rich Document Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (under review)*, 2024

Organized Competitions

ICDAR2023 Competition on Document UnderstanDing of Everything (DUDE), ICDAR, February-May, 2023, https://rrc.cvc.uab.es/?ch=23, Main organizer.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
LANGUAGE INTELLIGENCE & INFORMATION RETRIEVAL LAB
Celestijnenlaan 200A box 2402
B-3001 Leuven
jordy.vanlandeghem@cs.kuleuven.be
https://liir.cs.kuleuven.be/