Microsoft

ICDAR
LAUSANNE 2021

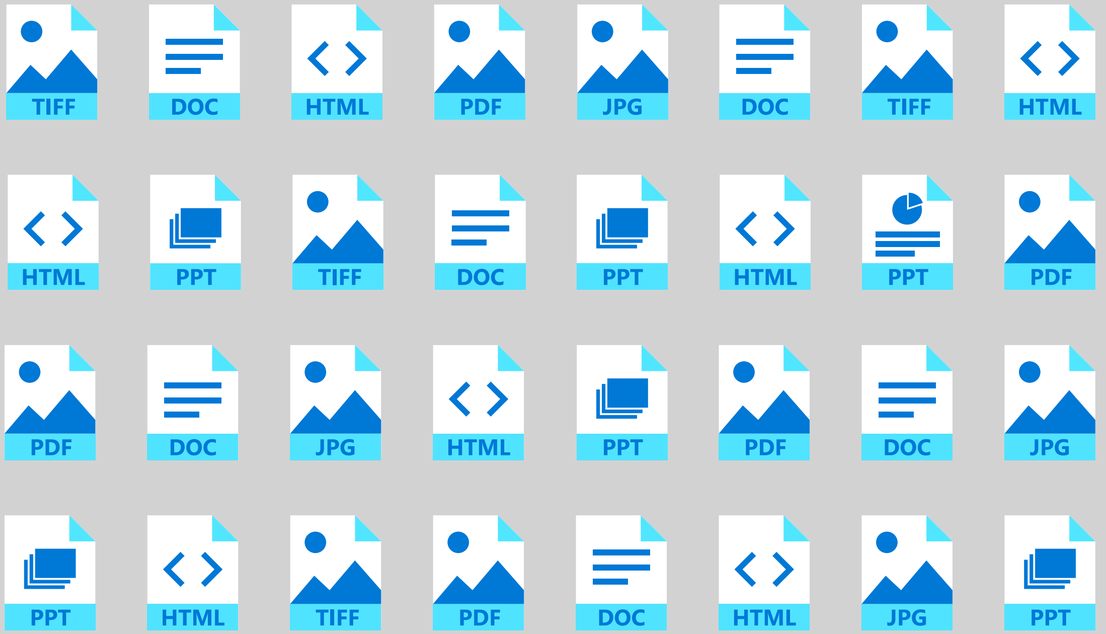
Document AI: Benchmarks, Models and Applications

Lei Cui
Microsoft Research Asia

DIL workshop, ICDAR 2021
2021-09-06

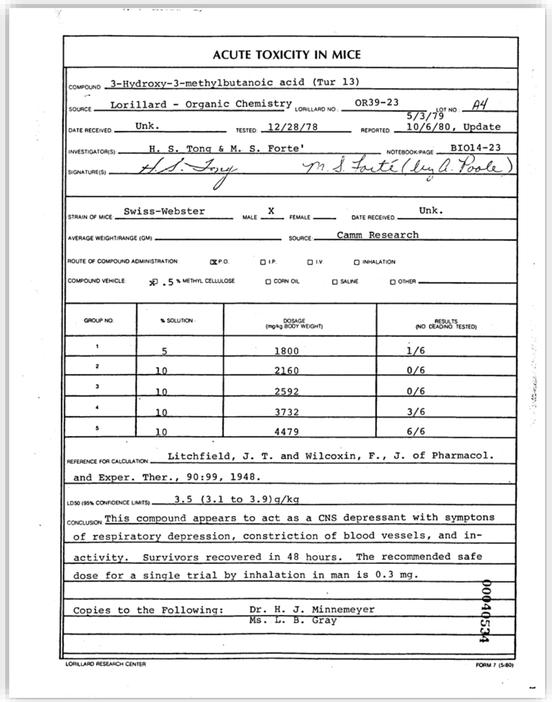
Document AI

- Analyze forms and documents
- Create intelligent search indexes
- Automate business workflows



Uncover latent insights from all your content

Document AI in Real World



ACUTE TOXICITY IN MICE

COMPOUND 3-Hydroxy-3-methylbutanoic acid (Tur 13)

SOURCE Lorillard - Organic Chemistry LORILLARD NO. OR39-23 LIST NO. A4

DATE RECEIVED Unk. TESTED 12/28/78 5/3/79 REPORTED 10/6/80, Update

INVESTIGATOR(S) H. S. Tong & M. S. Forte' NOTEBOOK/PAGE BIO14-23

SIGNATURE(S) H. S. Tong M. S. Forte (by G. Poole)

STRAIN OF MICE Swiss-Webster MALE X FEMALE DATE RECEIVED Unk.

AVERAGE WEIGHT/RANGE (GM) SOURCE Camm Research

ROUTE OF COMPOUND ADMINISTRATION ☒ P.O. ☐ I.P. ☐ I.V. ☐ INHALATION

COMPOUND VEHICLE ☒ .5 % METHYL CELLULOSE ☐ CORN OIL ☐ SALINE ☐ OTHER

GROUP NO.	% SOLUTION	DOSAGE (mg/kg BODY WEIGHT)	RESULTS (NO. DEADING TESTED)
1	5	1800	1/6
2	10	2160	0/6
3	10	2592	0/6
4	10	3732	3/6
5	10	4479	6/6

REFERENCE FOR CALCULATION Litchfield, J. T. and Wilcoxin, F., J. of Pharmacol. and Exper. Ther., 90:99, 1948.

LD50 (95% CONFIDENCE LIMITS) 3.5 (3.1 to 3.9)g/kg

CONCLUSION This compound appears to act as a CNS depressant with symptons of respiratory depression, constriction of blood vessels, and inactivity. Survivors recovered in 48 hours. The recommended safe dose for a single trial by inhalation in man is 0.3 mg.

Copies to the Following: Dr. H. J. Minnemeyer
Ms. L. B. Gray

00045524

LORILLARD RESEARCH CENTER FORM 7 (5-80)

Morton's The Steakhouse
735 S Figueroa Street
Los Angeles, CA 90017
(213) 553-4566

Server: Sally DOB: 09/08/2016
06:23 PM 09/08/2016
404/1 1/10084

SALE

Visa 6291458
Card #XXXXXXXXXXXX8698
Magnetic card present: SICKAFOOSE/DANNY
Card Entry Method: S

Approval: 04702B

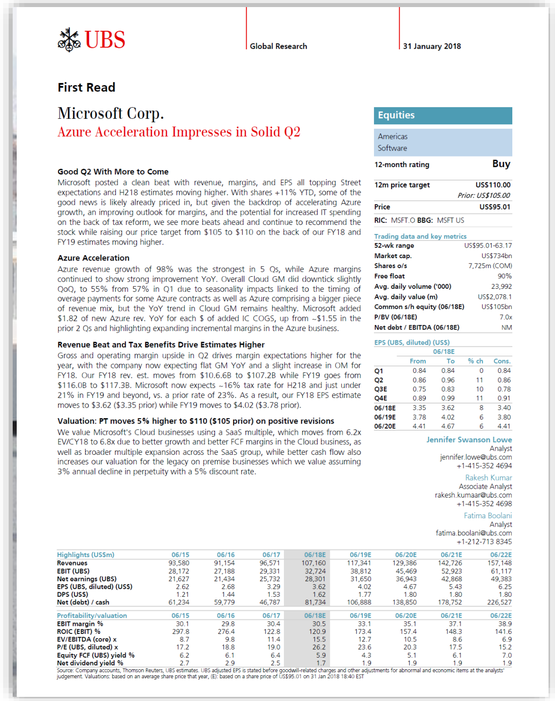
Amount: $33.79
+ Gratuity Not Inc: 16
= Balance Due: 39.79

I agree to pay the above total amount according to the card issuer agreement.

X

For banquet events, balance due includes suggested gratuity if accepted.

Guest Copy



UBS Global Research 31 January 2018

First Read

Microsoft Corp.

Azure Acceleration Impresses in Solid Q2

Equities
Americas
Software

12-month rating Buy

12m price target US$110.00
Prior: US$105.00
Price US$95.01

RIC: MSFT.O BBG: MSFT US

Trading data and key metrics

52-wk range	US$95.01-63.17
Market cap.	US$734bn
Shares o/s	7,725m (COM)
Free float	90%
Avg. daily volume ('000)	23,992
Avg. daily value (m)	US$2,078.1
Common s/h equity (06/18E)	US$105bn
P/BV (06/18E)	7.0x
Net debt / EBITDA (06/18E)	NM

EPS (UBS, diluted) (US$)

	06/18E From	To	% ch	Cons.
Q1	0.84	0.84	0	0.84
Q2	0.86	0.96	11	0.86
Q3E	0.75	0.83	10	0.78
Q4E	0.89	0.99	11	0.91
06/18E	3.35	3.62	8	3.40
06/19E	3.78	4.02	6	3.80
06/20E	4.41	4.67	6	4.41

Good Q2 With More to Come

Microsoft posted a clean beat with revenue, margins, and EPS all topping Street expectations and H218 estimates moving higher. With shares +11% YTD, some of the good news is likely already priced in, but given the backdrop of accelerating Azure growth, an improving outlook for margins, and the potential for increased IT spending on the back of tax reform, we see more beats ahead and continue to recommend the stock while raising our price target from $105 to $110 on the back of our FY18 and FY19 estimates moving higher.

Azure Acceleration

Azure revenue growth of 98% was the strongest in 5 Qs, while Azure margins continued to show strong improvement YoY. Overall Cloud GM did downtick slightly QoQ, to 55% from 57% in Q1 due to seasonality impacts linked to the timing of overage payments for some Azure contracts as well as Azure comprising a bigger piece of revenue mix, but the YoY trend in Cloud GM remains healthy. Microsoft added $1.82 of new Azure rev. YoY for each $ of added IC COGS, up from ~$1.55 in the prior 2 Qs and highlighting expanding incremental margins in the Azure business.

Revenue Beat and Tax Benefits Drive Estimates Higher

Gross and operating margin upside in Q2 drives margin expectations higher for the year, with the company now expecting flat GM YoY and a slight increase in OM for FY18. Our FY18 rev. est. moves from $10.6-6B to $107.2B while FY19 goes from $116.0B to $117.3B. Microsoft now expects ~16% tax rate for H218 and just under 21% in FY19 and beyond, vs. a prior rate of 23%. As a result, our FY18 EPS estimate moves to $3.62 ($3.35 prior) while FY19 moves to $4.02 ($3.78 prior).

Valuation: PT moves 5% higher to $110 ($105 prior) on positive revisions

We value Microsoft's Cloud businesses using a SaaS multiple, which moves from 6.2x EVICY18 to 6.8x due to better growth and better FCF margins in the Cloud business, as well as broader multiple expansion across the SaaS group, while better cash flow also increases our valuation for the legacy on premise businesses which we value assuming 3% annual decline in perpetuity with a 5% discount rate.

Jennifer Swanson Lowe
Analyst
jennifer.lowe@ubs.com
+1-415-352 4694

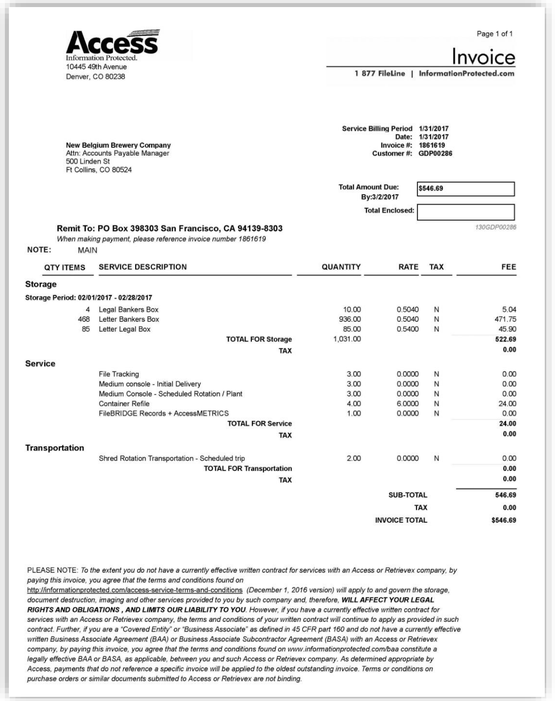
Rakesh Kumar
Associate Analyst
rakesh.kumar@ubs.com
+1-415-352 4698

Fatima Boolani
Analyst
fatima.boolani@ubs.com
+1-212-713 8345

Highlights (US$m)	06/15	06/16	06/17	06/18E	06/19E	06/20E	06/21E	06/22E
Revenues	93,580	91,154	96,571	107,160	117,341	129,386	142,726	157,148
EBIT (UBS)	28,172	27,188	29,331	32,724	38,812	45,469	52,923	61,117
Net earnings (UBS)	21,627	21,434	25,732	28,301	31,650	36,943	42,868	49,383
EPS (UBS, diluted) (US$)	2.62	2.68	3.29	3.62	4.02	4.67	5.43	6.25
DPS (US$)	1.21	1.44	1.53	1.62	1.77	1.80	1.80	1.80
Net (debt) / cash	61,234	59,779	46,787	81,734	106,888	138,850	178,752	226,527

Profitability/valuation	06/15	06/16	06/17	06/18E	06/19E	06/20E	06/21E	06/22E
EBIT margin %	30.1	29.8	30.4	30.5	33.1	35.1	37.1	38.9
ROIC (EBIT) %	297.8	276.4	122.8	120.9	173.4	157.4	148.3	141.6
EV/EBITDA (core) x	8.7	9.8	11.4	15.5	12.7	10.5	8.6	6.9
P/E (UBS, diluted) x	17.2	18.8	19.0	26.2	23.6	20.3	17.5	15.2
Equity FCF (UBS) yield %	6.2	6.1	6.4	5.9	4.3	5.1	6.1	7.0
Net dividend yield %	2.7	2.9	2.5	1.7	1.9	1.9	1.9	1.9

Source: Company accounts, Thomson Reuters, UBS estimates. UBS adjusted EPS is stated before goodwill-related charges and other adjustments for abnormal and economic items at the analysts' judgement. Valuations based on an average share price that year, (E) based on a share price of US$95.01 on 31 Jan 2018 18:40 EST



Access
Information Protected.
10445 40th Avenue
Denver, CO 80238

Page 1 of 1

Invoice

1 877 FileLine | InformationProtected.com

New Belgium Brewery Company
Attn: Accounts Payable Manager
500 Linden St
Ft Collins, CO 80524

Service Billing Period 1/31/2017
Date: 1/31/2017
Invoice #: 1861619
Customer #: GDP00286

Total Amount Due: $546.69
By:3/2/2017
Total Enclosed:

130GDP00286

Remit To: PO Box 398303 San Francisco, CA 94139-8303
When making payment, please reference invoice number 1861619

NOTE: MAIN

QTY ITEMS	SERVICE DESCRIPTION	QUANTITY	RATE	TAX	FEE
Storage					
Storage Period: 02/01/2017 - 02/28/2017					
4	Legal Bankers Box	10.00	0.5040	N	5.04
468	Letter Bankers Box	936.00	0.5040	N	471.75
85	Letter Legal Box	85.00	0.5400	N	45.90
	TOTAL FOR Storage	1,031.00			**522.69**
	TAX				**0.00**
Service					
	File Tracking	3.00	0.0000	N	0.00
	Medium console - Initial Delivery	3.00	0.0000	N	0.00
	Medium Console - Scheduled Rotation / Plant	3.00	0.0000	N	0.00
	Container Refile	4.00	6.0000	N	24.00
	FileBRIDGE Records + AccessMETRICS	1.00	0.0000	N	0.00
	TOTAL FOR Service				**24.00**
	TAX				**0.00**
Transportation					
	Shred Rotation Transportation - Scheduled trip	2.00	0.0000	N	0.00
	TOTAL FOR Transportation				**0.00**
	TAX				**0.00**
			SUB-TOTAL		**546.69**
			TAX		**0.00**
			INVOICE TOTAL		**$546.69**

PLEASE NOTE: *To the extent you do not have a currently effective written contract for services with an Access or Retrievex company, by paying this invoice, you agree that the terms and conditions found on* http://informationprotected.com/access-service-terms-and-conditions *(December 1, 2016 version) will apply to and govern the storage, document destruction, imaging and other services provided to you by such company and, therefore,* ***WILL AFFECT YOUR LEGAL RIGHTS AND OBLIGATIONS , AND LIMITS OUR LIABILITY TO YOU****. However, if you have a currently effective written contract for services with an Access or Retrievex company, the terms and conditions of your written contract will continue to apply as provided in such contract. Further, if you are a "Covered Entity" or "Business Associate" as defined in 45 CFR part 160 and do not have a currently effective written Business Associate Agreement (BAA) or Business Associate Subcontractor Agreement (BASA) with an Access or Retrievex company, by paying this invoice, you agree that the terms and conditions found on www.informationprotected.com/baa constitute a legally effective BAA or BASA, as applicable, between you and such Access or Retrievex company. As determined appropriate by Access, payments that do not reference a specific invoice will be applied to the oldest outstanding invoice. Terms or conditions on purchase orders or similar documents submitted to Access or Retrievex are not binding.*

Form

Receipt

Report

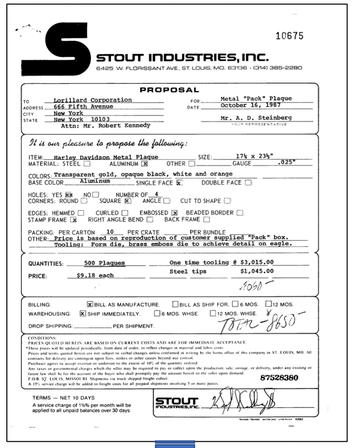
Invoice

Scanned documents (.jpg, .png, …)

Digital-born documents (.pdf, .docx, …)

Layout invariance (key-value, tabular, etc.) among visually-rich documents

Document AI Tasks



Key	Value
TO	Lorillard Corporation
ADDRESS	666 Fifth Avenue
CITY	New York
…	…

Form Understanding

Key	Value
Total	
Company	
Address	
Date	

STARBUCKS Store #10208
11302 Euclid Avenue
Cleveland, OH (216) 229-0749

CHK 664290
12/07/2014 06:43 PM
1912003 Drawer: 2 Reg: 2

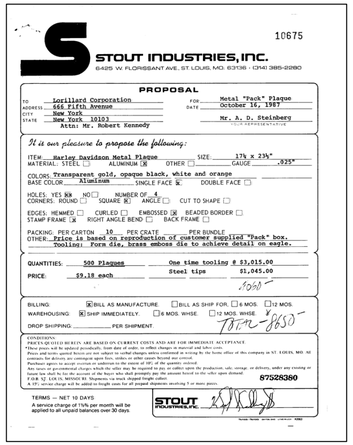
Vt Pep Mocha	4.95
Sbux Card	4.95
XXXXXXXXXXXX3228	
Subtotal	$4.95
Total	$4.95
Change Due	$0.00

Check Closed
12/07/2014 06:43 PM

SBUX Card x3228 New Balance: 37.45
Card is registered.

Key	Value
Total	4.95
Company	Starbucks Store
Address	11302 Euclid Avenue Cleveland, OH
Date	12/07/2014

Receipt Understanding



Category: Form

Document Image Classification

Document AI Products

Applications

Key-value Extraction
Document Classification
Document VQA

Downstream Tasks

Table Detection
Page Object Detection
Reading Order Detection

Benchmarks

TableBank (LREC'20)

DocBank (COLING'20)

ReadingBank (EMNLP'21)

XFUND Benchmark

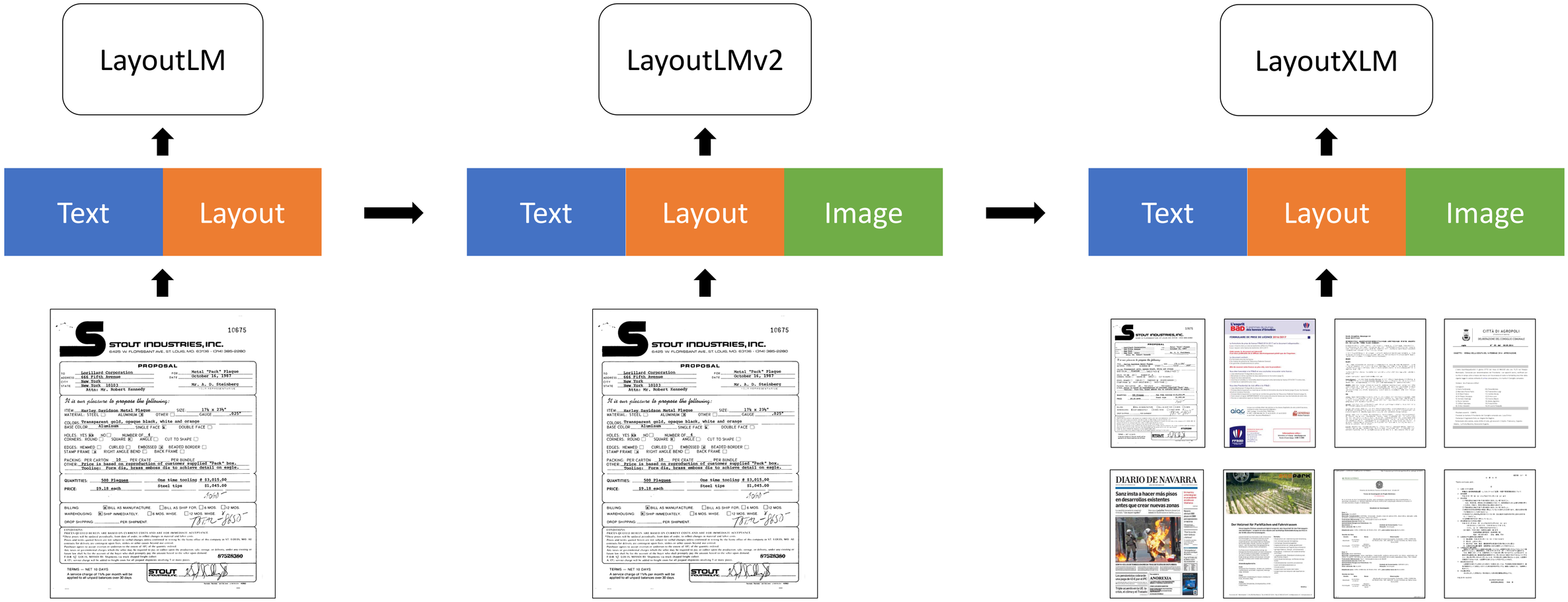
Foundation Models

LayoutLM/LayoutLMv2/LayoutXLM (KDD'20, ACL'21)

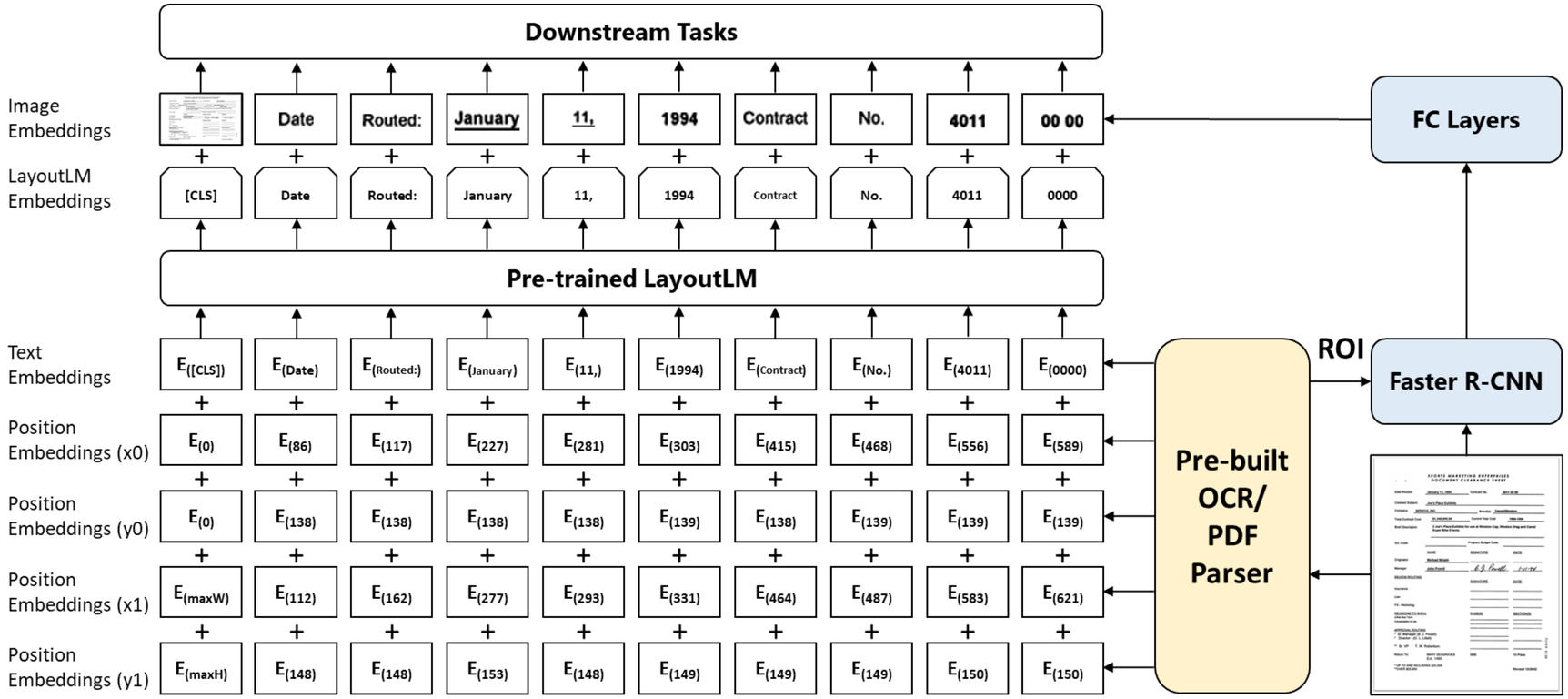
https://aka.ms/document-ai/

LayoutLM Model Family

LayoutLM -> LayoutLMv2 -> LayoutXLM



LayoutLM



* Text embeddings initialized by BERT/UniLM

LayoutLM: Pre-training of Text and Layout for Document Image Understanding, KDD'2020,

SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET

Date Routed: January 11, 1994 Contract No. 4011 00 00

Contract Subject: Joe's Place Exhibits

Company SPEVCO, INC. Brand(s) Camel/Winston

Total Contract Cost $1,340,000.00 Current Year Cost 1994-1995

Brief Description 2 Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.

G/L Code: Program Budget Code

	NAME	SIGNATURE	DATE
Originator	Michael Wright		
Manager	John Powell	B. J. Powell	1-11-94

REVIEW ROUTING

	SIGNATURE	DATE
Insurance		
Law		
FS - Marketing		

REVISIONS TO SHELL (Other than Term, Compensation or Job)	PAGE(S)	SECTION(S)

APPROVAL ROUTING
* Sr. Manager (B. J. Powell)
* Director - (G. L. Littell)

** Sr. VP T. W. Robertson

Return To: MARY SEAGRAVES Ext. 1485 SME 13 Plaza

* UP TO AND INCLUDING $25,000
**OVER $25,000

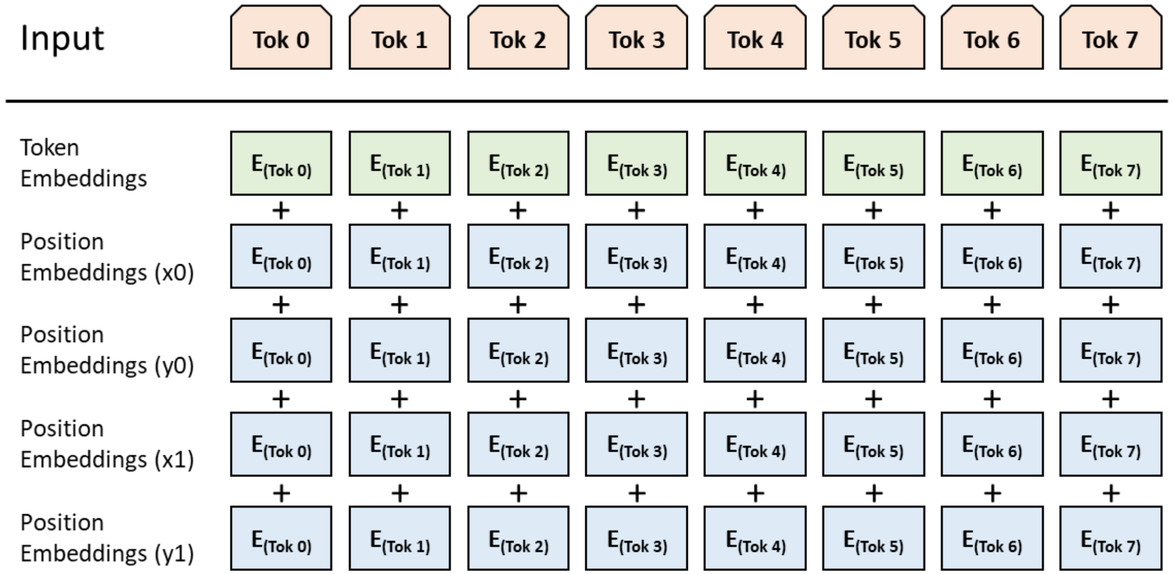
Revised 10/26/92

51669 8130

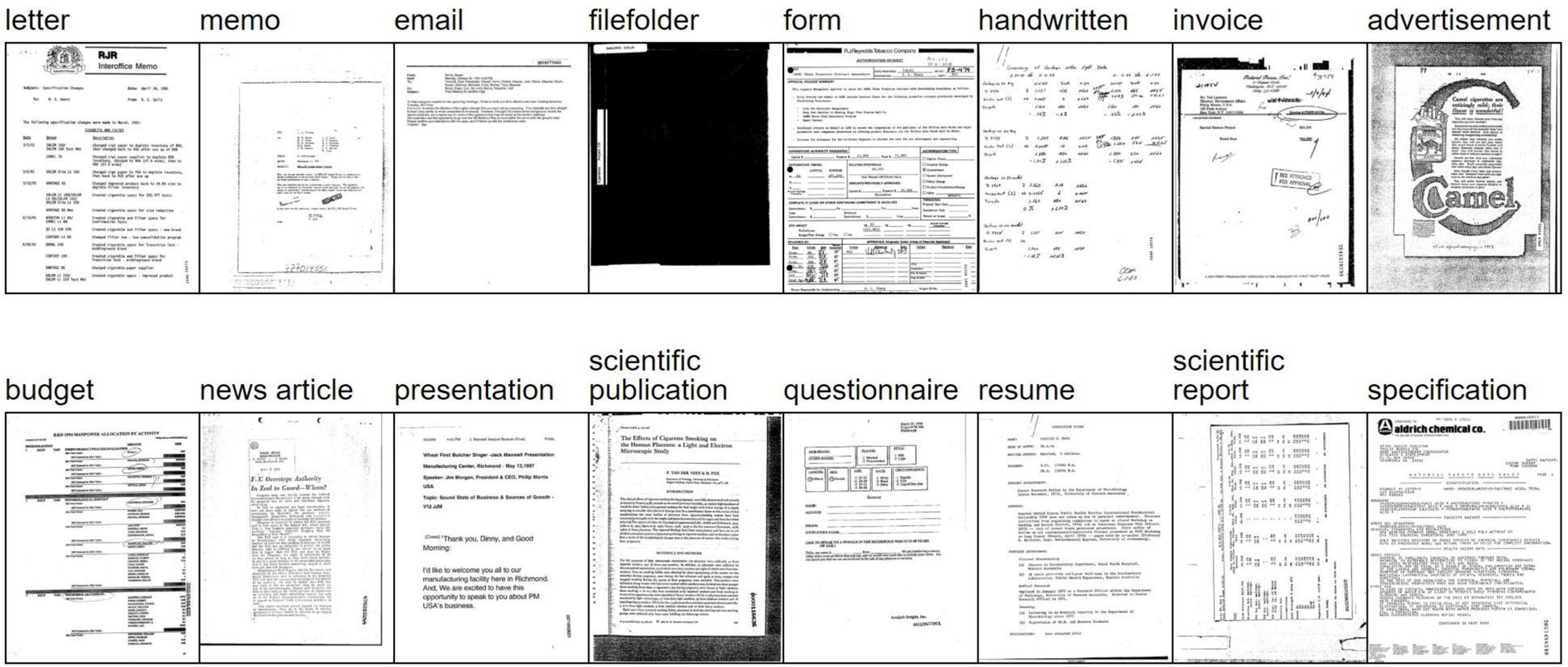
Date Routed: January 11, 1994 Contract No. 4011 00 00

OCR/PDF Parser

Image	Token	Bounding Box (x0,y0,x1,y1)
Date	Date	86 138 112 148
Routed:	Routed:	117 138 162 148
January	January	227 138 277 153
11	11,	281 138 293 148
1994	1994	303 139 331 149
Contract	Contract	415 138 464 149
No.	No.	468 139 487 149
4011	4011	556 139 583 150
00 00	0000	589 139 621 150



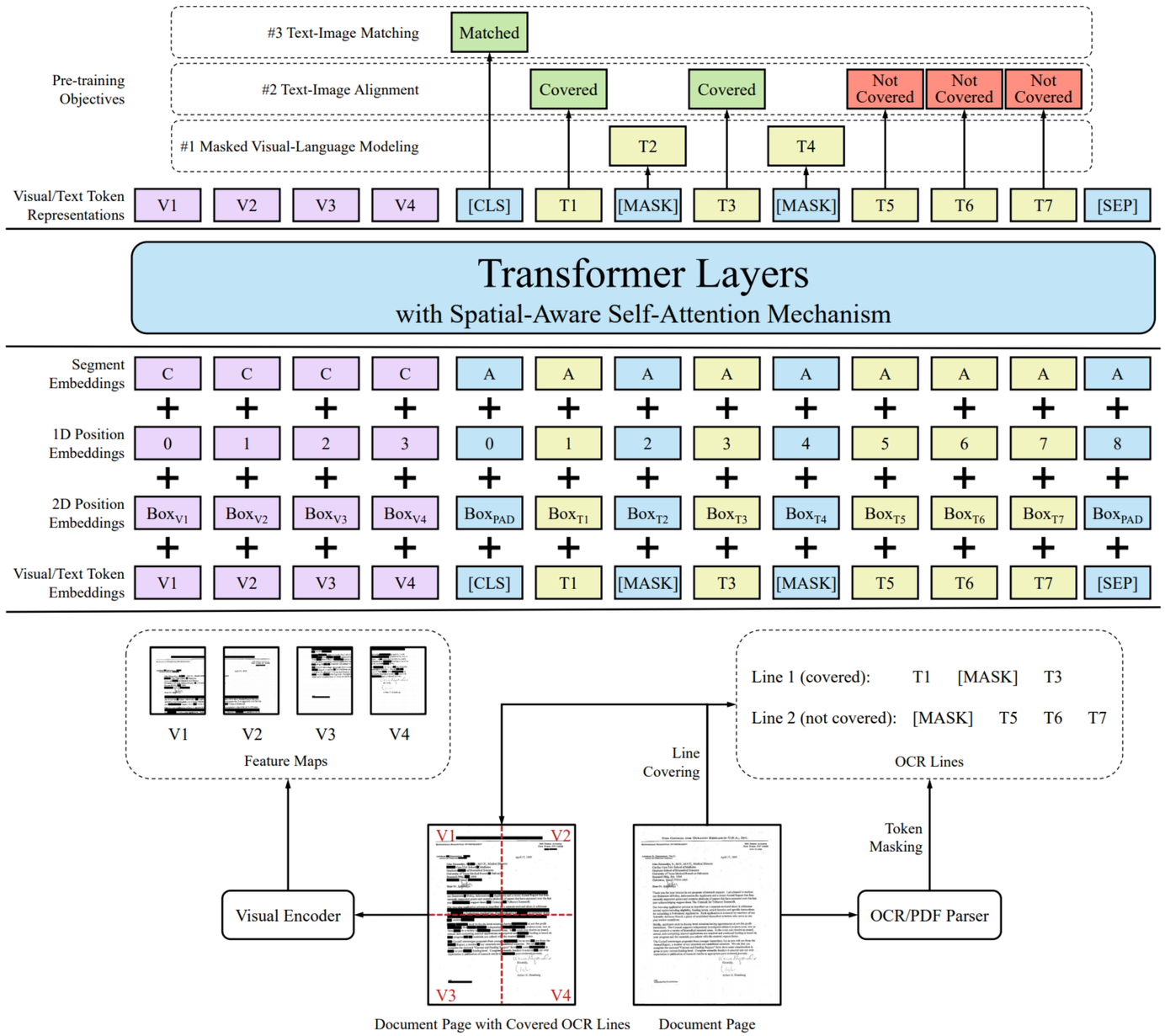
Pre-training Data



11 million scanned document images from IIT-CDIP Test Collection 1.0

https://ir.nist.gov/cdip/

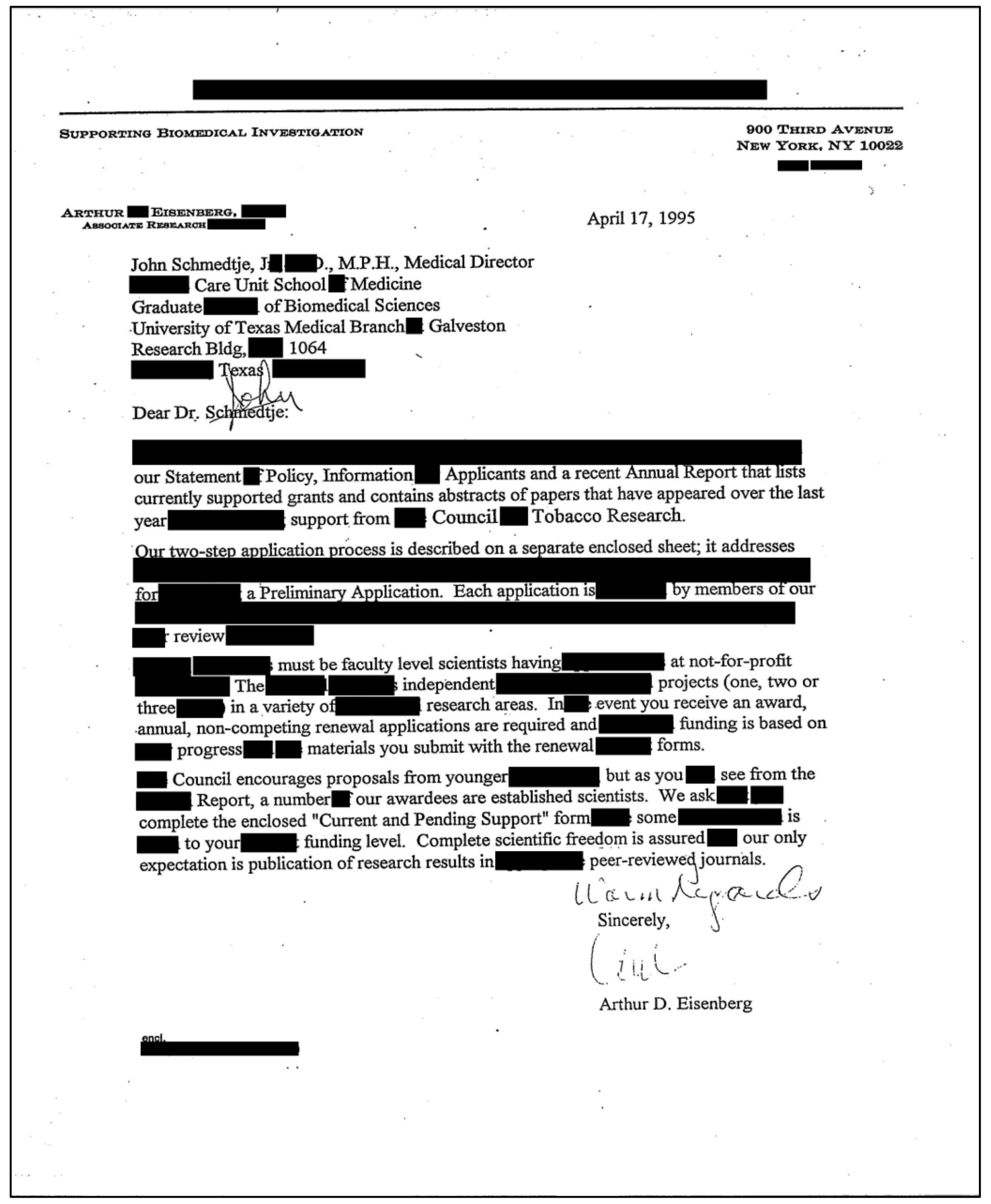
LayoutLMv2



[LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding](), ACL'2021

Pre-training Tasks

- Masked Visual-Language Modeling
- Text-Image Matching
- Text-Image Alignment



[redacted]

SUPPORTING BIOMEDICAL INVESTIGATION

900 THIRD AVENUE
NEW YORK, NY 10022
[redacted]

ARTHUR [redacted] EISENBERG, [redacted]
ASSOCIATE RESEARCH [redacted]

April 17, 1995

John Schmedtje, J[redacted]D., M.P.H., Medical Director
[redacted] Care Unit School [redacted] Medicine
Graduate [redacted] of Biomedical Sciences
University of Texas Medical Branch [redacted] Galveston
Research Bldg, [redacted] 1064
[redacted] Texas [redacted]

Dear Dr. Schmedtje:

[redacted] our Statement [redacted] Policy, Information [redacted] Applicants and a recent Annual Report that lists currently supported grants and contains abstracts of papers that have appeared over the last year [redacted] support from [redacted] Council [redacted] Tobacco Research.

Our two-step application process is described on a separate enclosed sheet; it addresses [redacted] for [redacted] a Preliminary Application. Each application is [redacted] by members of our [redacted] review [redacted]

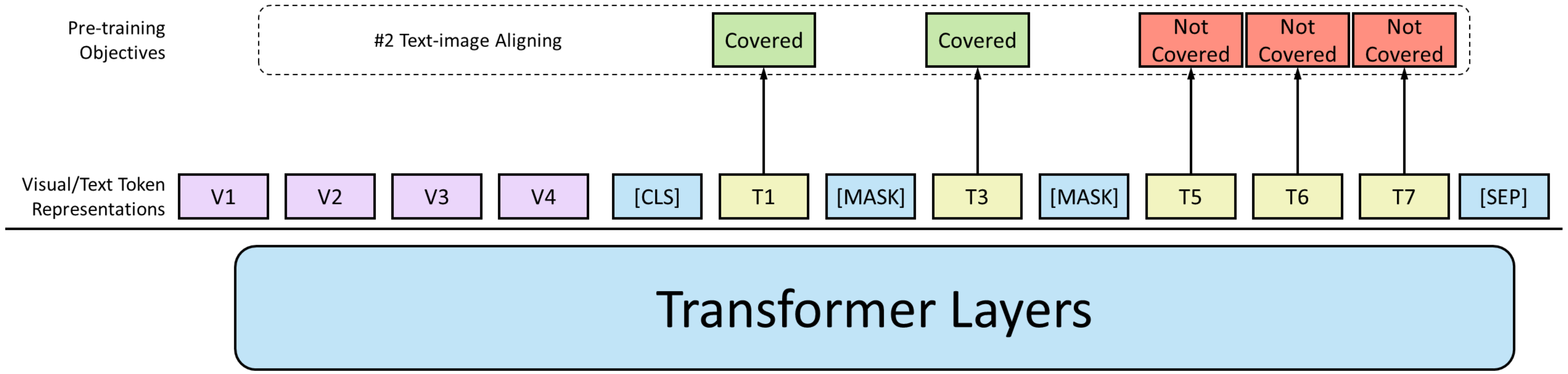
[redacted] must be faculty level scientists having [redacted] at not-for-profit [redacted] The [redacted] independent [redacted] projects (one, two or three [redacted] in a variety of [redacted] research areas. In [redacted] event you receive an award, annual, non-competing renewal applications are required and [redacted] funding is based on [redacted] progress [redacted] materials you submit with the renewal [redacted] forms.

[redacted] Council encourages proposals from younger [redacted] but as you [redacted] see from the [redacted] Report, a number [redacted] our awardees are established scientists. We ask [redacted] complete the enclosed "Current and Pending Support" form [redacted] some [redacted] is [redacted] to your [redacted] funding level. Complete scientific freedom is assured [redacted] our only expectation is publication of research results in [redacted] peer-reviewed journals.

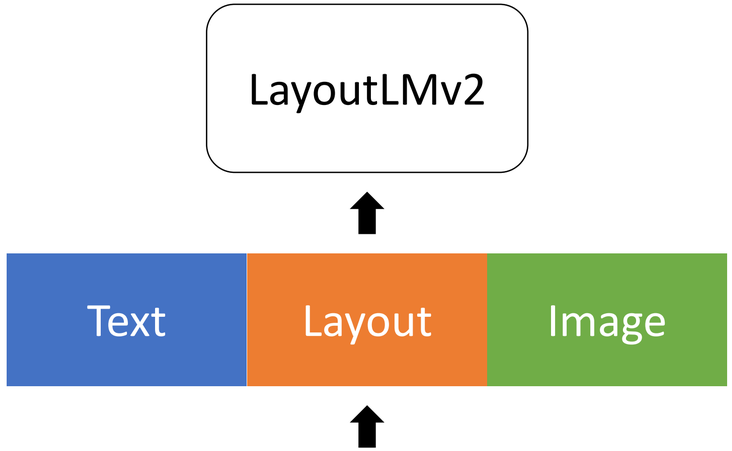
Sincerely,

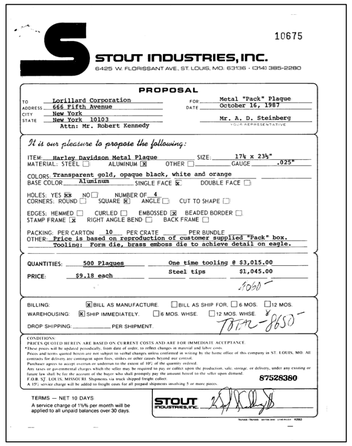
Arthur D. Eisenberg

encl.
[redacted]

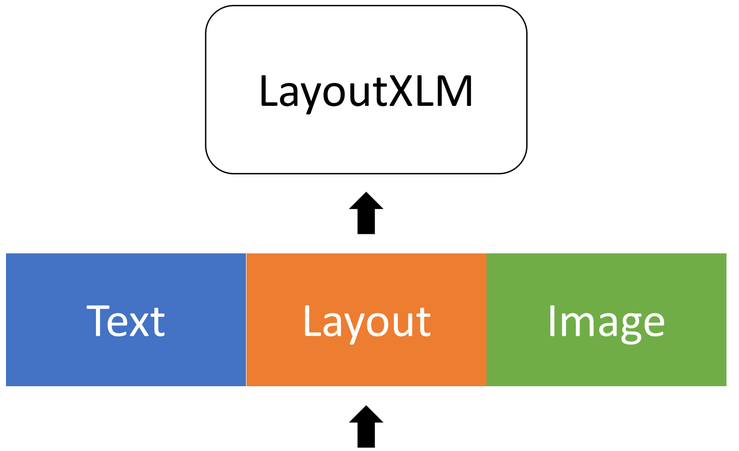


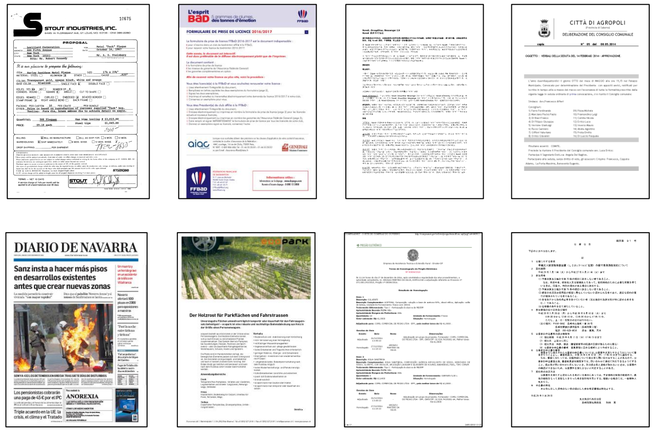
LayoutLMv2 -> LayoutXLM





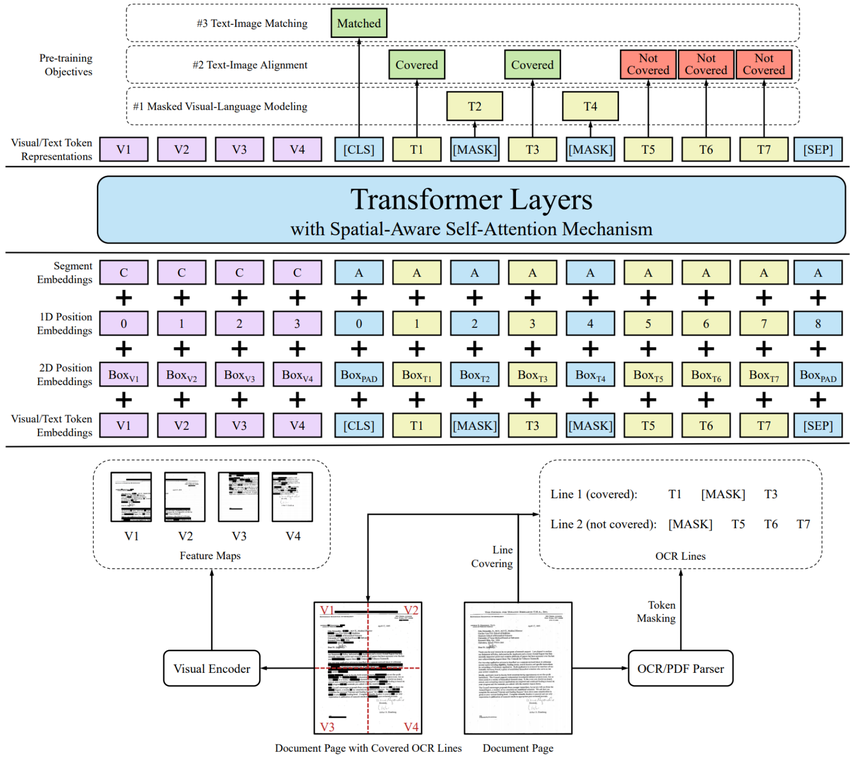
Multi-modal





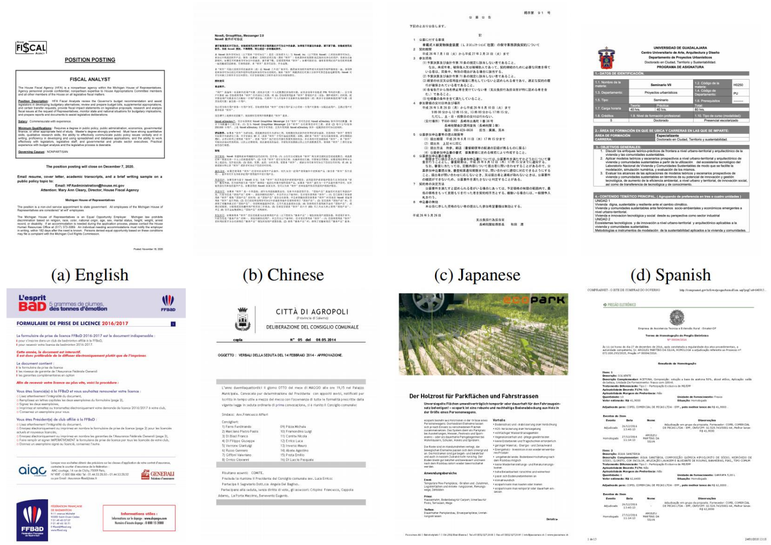
Multi-modal + Multi-lingual

LayoutXLM



LayoutLMv2 + InfoXLM

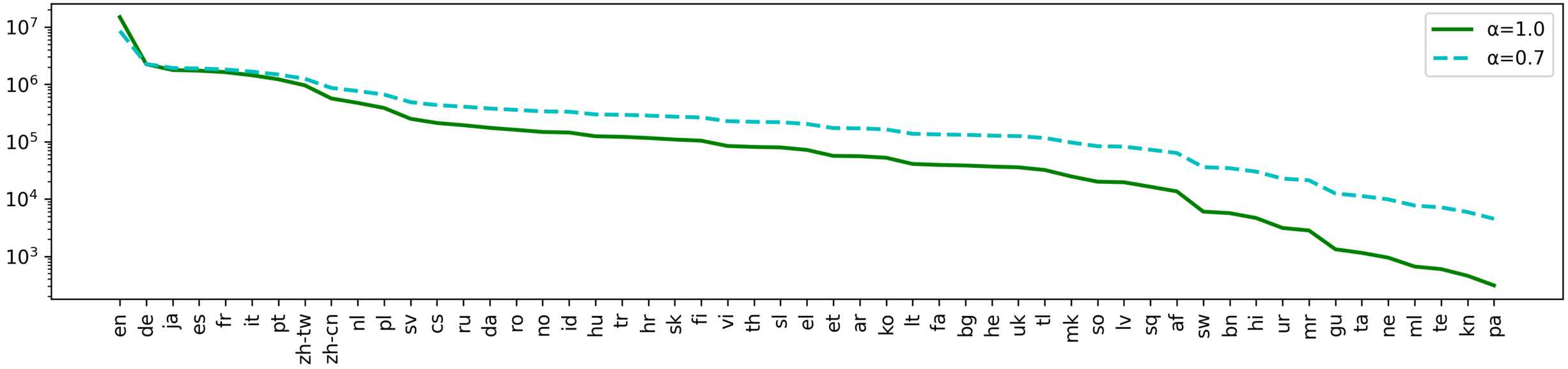
+



Multi-lingual documents (50+ languages)

[LayoutXLM: Multi-modal Pre-training for Multi-lingual Document Understanding](), Preprint

Language Distribution for Pre-training



Totally ***30M document images with 50+ languages*** used for pre-training the ***LayoutXLM***

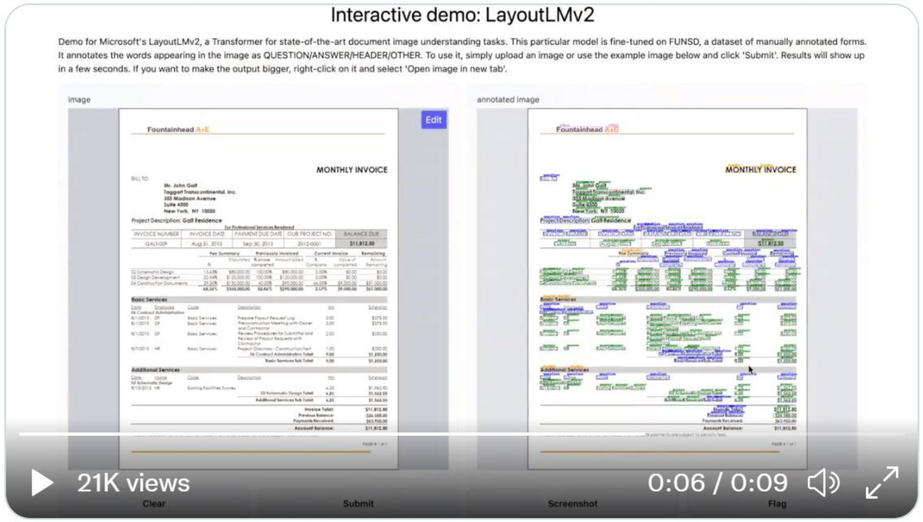
Hugging Face
@huggingface

Document parsing meets 🤗 Transformers!

📄#LayoutLMv2 and #LayoutXLM by @MSFTResearch are now available! 🔥

They're capable of parsing document images (like PDFs) by incorporating text, layout, and visual information, as in the @gradio demo below ⬇️

huggingface.co/spaces/nielsr/...



10:52 PM · Aug 31, 2021 · Twitter Web App

193 Retweets **18** Quote Tweets **716** Likes

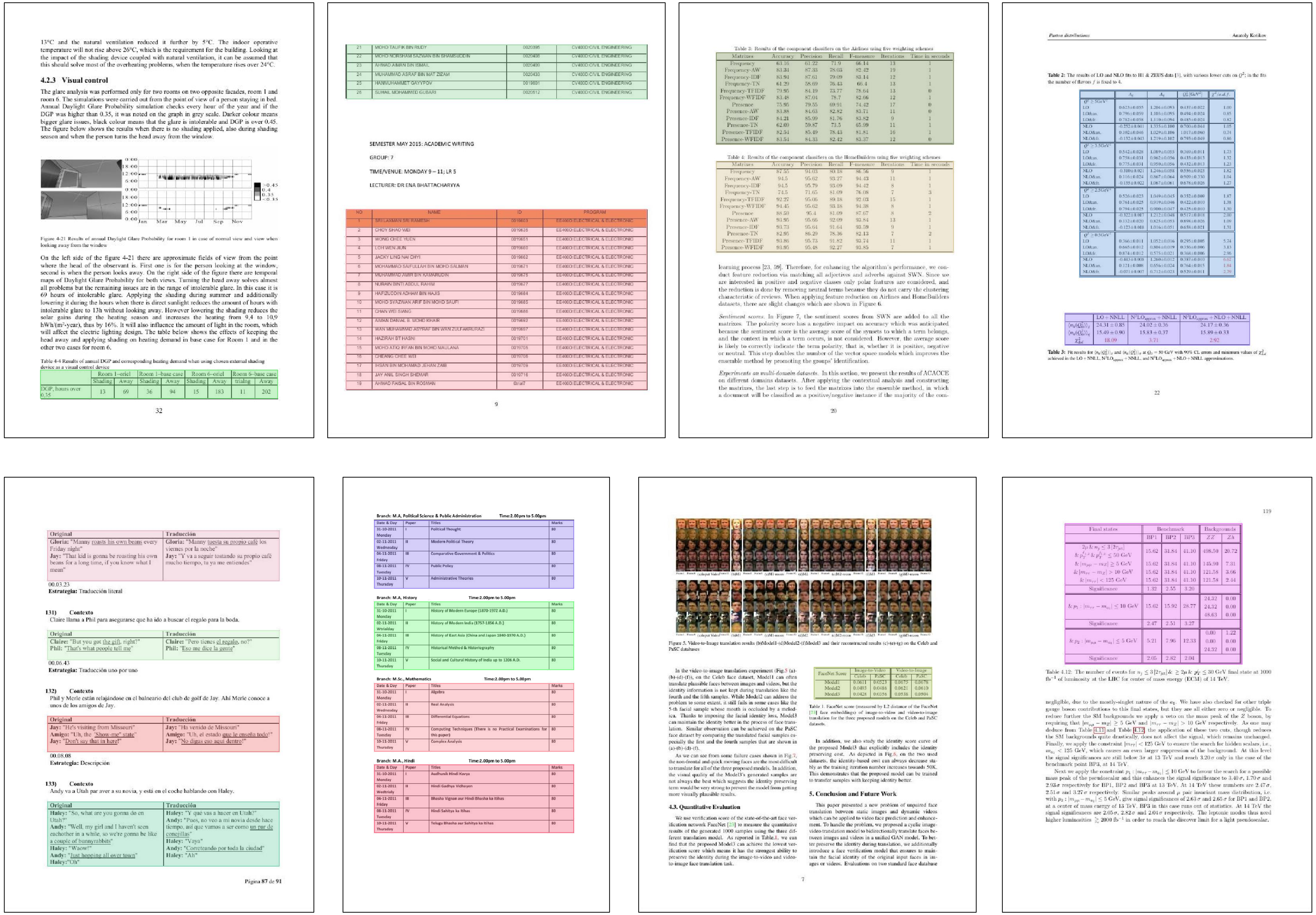
microsoft/layoutlm-base-uncased
Updated Aug 11 • 448k • ♥ 3

microsoft/layoutlmv2-base-uncased
Updated Aug 16 • 54.6k • ♥ 5

microsoft/layoutxlm-base
Updated Aug 26 • 9.12k • ♥ 6

Benchmark Datasets

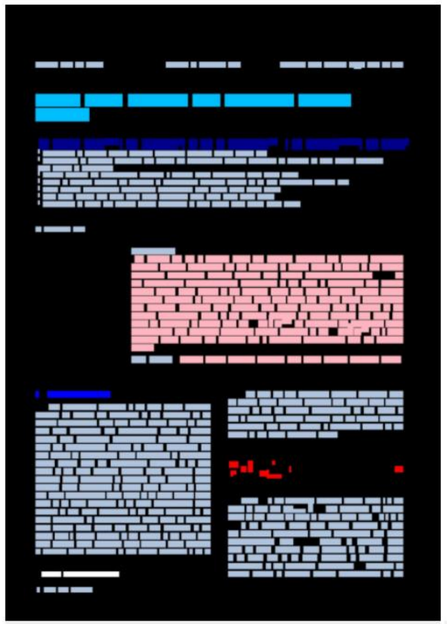
TableBank



[TableBank: A Benchmark Dataset for Table Detection and Recognition](https://aka.ms/tablebank), LREC'2020

Download: https://aka.ms/tablebank

DocBank



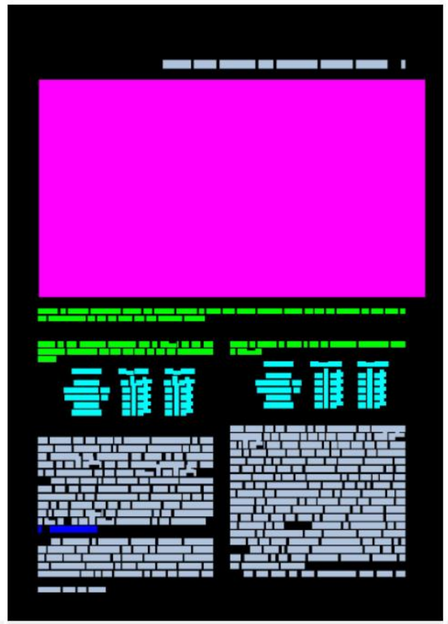
Cosmic String Detection with Tree-Based Machine Learning

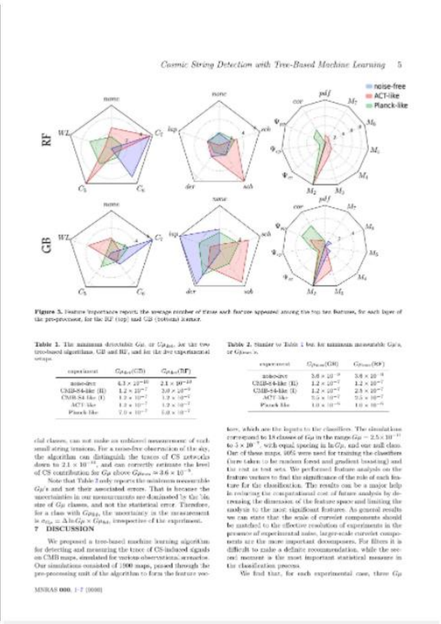
A. Vafaei Sadr[1,2,3], M. Farhang[1], S. M. S. Movahed[1,4] *, B. Bassett[3,5,6,7], M. Kunz[2]

ABSTRACT

We explore the use of random forest and gradient boosting, two powerful tree-based machine learning algorithms, for the detection of cosmic strings in maps of the cosmic microwave background (CMB), through their unique Gott-Kaiser-Stebbins effect on the temperature anisotropies.The information in the maps is compressed into feature vectors before being passed to the learning units. The feature vectors contain various statistical measures of processed CMB maps that boost the cosmic string detectability. Our proposed classifiers, after training, give results improved over or similar to the claimed detectability levels of the existing methods for string tension, $G\mu$. They can make 3σ detection of strings with $G\mu \geq 2.1 \times 10^{-10}$ for noise-free, 0.9′-resolution CMB observations. The minimum detectable tension increases to $G\mu \geq 3.0 \times 10^{-8}$ for a more realistic, CMB S4-like (B) strategy, still a significant improvement over the previous results.

Key words: Cosmic string, Machine learning, Tree based models, Curvelet, CMB.





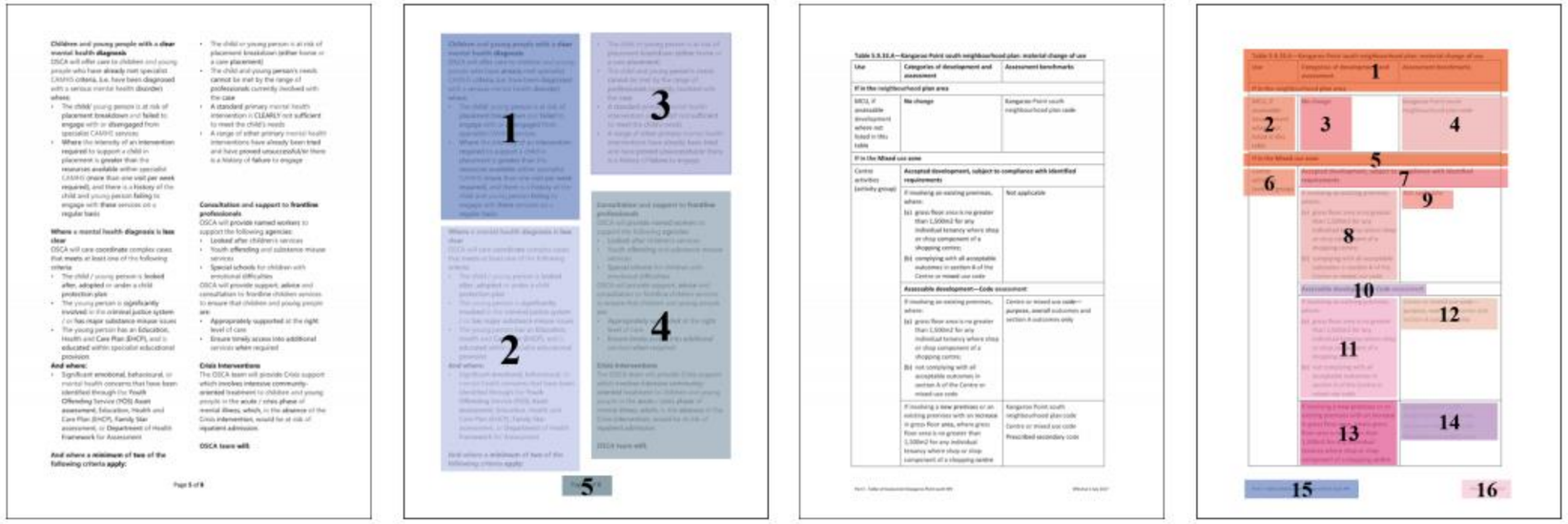
The colors of semantic structure labels are:

Abstract | Author | Caption | Equation | Figure | Footer | List | Paragraph | Reference | Section | Table | Title

[DocBank: A Benchmark Dataset for Document Layout Analysis](), COLING'2020

Download: https://aka.ms/docbank

ReadingBank

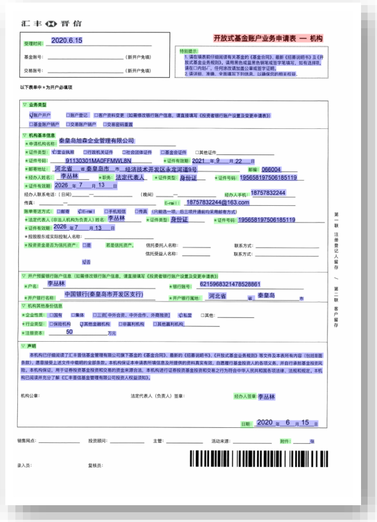


(a) (b) (c) (d)

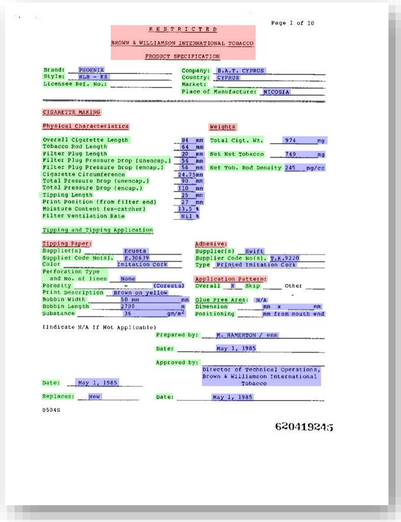
[LayoutReader: Pre-training of Text and Layout for Reading Order Detection](), EMNLP'2021

Download: https://aka.ms/readingbank

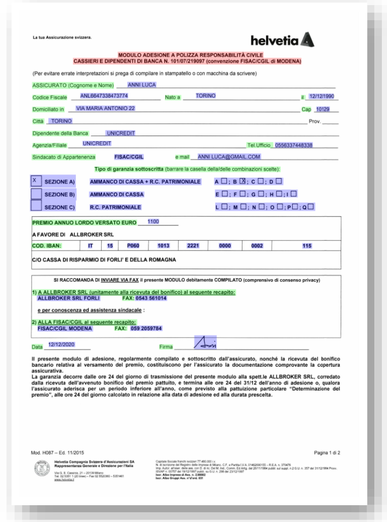
XFUND



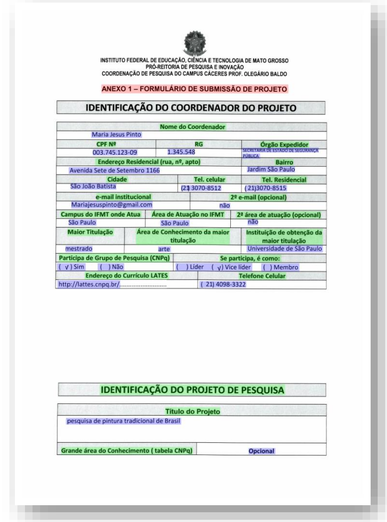
Chinese



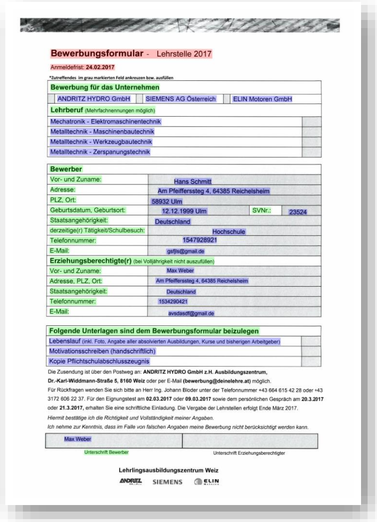
English



Italian

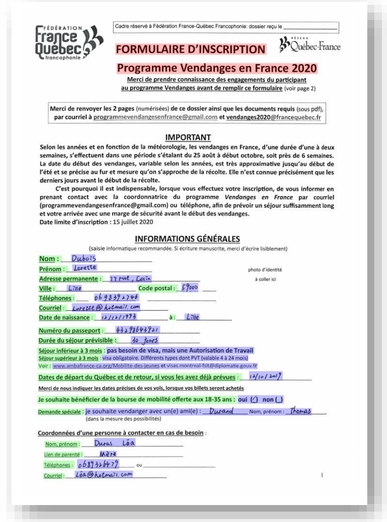


Portuguese

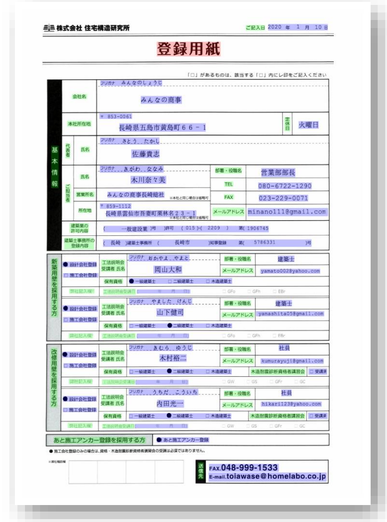


German

Spanish



French



Japanese

[LayoutXLM: Multi-modal Pre-training for Multi-lingual Document Understanding](), Preprint

Download: https://aka.ms/xfund

Applications

Applications

- Information Extraction
  - Form Understanding (***FUNSD***)
    - https://guillaumejaume.github.io/FUNSD/
  - Receipt Understanding (***SROIE, CORD***)
    - https://rrc.cvc.uab.es/?ch=13
    - https://github.com/clovaai/cord
  - Document Information Extraction (***Kleister-NDA***)
    - https://github.com/applicaai/kleister-nda
  - Multi-lingual Form Understanding (***XFUND***)
    - https://aka.ms/xfund
- Classification
  - Document Image Classification (***RVL-CDIP***)
    - https://www.cs.cmu.edu/~aharley/rvl-cdip/
- VQA
  - Document Visual Question Answering (***DocVQA***)
    - https://rrc.cvc.uab.es/?ch=17
- Layout Analysis
  - Table Detection (***TableBank***)
    - https://aka.ms/tablebank
  - Page Object Detection (***DocBank***)
    - https://aka.ms/docbank
  - Reading Order Detection (***ReadingBank***)
    - https://aka.ms/readingbank

Semantic Entity Recognition

Model	FUNSD	CORD	SROIE	Kleister-NDA
$\text{BERT}_{\text{BASE}}$	0.6026	0.8968	0.9099	0.7790
$\text{UniLMv2}_{\text{BASE}}$	0.6648	0.9092	0.9459	0.7950
$\text{BERT}_{\text{LARGE}}$	0.6563	0.9025	0.9200	0.7910
$\text{UniLMv2}_{\text{LARGE}}$	0.7072	0.9205	0.9488	0.8180
$\text{LayoutLM}_{\text{BASE}}$	0.7866	0.9472	0.9438	0.8270
$\text{LayoutLM}_{\text{LARGE}}$	0.7895	0.9493	0.9524	0.8340
$\text{LayoutLMv2}_{\text{BASE}}$	0.8276	0.9495	0.9625	0.8330
$\text{LayoutLMv2}_{\text{LARGE}}$	**0.8420**	**0.9601**	**0.9781**	**0.8520**
BROS (Hong et al., 2021)	0.8121	0.9536	0.9548	–
SPADE (Hwang et al., 2020)	–	0.9150	–	–
PICK (Yu et al., 2020)	–	–	0.9612	–
TRIE (Zhang et al., 2020)	–	–	0.9618	–
Top-1 on SROIE Leaderboard (until 2020-12-24)	–	–	0.9767	–
$\text{RoBERTa}_{\text{BASE}}$ in (Graliński et al., 2020)	–	–	–	0.7930

Table 2: Entity-level F1 scores of the four entity extraction tasks: FUNSD, CORD, SROIE and Kleister-NDA.

Document Image Classification

Model	Accuracy	#Parameters
$\text{BERT}_{\text{BASE}}$	89.81%	110M
$\text{UniLMv2}_{\text{BASE}}$	90.06%	125M
$\text{BERT}_{\text{LARGE}}$	89.92%	340M
$\text{UniLMv2}_{\text{LARGE}}$	90.20%	355M
$\text{LayoutLM}_{\text{BASE}}$ (w/ image)	94.42%	160M
$\text{LayoutLM}_{\text{LARGE}}$ (w/ image)	94.43%	390M
$\text{LayoutLMv2}_{\text{BASE}}$	95.25%	200M
$\text{LayoutLMv2}_{\text{LARGE}}$	**95.64%**	426M
VGG-16 (Afzal et al., 2017)	90.97%	-
Single model (Das et al., 2018)	91.11%	-
Ensemble (Das et al., 2018)	92.21%	-
InceptionResNetV2[6] (Szegedy et al., 2016)	92.63%	-
LadderNet (Sarkhel & Nandi, 2019)	92.77%	-
Single model (Dauphinee et al., 2019)	93.03%	-
Ensemble (Dauphinee et al., 2019)	93.07%	-

Table 5: Classification accuracy on the RVL-CDIP dataset

Document VQA

Model	Fine-tuning set	ANLS	#Parameters
$\text{BERT}_{\text{BASE}}$	train	0.6354	110M
$\text{UniLMv2}_{\text{BASE}}$	train	0.7134	125M
$\text{BERT}_{\text{LARGE}}$	train	0.6768	340M
$\text{UniLMv2}_{\text{LARGE}}$	train	0.7709	355M
$\text{LayoutLM}_{\text{BASE}}$	train	0.6979	113M
$\text{LayoutLM}_{\text{LARGE}}$	train	0.7259	343M
$\text{LayoutLMv2}_{\text{BASE}}$	train	0.7808	200M
$\text{LayoutLMv2}_{\text{LARGE}}$	train	0.8348	426M
$\text{LayoutLMv2}_{\text{LARGE}}$	train + dev	0.8529	426M
$\text{LayoutLMv2}_{\text{LARGE}}$ + QG	train + dev	**0.8672**	426M
Top-1 on DocVQA Leaderboard (30 models ensemble)[7]	-	0.8506	-

Table 6: Average Normalized Levenshtein Similarity (ANLS) score on the DocVQA dataset (until 2020-12-24), "QG" denotes the data augmentation with the question generation dataset.

Table Detection

Models	Word			Latex			Word+Latex		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ResNeXt-101 (Word)	0.9496	0.8388	0.8908	0.9902	0.5948	0.7432	0.9594	0.7607	0.8486
ResNeXt-152 (Word)	0.9530	0.8829	**0.9166**	0.9808	0.6890	0.8094	0.9603	0.8209	0.8851
ResNeXt-101 (Latex)	0.8288	0.9395	0.8807	0.9854	0.9760	0.9807	0.8744	0.9512	0.9112
ResNeXt-152 (Latex)	0.8259	0.9562	0.8863	0.9867	0.9754	**0.9810**	0.8720	0.9624	0.9149
ResNeXt-101 (Word+Latex)	0.9557	0.8403	0.8943	0.9886	0.9694	0.9789	0.9670	0.8817	0.9224
ResNeXt-152 (Word+Latex)	0.9540	0.8639	0.9067	0.9885	0.9732	0.9808	0.9657	0.8989	**0.9311**

Table 2: Evaluation results on Word and Latex datasets with ResNeXt-{101,152} as the backbone networks

Models	Precision	Recall	F1
ICDAR 2013 (train)	0.9748	0.7997	0.8786
UNLV	0.9185	0.9639	0.9406
Marmot	0.7692	0.9844	0.8636
DeepFigures	0.8527	0.9348	0.8918
TableBank (ResNeXt-152, Word)	0.9725	0.8528	0.9087
TableBank (ResNeXt-152, Latex)	**0.9658**	**0.9594**	**0.9625**
TableBank (ResNeXt-152, Word + Latex)	0.9635	0.9039	0.9328
Tesseract	0.9439	0.7144	0.8133
Camelot	0.9785	0.6856	0.8063

Table 3: Evaluation results on ICDAR 2013 dataset

Page Object Detection

Models	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	Reference	Section	Table	Title	Macro average
$\text{BERT}_{\text{BASE}}$	0.9294	0.8484	0.8629	0.8152	1.0000	0.7805	0.7133	0.9619	0.9310	0.9081	0.8296	0.9442	0.8770
$\text{RoBERTa}_{\text{BASE}}$	0.9288	0.8618	0.8944	0.8248	1.0000	0.8014	0.7353	0.9646	0.9341	0.9337	0.8389	0.9511	0.8891
$\text{LayoutLM}_{\text{BASE}}$	**0.9816**	0.8595	0.9597	0.8947	1.0000	0.8957	0.8948	0.9788	0.9338	0.9598	0.8633	**0.9579**	0.9316
$\text{BERT}_{\text{LARGE}}$	0.9286	0.8577	0.8650	0.8177	1.0000	0.7814	0.6960	0.9619	0.9284	0.9065	0.8320	0.9430	0.8765
$\text{RoBERTa}_{\text{LARGE}}$	0.9479	0.8724	0.9081	0.8370	1.0000	0.8392	0.7451	0.9665	0.9334	0.9407	0.8494	0.9461	0.8988
$\text{LayoutLM}_{\text{LARGE}}$	0.9784	0.8783	0.9556	0.8974	**1.0000**	0.9146	0.9004	0.9790	0.9332	0.9596	0.8679	0.9552	0.9350
X101	0.9717	0.8227	0.9435	0.8938	0.8812	0.9029	0.9051	0.9682	0.8798	0.9412	0.8353	0.9158	0.9051
$\text{X101+LayoutLM}_{\text{BASE}}$	0.9815	0.8907	**0.9669**	0.9430	0.9990	0.9292	**0.9300**	0.9843	**0.9437**	0.9664	0.8818	0.9575	0.9478
$\text{X101+LayoutLM}_{\text{LARGE}}$	0.9802	**0.8964**	0.9666	**0.9440**	0.9994	**0.9352**	0.9293	**0.9844**	0.9430	**0.9670**	**0.8875**	0.9531	**0.9488**

Table 4: The performance of BERT, RoBERTa, LayoutLM and Faster R-CNN on the DocBank test set.

Reading Order Detection

Method	Encoder	Avg. Page-level BLEU ↑	ARD ↓
Heuristic Method	-	0.6972	8.46
LayoutReader (text only)	BERT	0.8510	12.08
	UniLM	0.8765	10.65
LayoutReader (layout only)	LayoutLM (layout only)	0.9732	2.31
LayoutReader	LayoutLM	**0.9819**	**1.75**

Table 2: Evaluation results of the LayoutReader on the reading order detection task, where the source-side of training/testing data is in the left-to-right and top-to-bottom order

Method	Avg. Page-level BLEU ↑	ARD ↓
Heuristic Method	0.3391	13.61
Tesseract OCR	0.7532	1.42
LayoutReader	**0.9360**	**0.27**

Table 5: Adaption to text lines of Tesseract OCR

Method	Avg. Page-level BLEU ↑	ARD ↓
Heuristic Method	0.3752	10.17
The commercial OCR	0.8530	2.40
LayoutReader	**0.9430**	**0.59**

Table 6: Adaption to text lines of the commercial OCR

Multilingual Document AI

Language-specific Fine-tuning

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	0.8774	0.7761	0.6105	0.6743	0.6687	0.6814	0.6818	0.7047
	InfoXLM$_{\text{BASE}}$	0.6852	0.8868	0.7865	0.6230	0.7015	0.6751	0.7063	0.7008	0.7207
	LayoutXLM$_{\text{BASE}}$	**0.794**	**0.8924**	**0.7921**	**0.7550**	**0.7902**	**0.8082**	**0.8222**	**0.7903**	**0.8056**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.8925	0.7817	0.6515	0.7170	0.7139	0.711	0.7241	0.7374
	InfoXLM$_{\text{LARGE}}$	0.7325	0.8955	0.7904	0.6740	0.7140	0.7152	0.7338	0.7212	0.7471
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.9161**	**0.8033**	**0.7830**	**0.8098**	**0.8275**	**0.8361**	**0.8273**	**0.8282**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.2659	0.5105	0.5800	0.5295	0.4965	0.5305	0.5041	0.3982	0.4769
	InfoXLM$_{\text{BASE}}$	0.2920	0.5214	0.6000	0.5516	0.4913	0.5281	0.5262	0.4170	0.4910
	LayoutXLM$_{\text{BASE}}$	**0.5483**	**0.7073**	**0.6963**	**0.6896**	**0.6353**	**0.6415**	**0.6551**	**0.5718**	**0.6432**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.6475	0.6798	0.6330	0.6080	0.6171	0.6189	0.5762	0.5910
	InfoXLM$_{\text{LARGE}}$	0.3679	0.6775	0.6604	0.6346	0.6096	0.6659	0.6057	0.5800	0.6002
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.7888**	**0.7255**	**0.7666**	**0.7102**	**0.7691**	**0.6843**	**0.6796**	**0.7206**

Table 2: Language-specific fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on X, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

SER: Semantic Entity Recognition (headers, keys, values)

RE: Relation extraction for key-value pairs

Zero-shot Transfer

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	0.4144	0.3023	0.3055	0.371	0.2767	0.3286	0.3936	0.3824
	InfoXLM$_{\text{BASE}}$	0.6852	0.4408	0.3603	0.3102	0.4021	0.2880	0.3587	0.4502	0.4119
	LayoutXLM$_{\text{BASE}}$	**0.794**	**0.6019**	**0.4715**	**0.4565**	**0.5757**	**0.4846**	**0.5252**	**0.539**	**0.5561**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.5205	0.3939	0.3627	0.4672	0.3398	0.418	0.4997	0.4637
	InfoXLM$_{\text{LARGE}}$	0.7325	0.5536	0.4132	0.3689	0.4909	0.3598	0.4363	0.5126	0.4835
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.6896**	**0.519**	**0.4976**	**0.6135**	**0.5517**	**0.5905**	**0.6077**	**0.6115**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.2659	0.1601	0.2611	0.2440	0.2240	0.2374	0.2288	0.1996	0.2276
	InfoXLM$_{\text{BASE}}$	0.2920	0.2405	0.2851	0.2481	0.2454	0.2193	0.2027	0.2049	0.2423
	LayoutXLM$_{\text{BASE}}$	**0.5483**	**0.4494**	**0.4408**	**0.4708**	**0.4416**	**0.4090**	**0.3820**	**0.3685**	**0.4388**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.2421	0.3037	0.2843	0.2897	0.2496	0.2617	0.2333	0.2765
	InfoXLM$_{\text{LARGE}}$	0.3679	0.3156	0.3364	0.3185	0.3189	0.2720	0.2953	0.2554	0.3100
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.5531**	**0.5696**	**0.5780**	**0.5615**	**0.5184**	**0.4890**	**0.4795**	**0.5487**

Table 3: Zero-shot transfer accuracy (F1) on the XFUND dataset (fine-tuning on FUNSD, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

Multitask Learning

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.6633	0.883	0.7786	0.6223	0.7035	0.6814	0.7146	0.6726	0.7149
	InfoXLM$_{\text{BASE}}$	0.6538	0.8741	0.7855	0.5979	0.7057	0.6826	0.7055	0.6796	0.7106
	LayoutXLM$_{\text{BASE}}$	**0.7924**	**0.8973**	**0.7964**	**0.7798**	**0.8173**	**0.821**	**0.8322**	**0.8241**	**0.8201**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7151	0.8967	0.7828	0.6615	0.7407	0.7165	0.7431	0.7449	0.7502
	InfoXLM$_{\text{LARGE}}$	0.7246	0.8919	0.7998	0.6702	0.7376	0.7180	0.7523	0.7332	0.7534
	LayoutXLM$_{\text{LARGE}}$	**0.8068**	**0.9155**	**0.8216**	**0.8055**	**0.8384**	**0.8372**	**0.853**	**0.8650**	**0.8429**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.3638	0.6797	0.6829	0.6828	0.6727	0.6937	0.6887	0.6082	0.6341
	InfoXLM$_{\text{BASE}}$	0.3699	0.6493	0.6473	0.6828	0.6831	0.6690	0.6384	0.5763	0.6145
	LayoutXLM$_{\text{BASE}}$	**0.6671**	**0.8241**	**0.8142**	**0.8104**	**0.8221**	**0.8310**	**0.7854**	**0.7044**	**0.7823**
	XLM-RoBERTa$_{\text{LARGE}}$	0.4246	0.7316	0.7350	0.7513	0.7532	0.7520	0.7111	0.6582	0.6896
	InfoXLM$_{\text{LARGE}}$	0.4543	0.7311	0.7510	0.7644	0.7549	0.7504	0.7356	0.6875	0.7037
	LayoutXLM$_{\text{LARGE}}$	**0.7683**	**0.9000**	**0.8621**	**0.8592**	**0.8669**	**0.8675**	**0.8263**	**0.8160**	**0.8458**

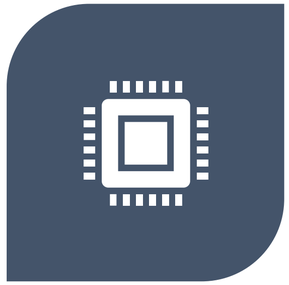
Table 4: Multitask fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on 8 languages all, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

Challenges in Document AI

MODEL LIMITATIONS

DATA QUALITY IN REAL-WORD

TASK CORRELATIONS



DATA/COMPUTATION INSUFFICIENCY

Document AI @MSRA

- Multimodal Pre-trained Models
  - ***LayoutLM*** (KDD'2020)
  - ***LayoutLMv2*** (ACL'2021)
  - ***LayoutXLM*** (Preprint)

- Benchmark Datasets
  - ***TableBank*** (LREC'2020)
  - ***DocBank*** (COLING'2020)
  - ***ReadingBank*** (EMNLP'2021)
  - ***XFUND*** (with LayoutXLM)

- Our paper "***Document AI: Benchmarks, Models and Applications***" will be publicly available soon