Microsoft

ICDAR
LAUSANNE 2021

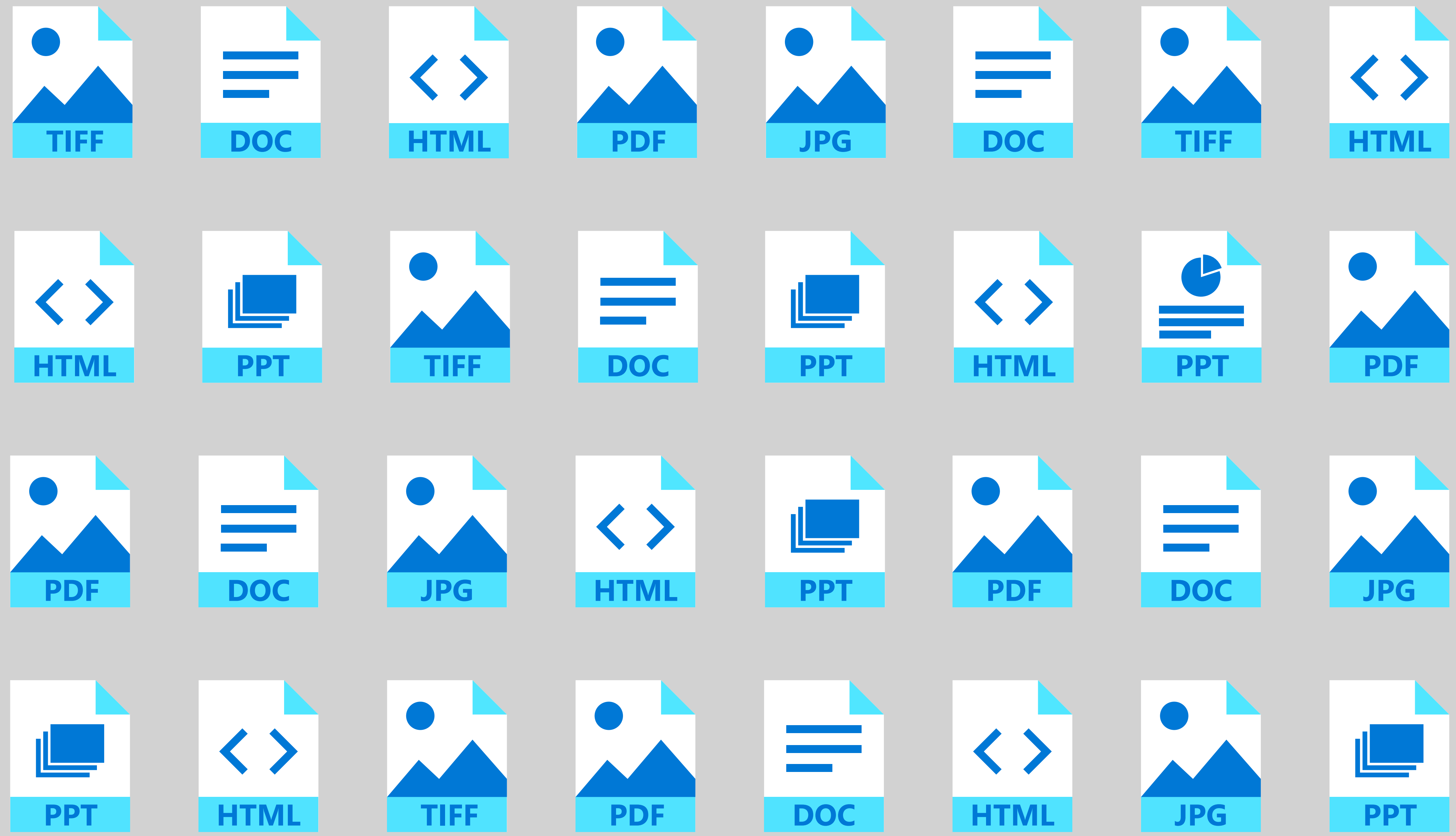
Document AI: Benchmarks, Models and Applications

Lei Cui
Microsoft Research Asia

DIL workshop, ICDAR 2021
2021-09-06

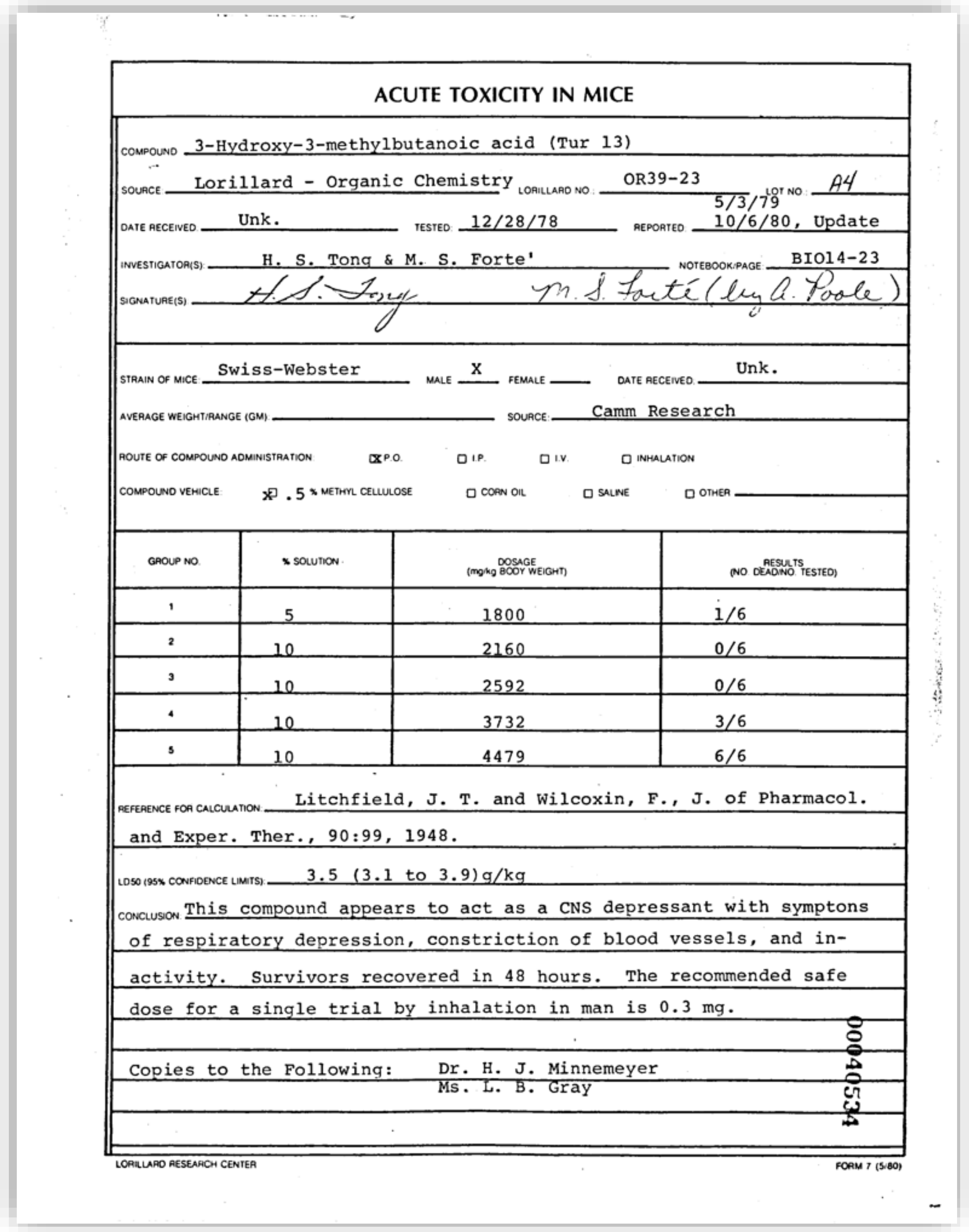
Document AI

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



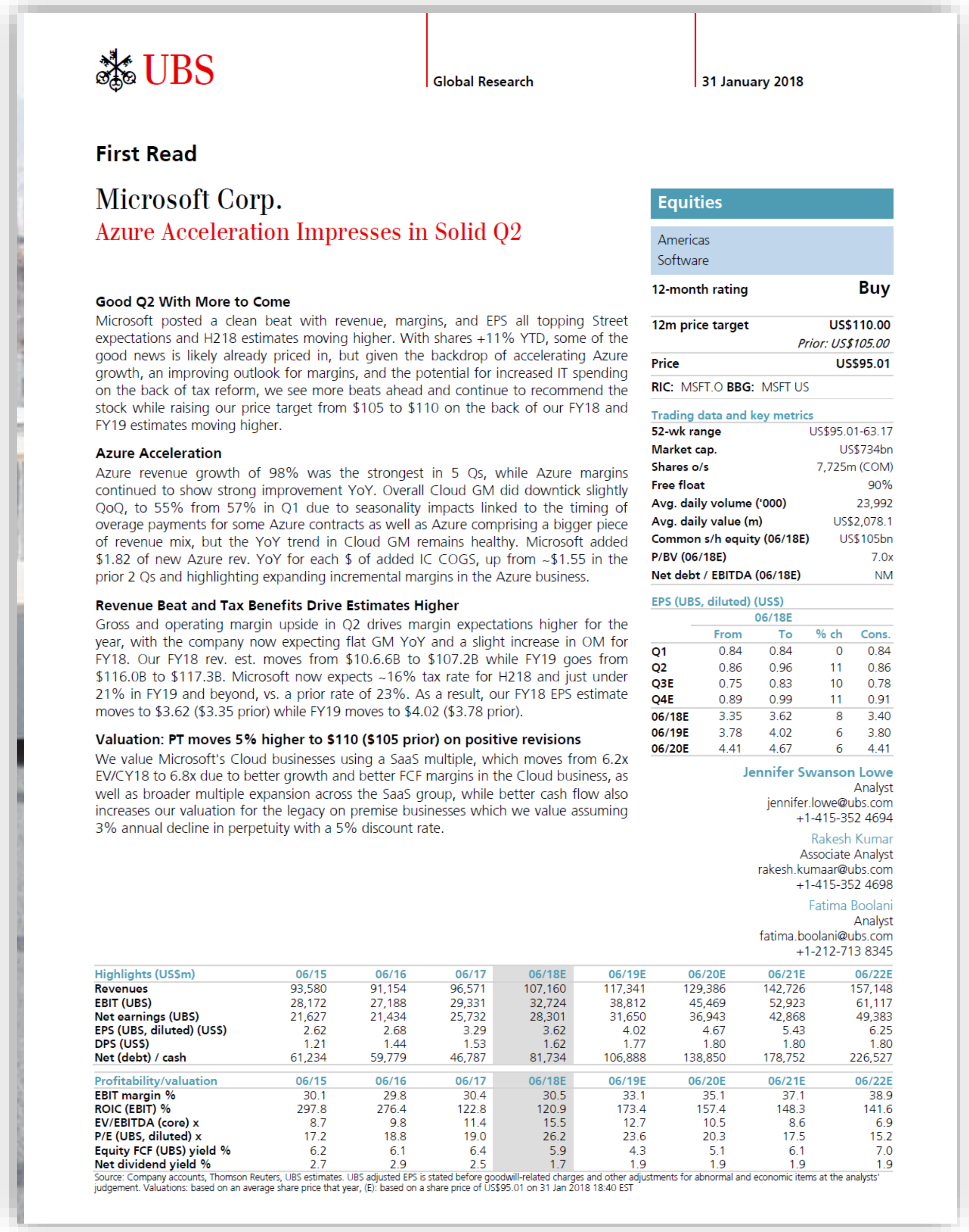
Uncover latent insights from all your content

Document AI in Real World

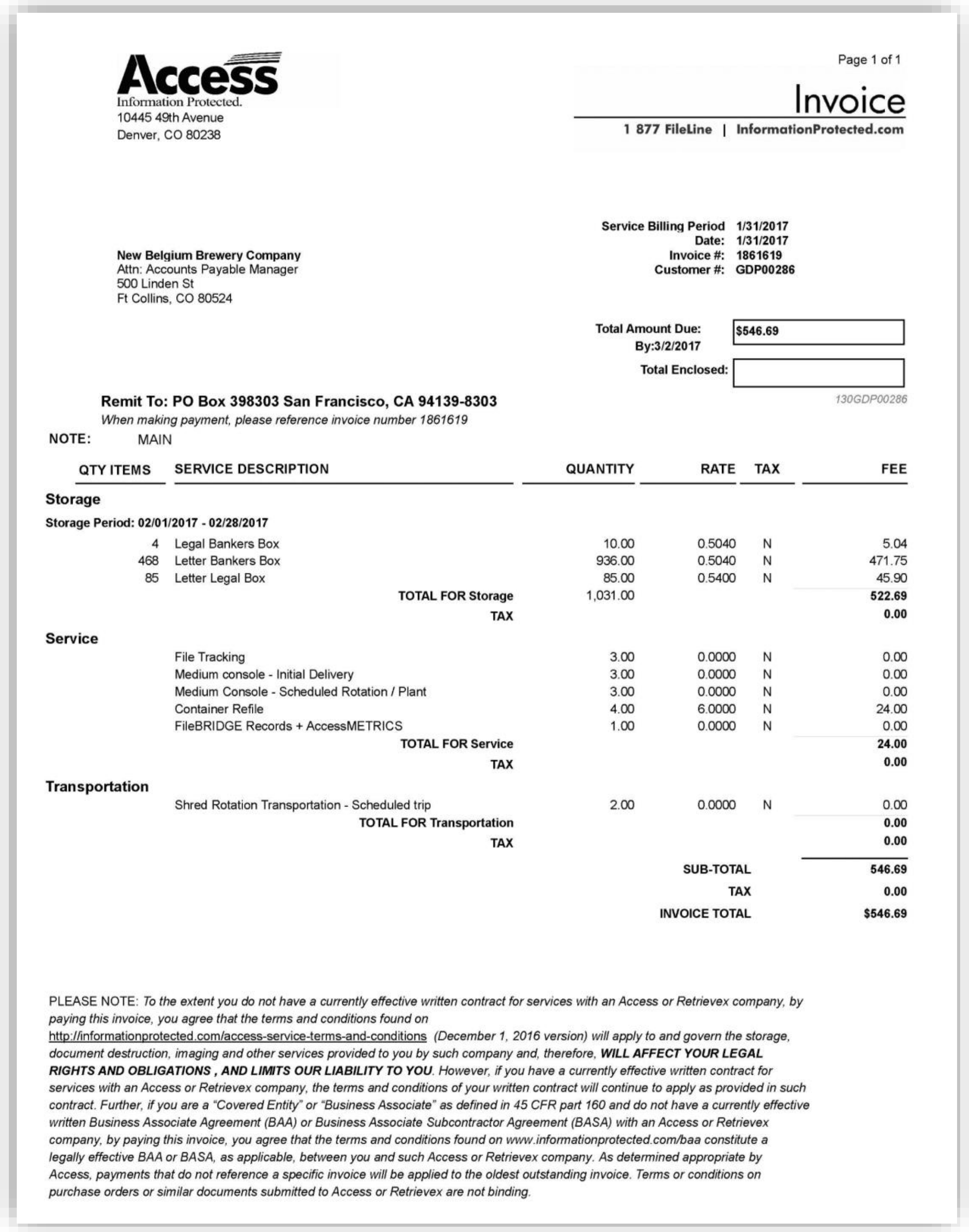


Form

Receipt



Report



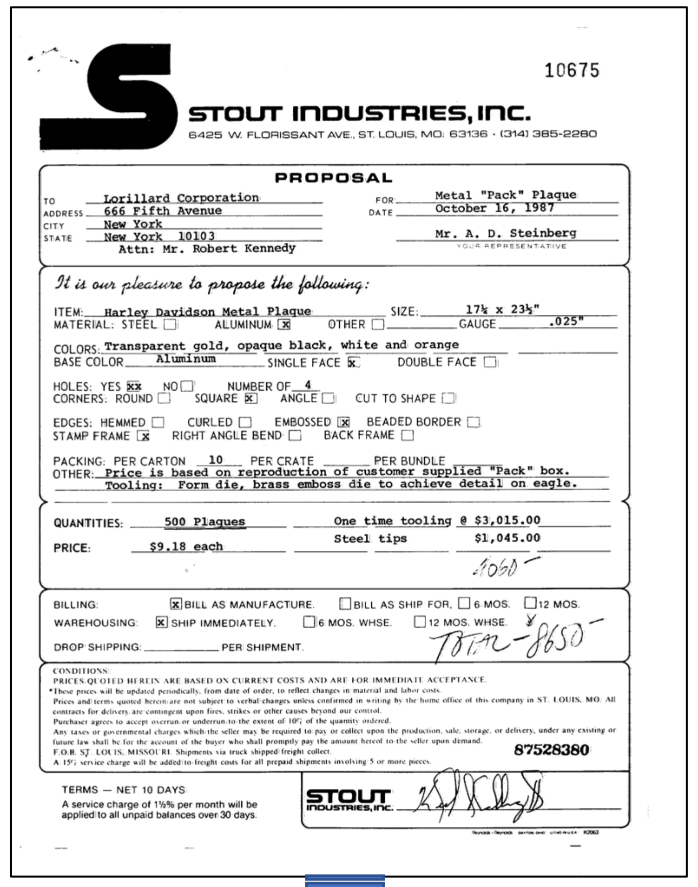
Invoice

Scanned documents (.jpg, .png, ...)

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

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

Document AI Tasks



10675

STOUT INDUSTRIES, INC.

PROPOSAL

TO Lorillard Corporation
ADDRESS 666 Fifth Avenue
CITY New York
STATE New York 10103
Attn: Mr. Robert Kennedy

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

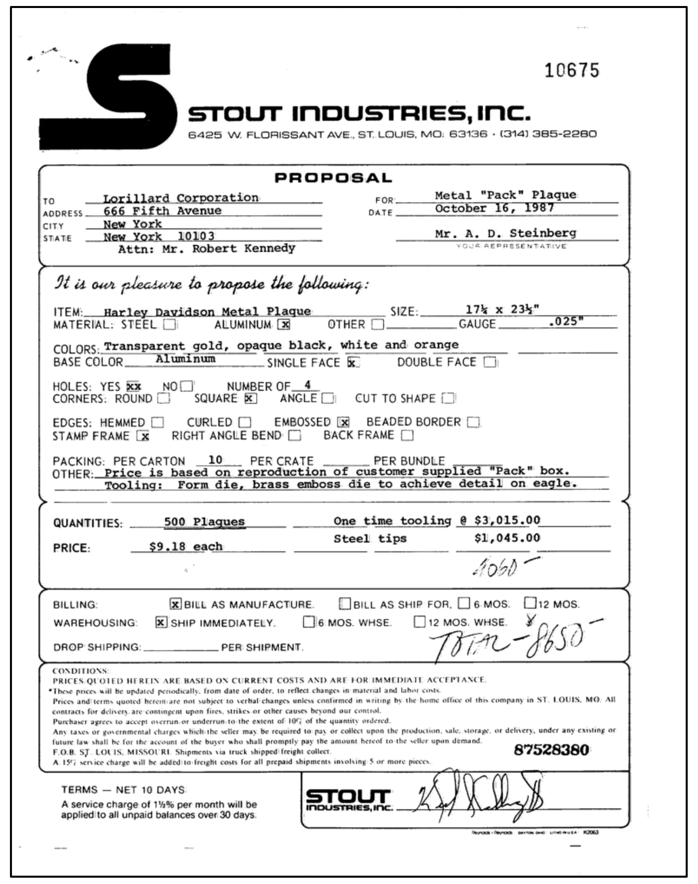
Vt Pep Mocha	4.95
Sbux Card	4.95
XXXXXXXXXXXX3228	
Subtotal	$4.95
Total	$4.95
Change Due	$0.00

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



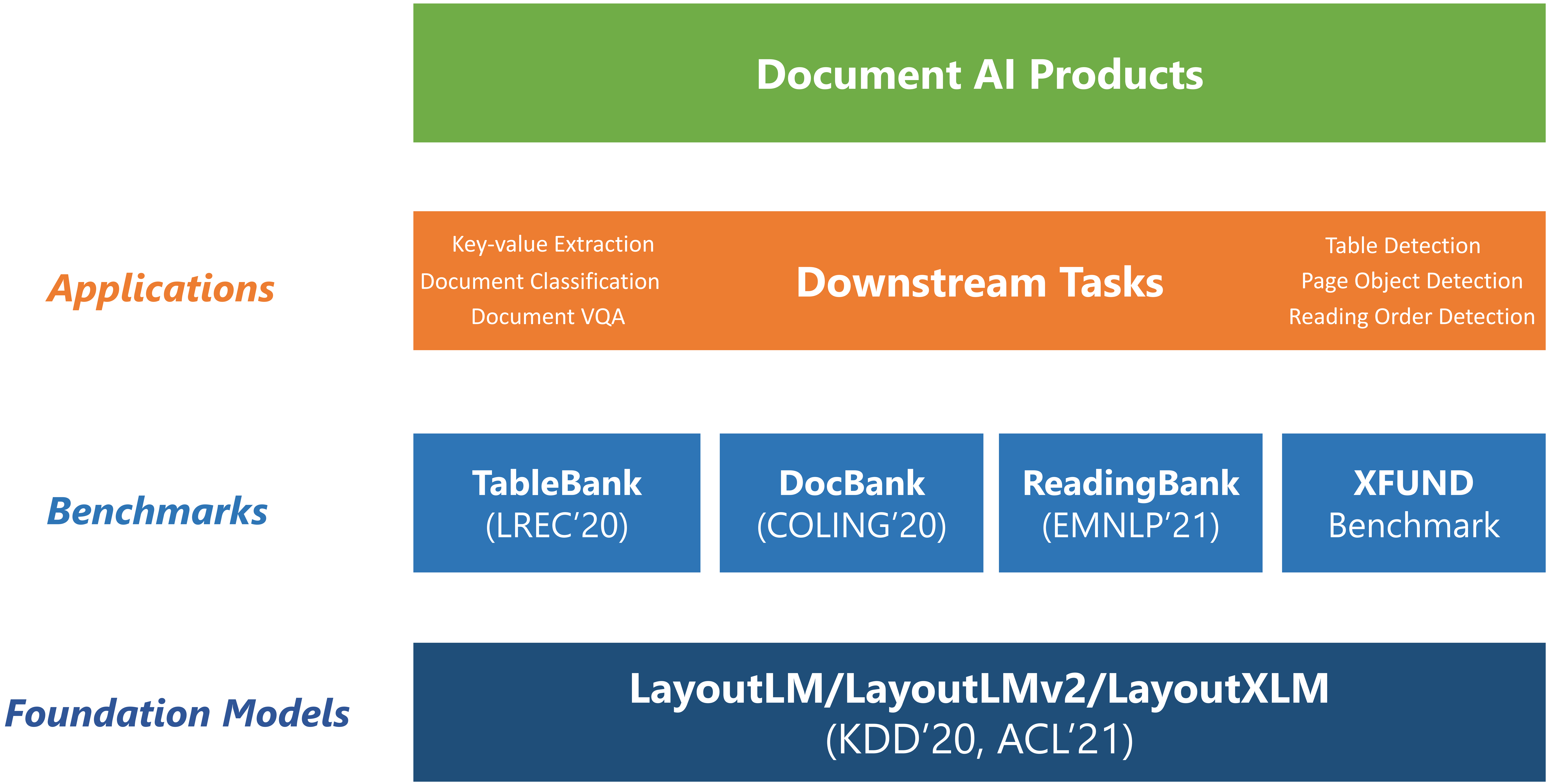
10675

STOUT INDUSTRIES, INC.

PROPOSAL

Category: Form

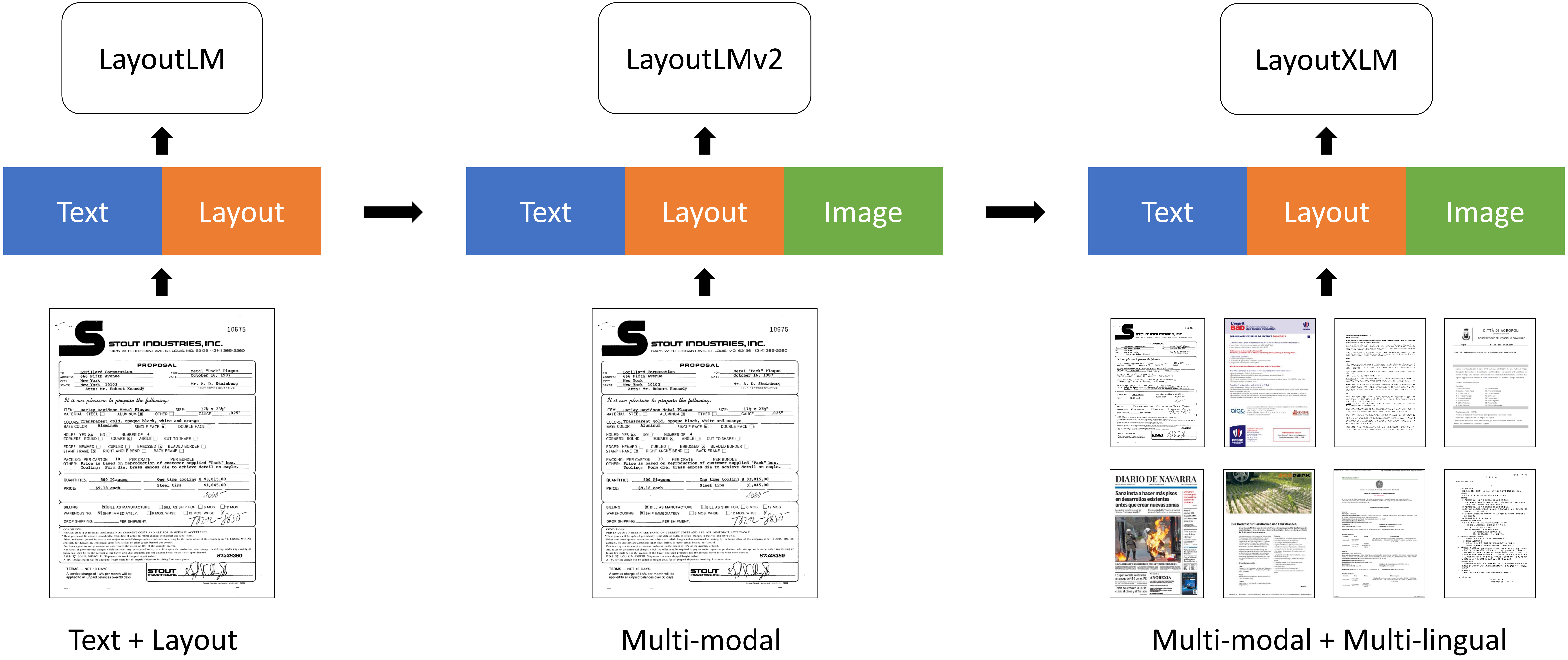
Document Image Classification



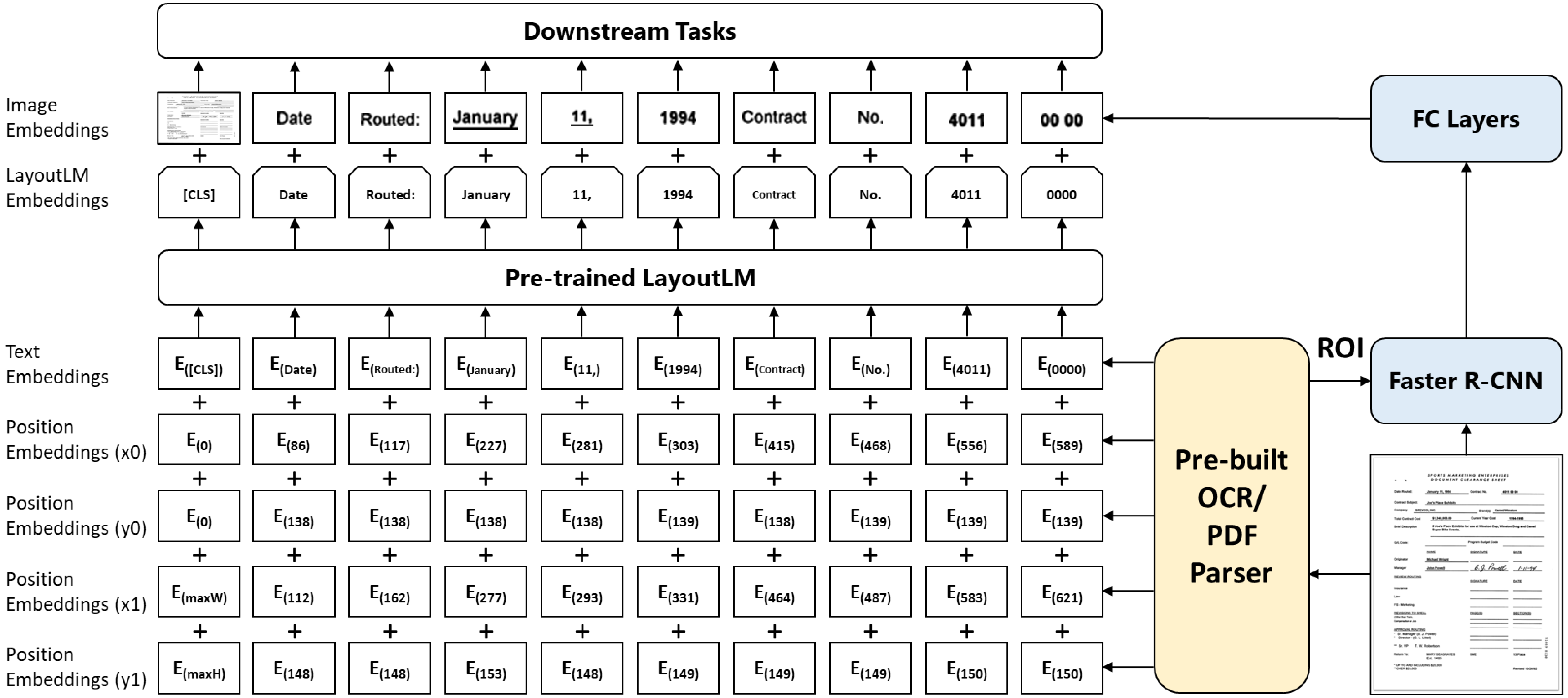
https://aka.ms/document-ai/

LayoutLM Model Family

LayoutLM -> LayoutLMv2 -> LayoutXLM

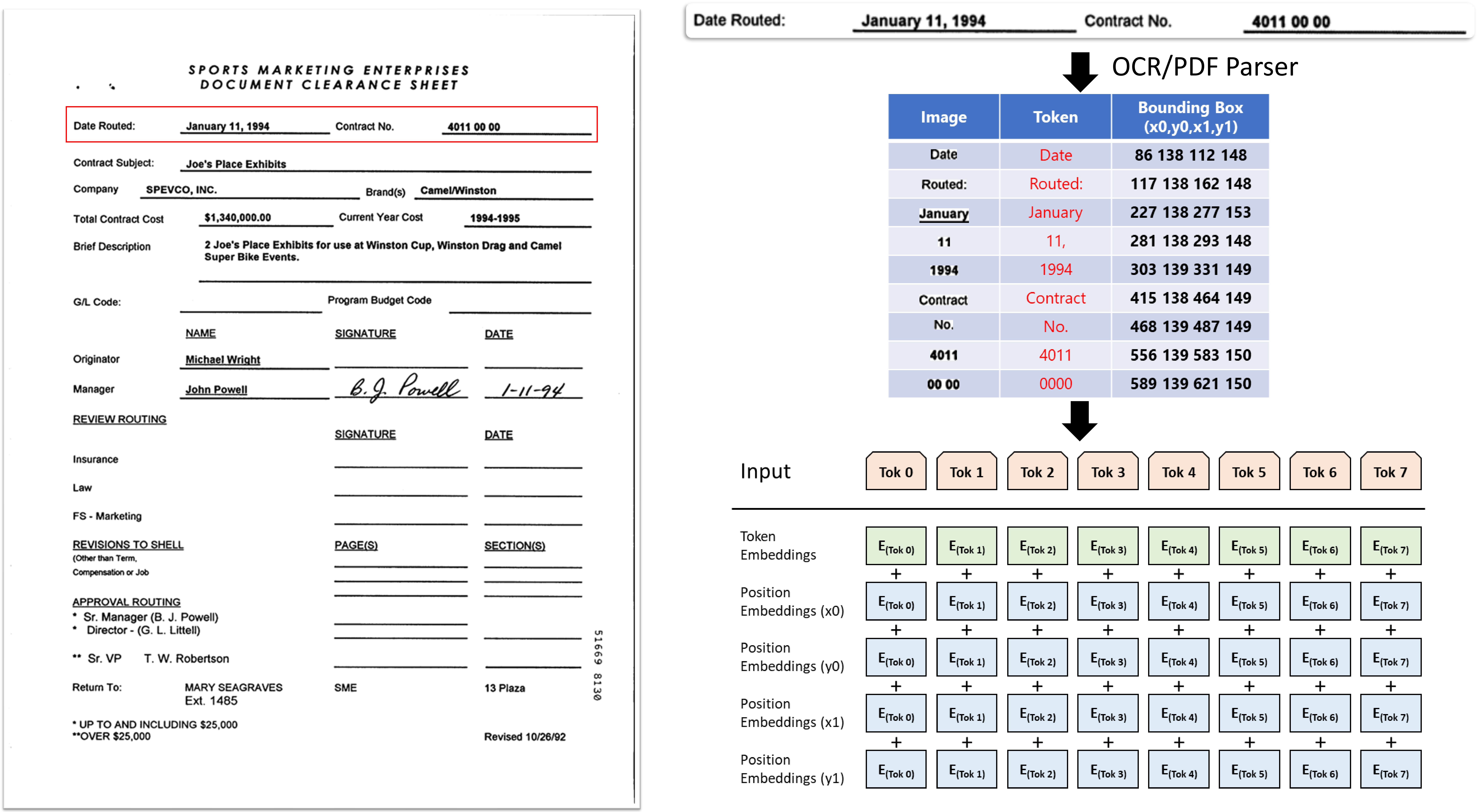


LayoutLM

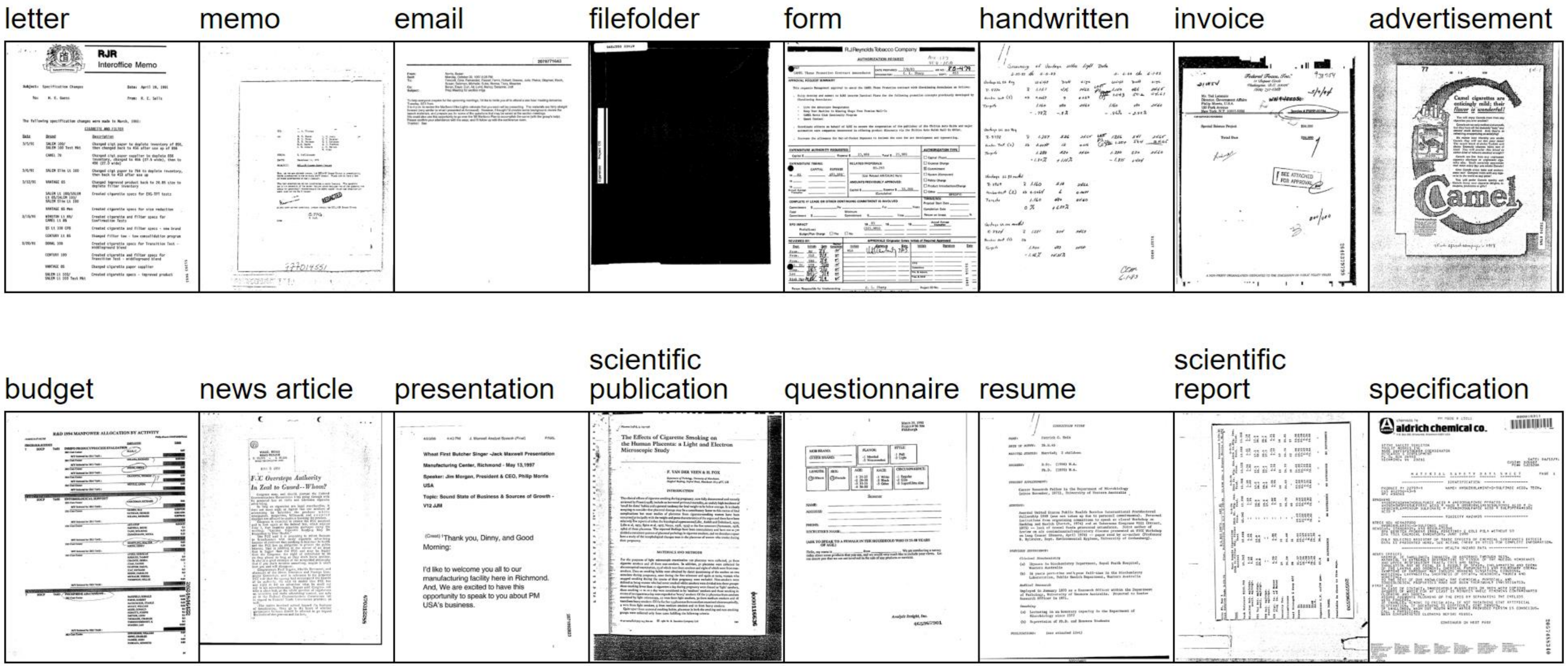


* Text embeddings initialized by BERT/UniLM

[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
Input
Tok 0 Tok 1 Tok 2 Tok 3 Tok 4 Tok 5 Tok 6 Tok 7
Token Embeddings
Position Embeddings (x0)
Position Embeddings (y0)
Position Embeddings (x1)
Position Embeddings (y1)
E(Tok 0) E(Tok 1) E(Tok 2) E(Tok 3) E(Tok 4) E(Tok 5) E(Tok 6) E(Tok 7)
+

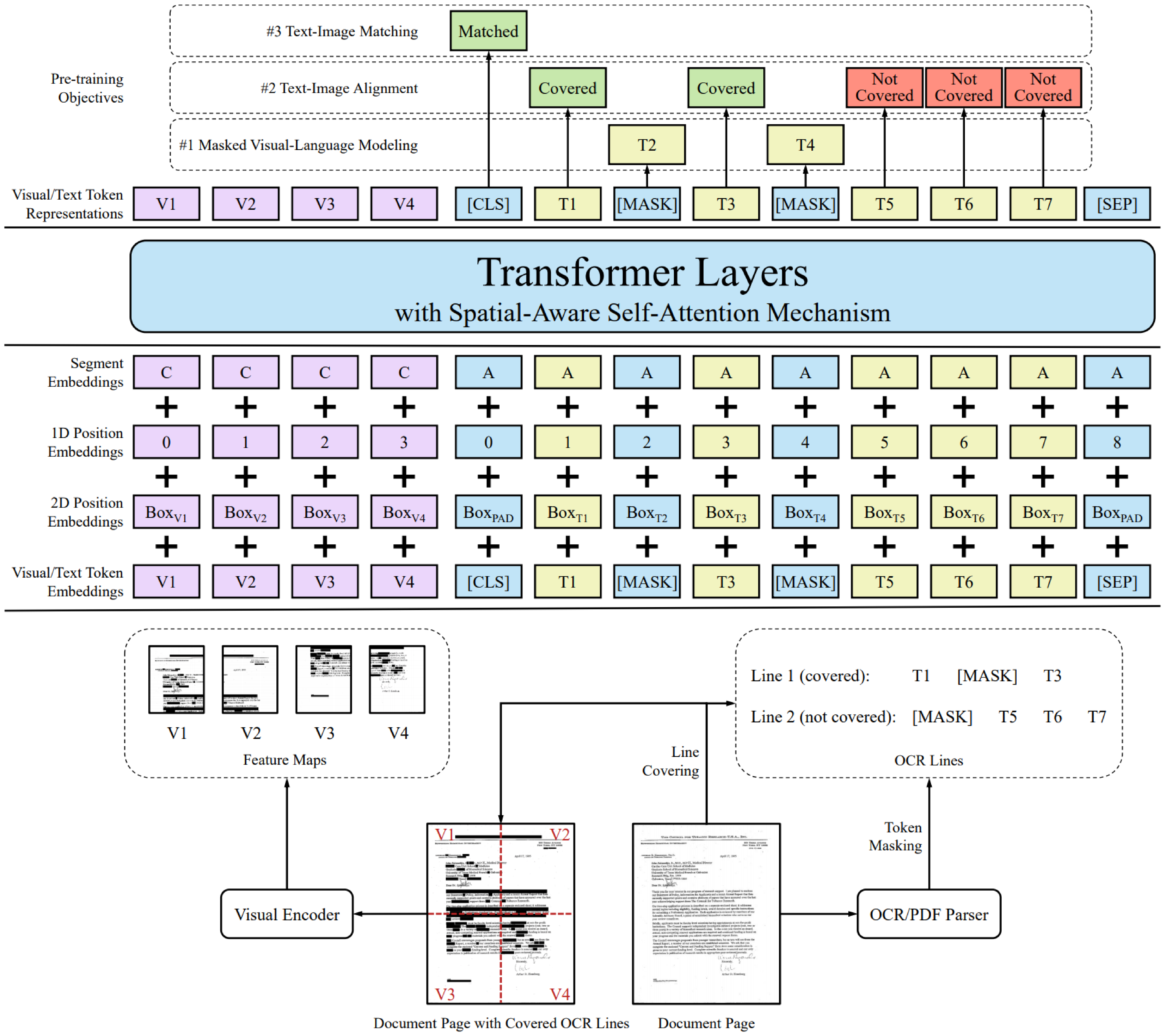
Pre-training Data



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

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

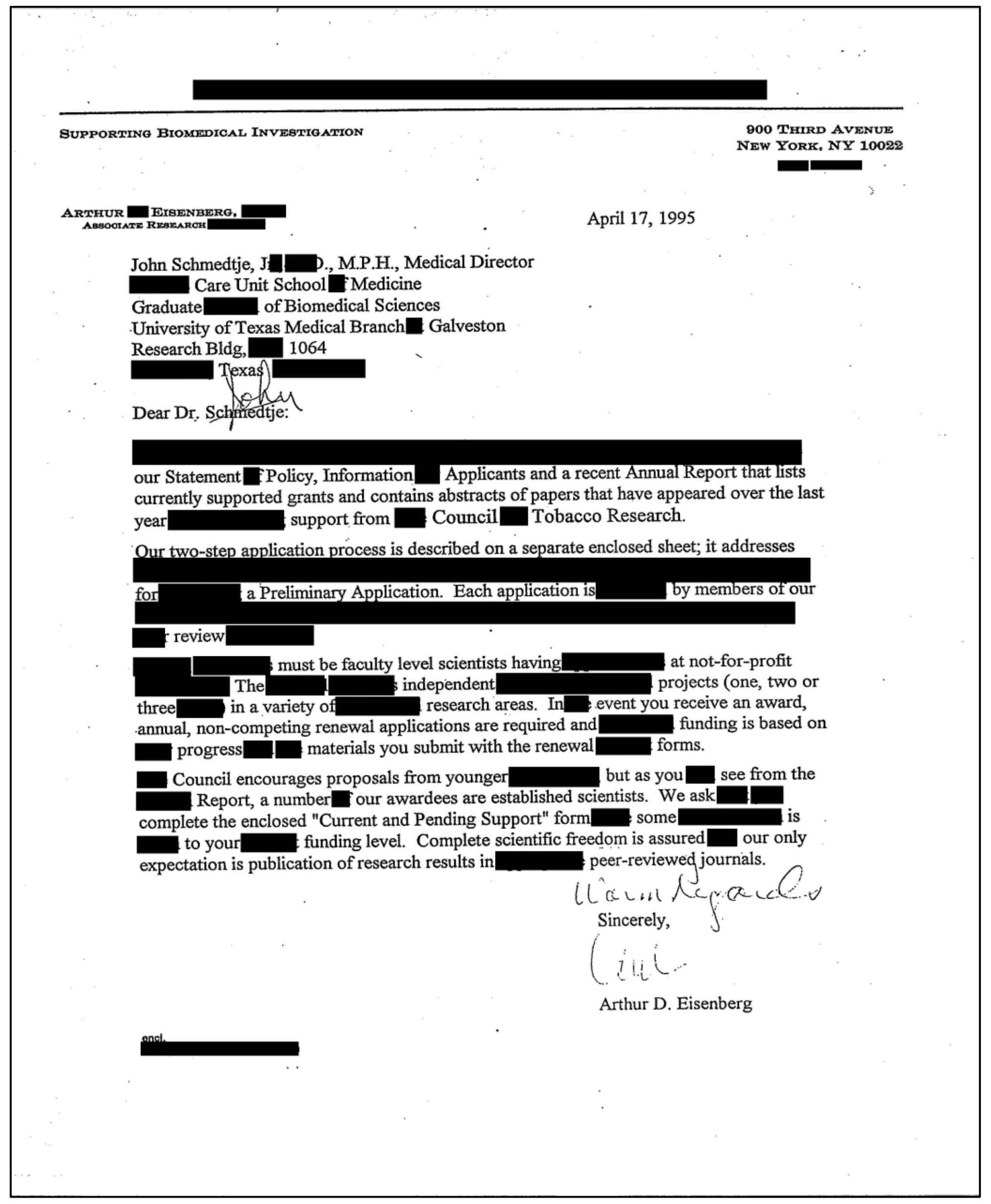
LayoutLMv2



[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]

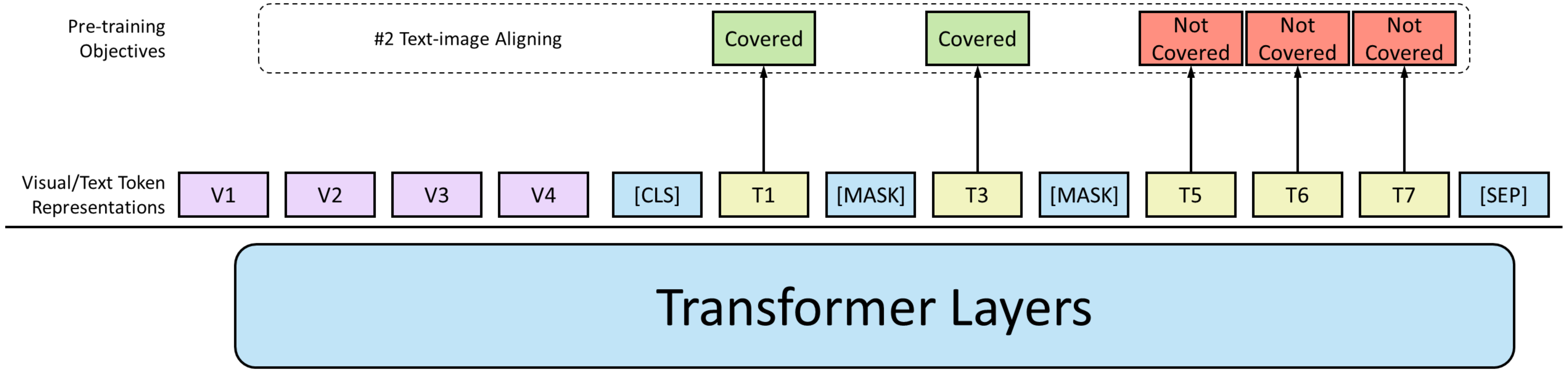
[redacted] must be faculty level scientists having [redacted] at not-for-profit [redacted] The [redacted] independent [redacted] projects (one, two or three [redacted] in a variety of [redacted] research areas. In [redacted] event you receive an award, annual, non-competing renewal applications are required 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.

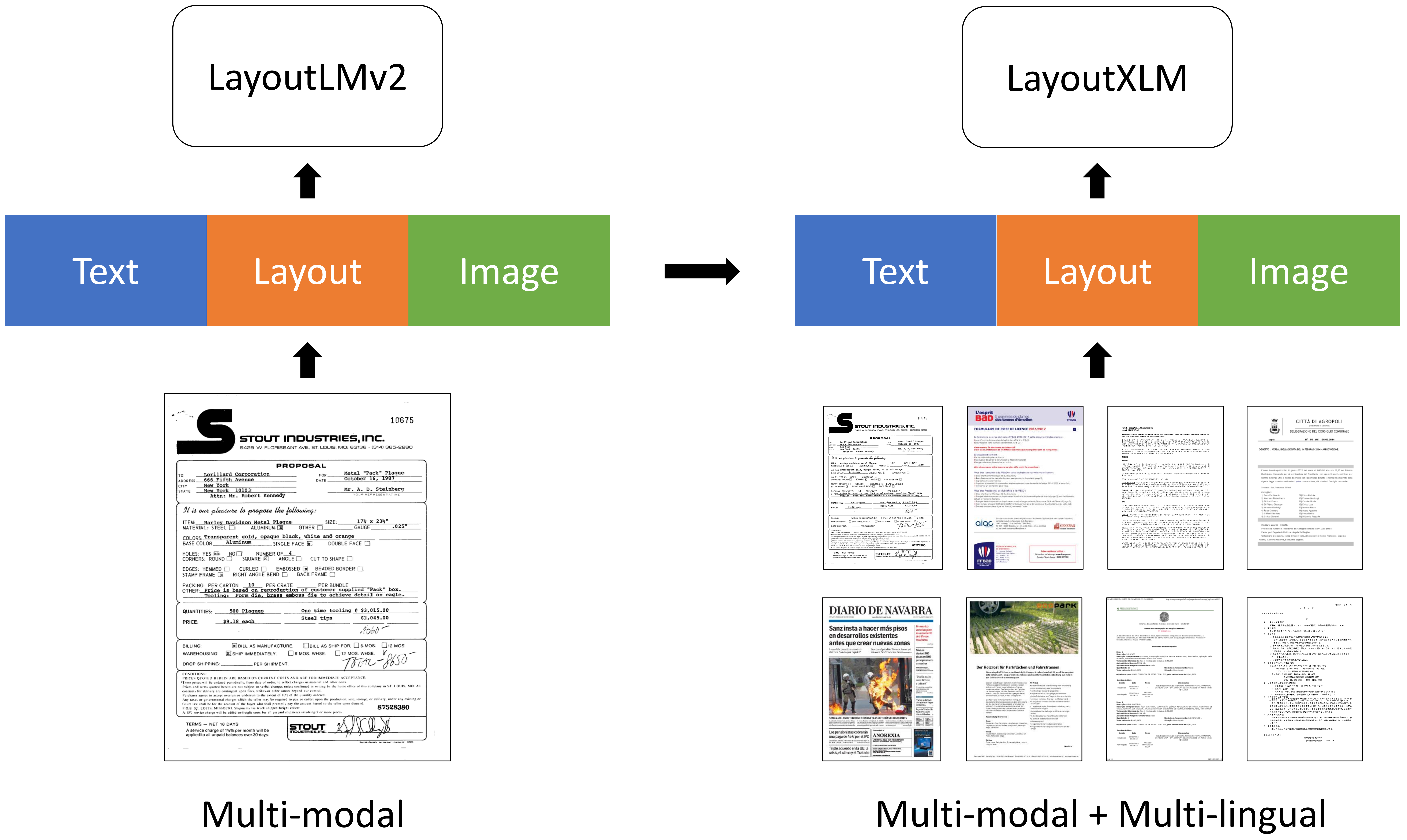
Sincerely,

Arthur D. Eisenberg

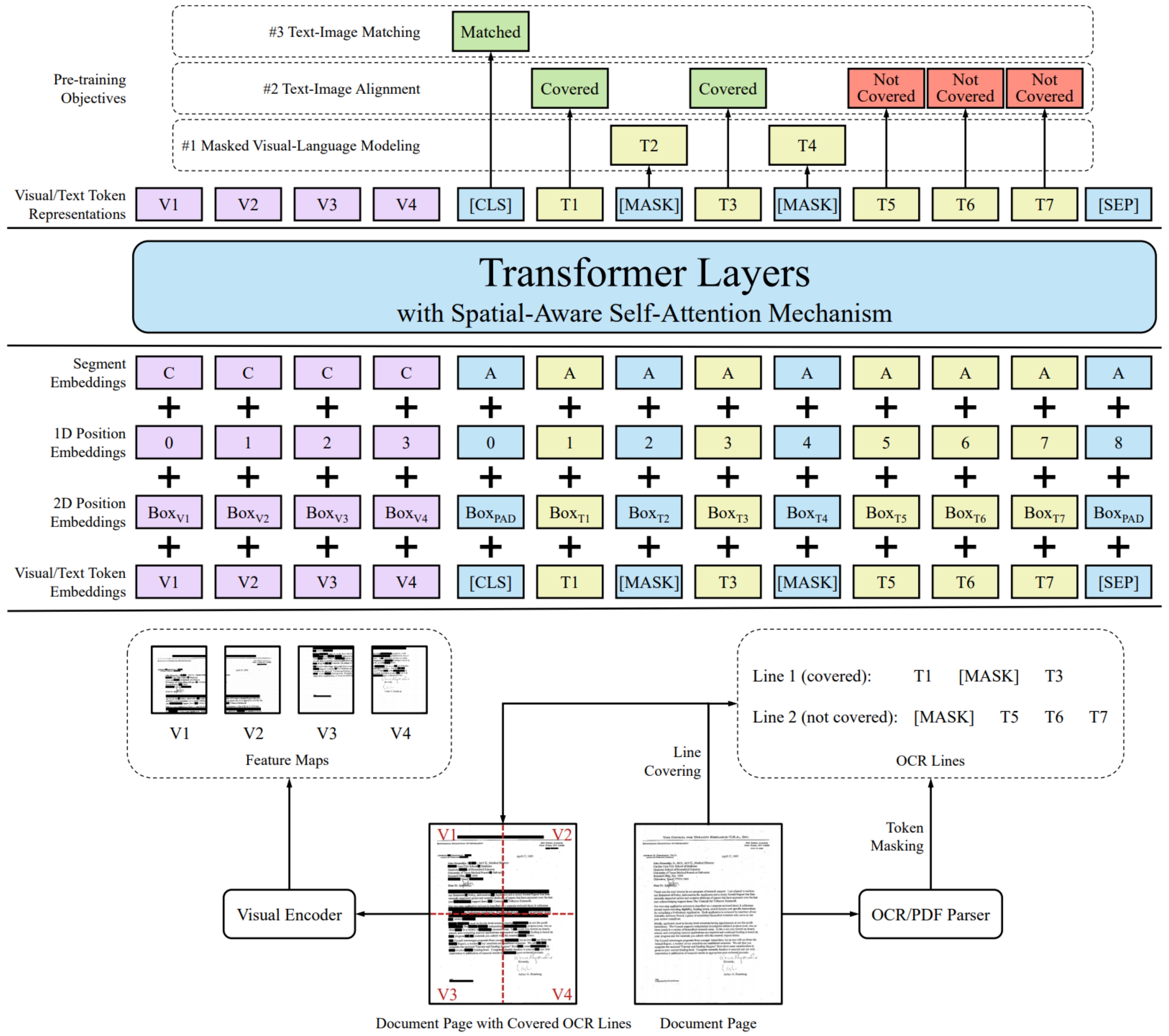
encl.
[redacted]



LayoutLMv2 -> LayoutXLM

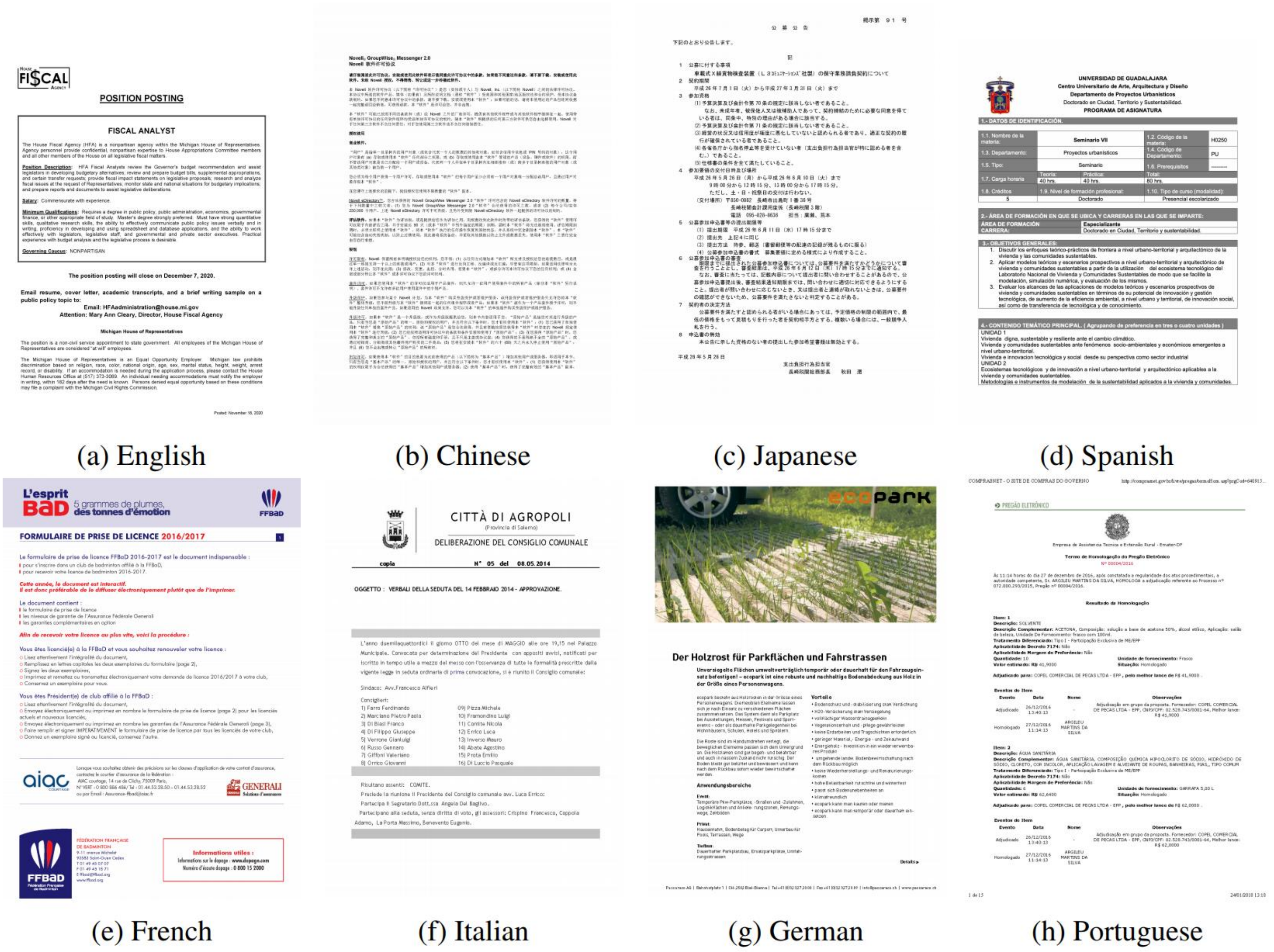


LayoutXLM



LayoutLMv2 + InfoXLM

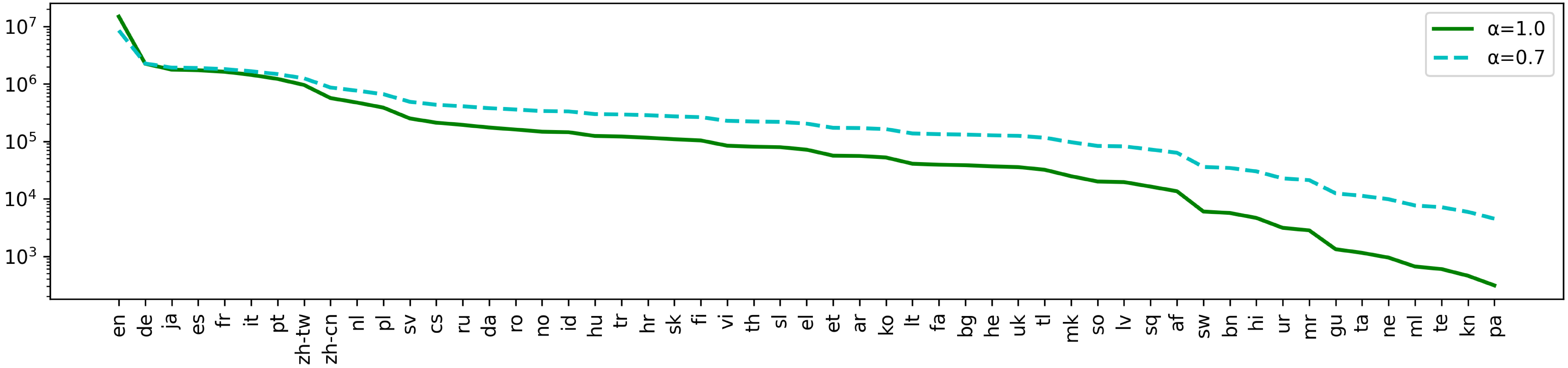
+



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***

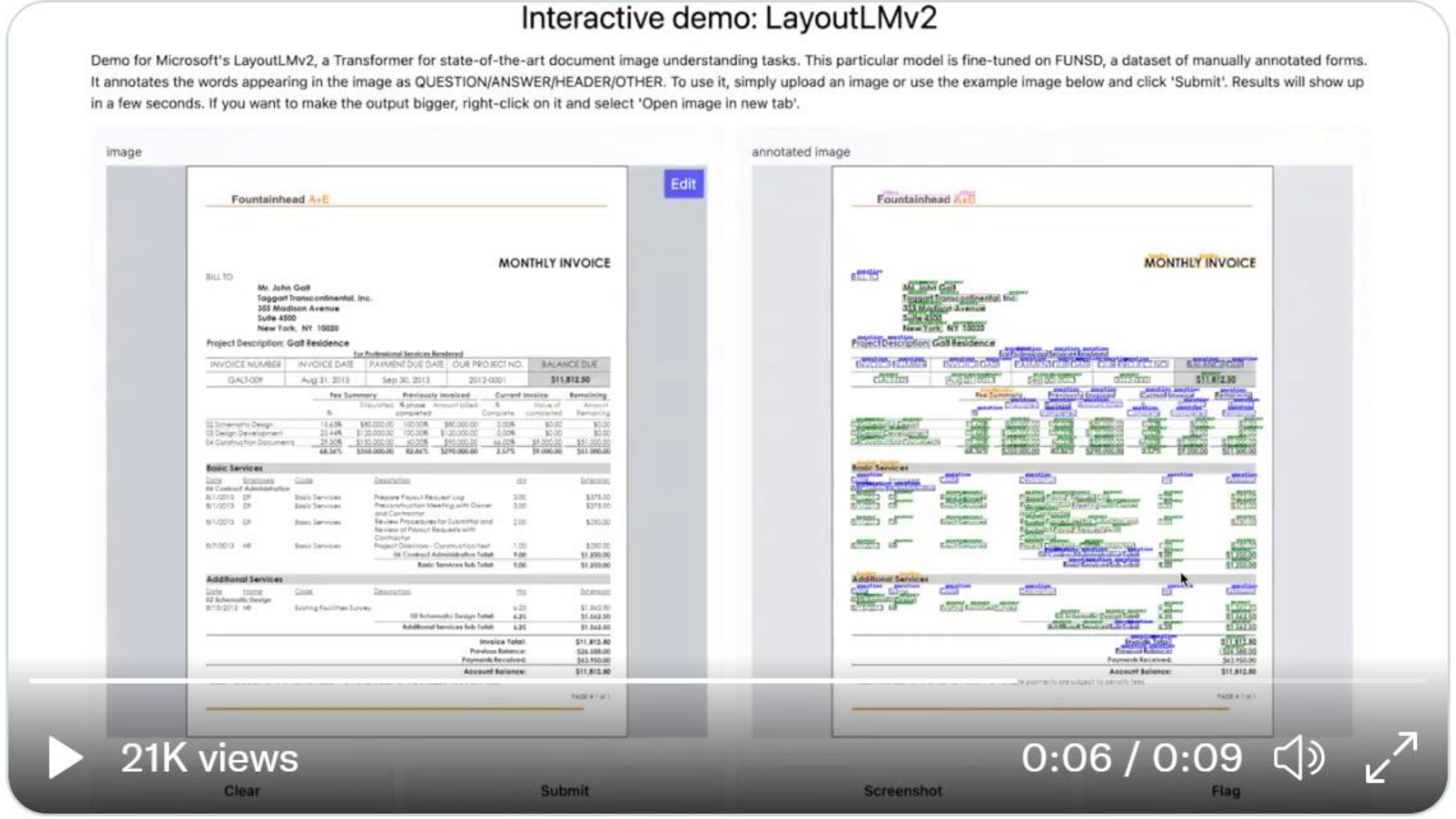
Hugging Face
@huggingface

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

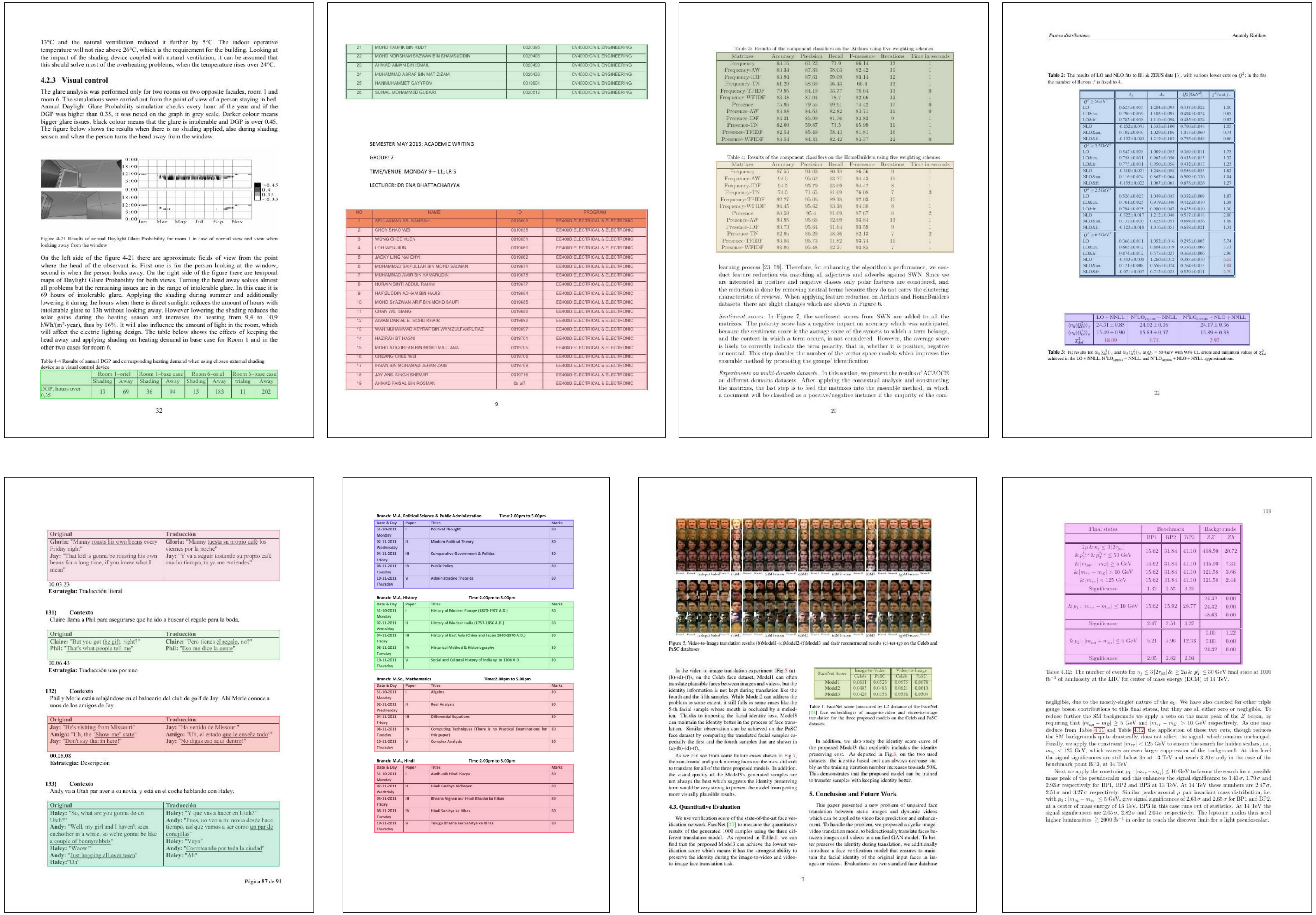
microsoft/layoutlm-base-uncased
Updated Aug 11 • 448k • ♥ 3

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

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

Benchmark Datasets

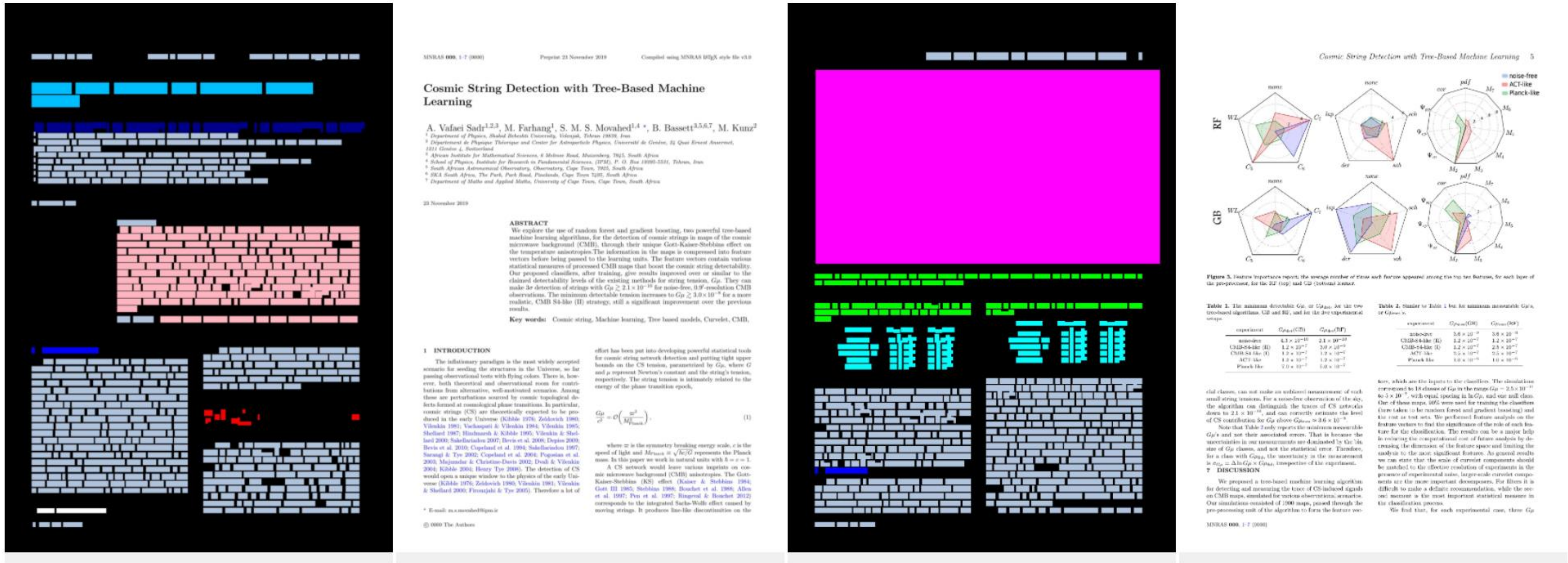
TableBank



[TableBank: A Benchmark Dataset for Table Detection and Recognition](), LREC'2020

Download: https://aka.ms/tablebank

DocBank

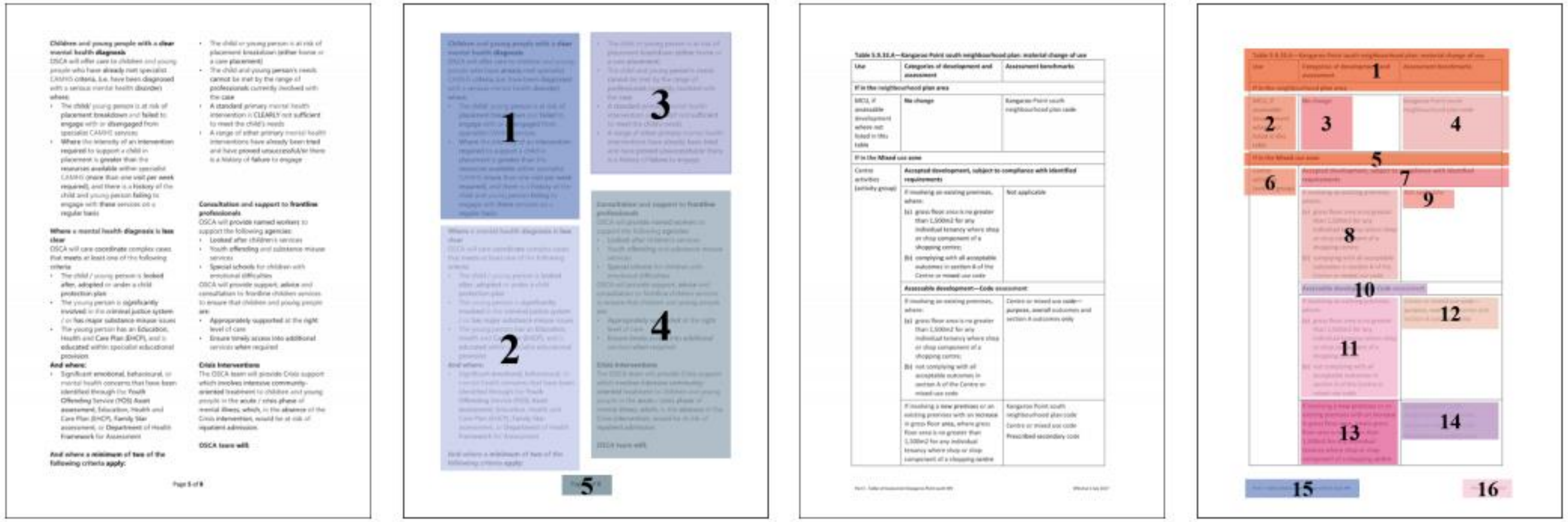


The colors of semantic structure labels are:

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

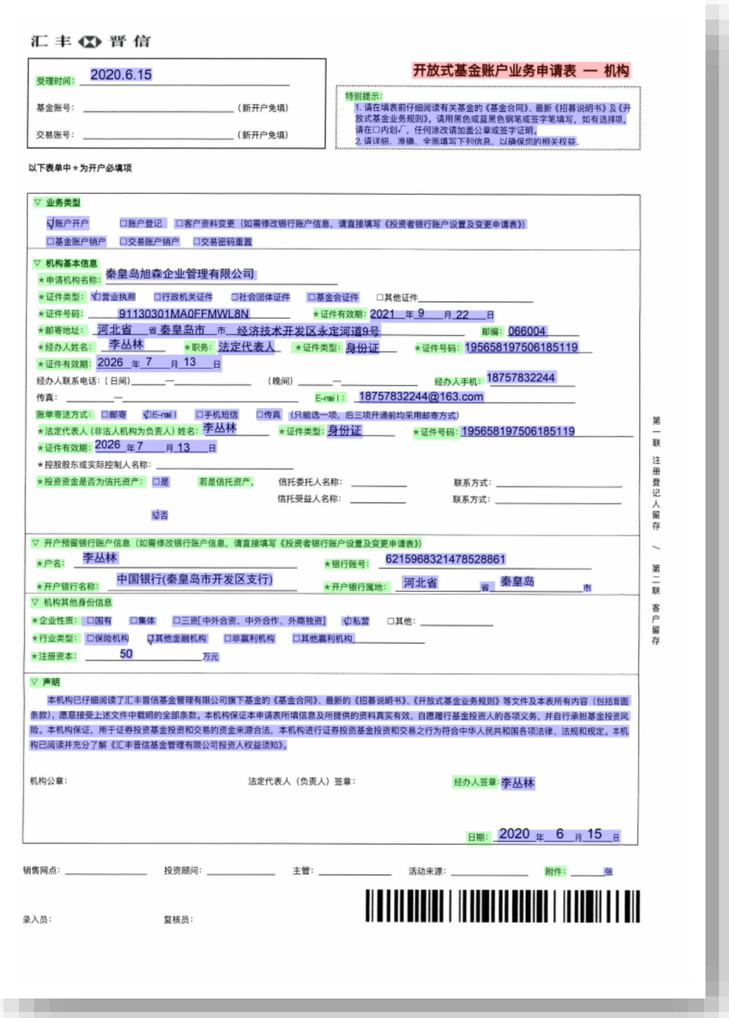
ReadingBank

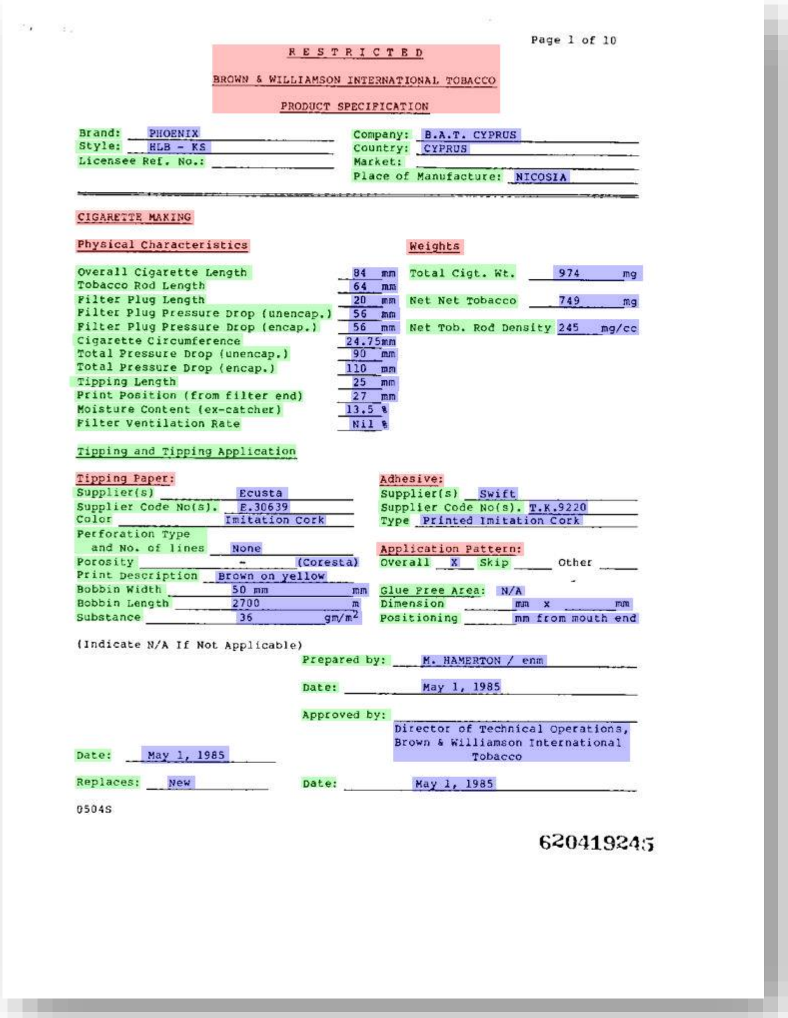


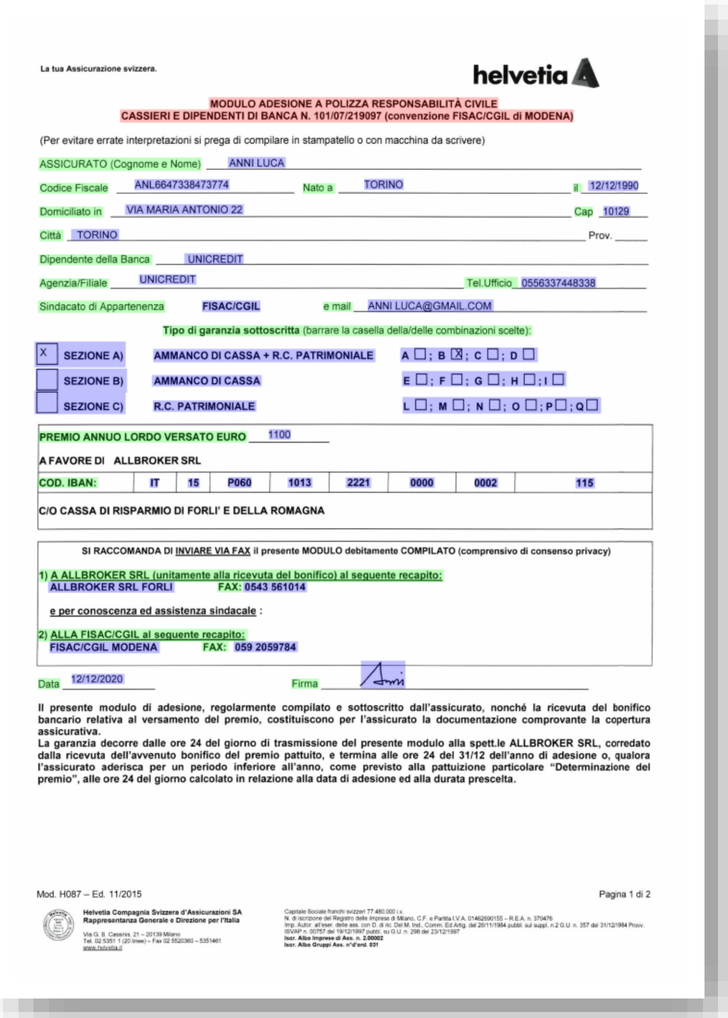
(a) (b) (c) (d)

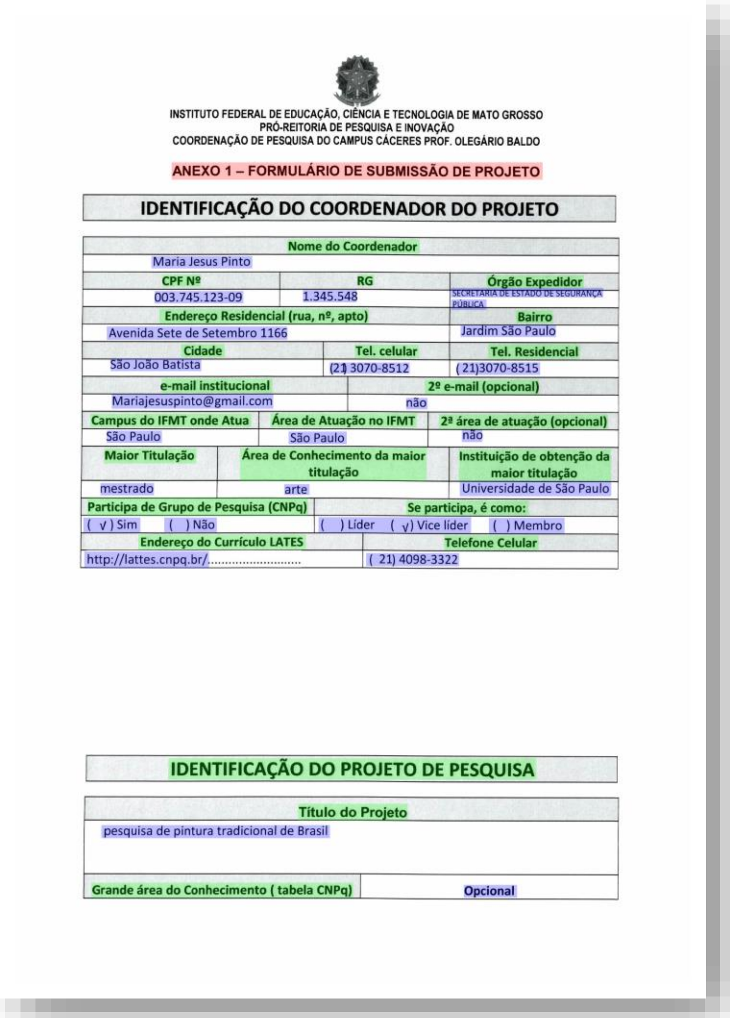
LayoutReader: Pre-training of Text and Layout for Reading Order Detection, EMNLP'2021
Download: https://aka.ms/readingbank

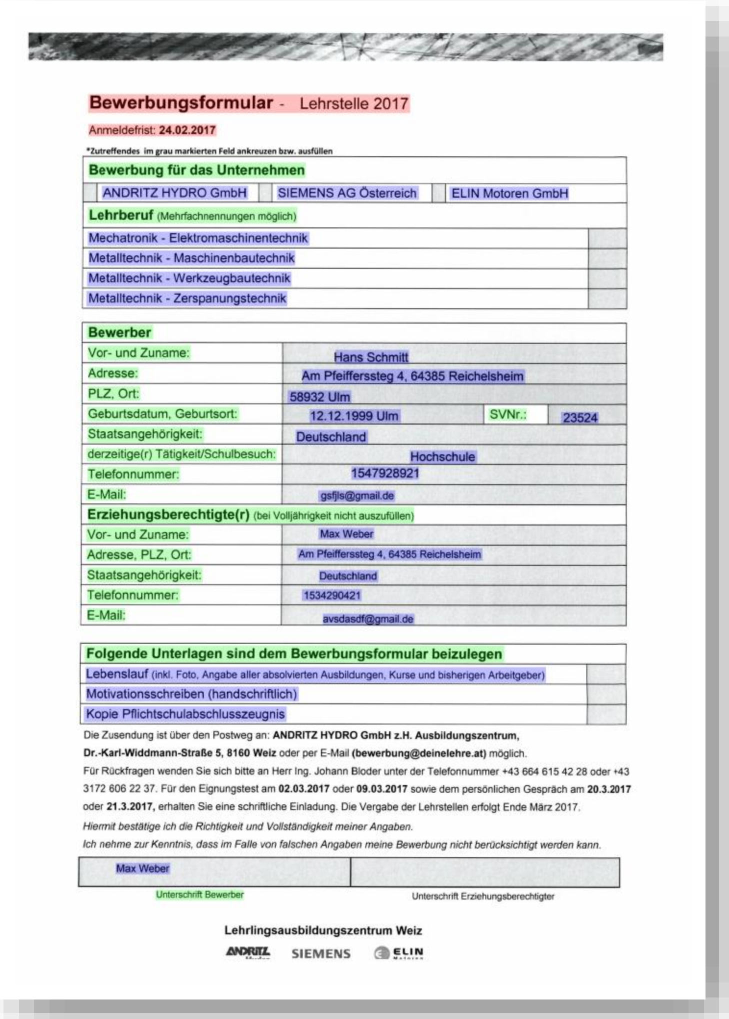
XFUND


Chinese

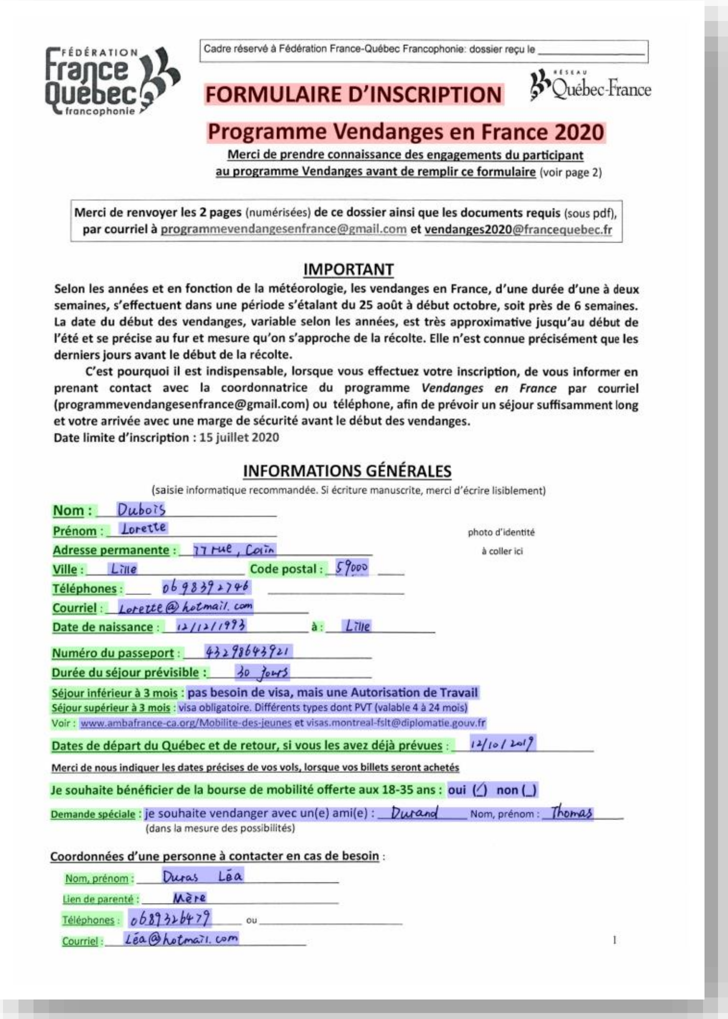

English


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
$BERT_{BASE}$	89.81%	110M
$UniLMv2_{BASE}$	90.06%	125M
$BERT_{LARGE}$	89.92%	340M
$UniLMv2_{LARGE}$	90.20%	355M
$LayoutLM_{BASE}$ (w/ image)	94.42%	160M
$LayoutLM_{LARGE}$ (w/ image)	94.43%	390M
$LayoutLMv2_{BASE}$	95.25%	200M
$LayoutLMv2_{LARGE}$	**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
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
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
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
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
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
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
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
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$_{BASE}$	0.667	0.8774	0.7761	0.6105	0.6743	0.6687	0.6814	0.6818	0.7047
	InfoXLM$_{BASE}$	0.6852	0.8868	0.7865	0.6230	0.7015	0.6751	0.7063	0.7008	0.7207
	LayoutXLM$_{BASE}$	**0.794**	**0.8924**	**0.7921**	**0.7550**	**0.7902**	**0.8082**	**0.8222**	**0.7903**	**0.8056**
	XLM-RoBERTa$_{LARGE}$	0.7074	0.8925	0.7817	0.6515	0.7170	0.7139	0.711	0.7241	0.7374
	InfoXLM$_{LARGE}$	0.7325	0.8955	0.7904	0.6740	0.7140	0.7152	0.7338	0.7212	0.7471
	LayoutXLM$_{LARGE}$	**0.8225**	**0.9161**	**0.8033**	**0.7830**	**0.8098**	**0.8275**	**0.8361**	**0.8273**	**0.8282**
RE	XLM-RoBERTa$_{BASE}$	0.2659	0.5105	0.5800	0.5295	0.4965	0.5305	0.5041	0.3982	0.4769
	InfoXLM$_{BASE}$	0.2920	0.5214	0.6000	0.5516	0.4913	0.5281	0.5262	0.4170	0.4910
	LayoutXLM$_{BASE}$	**0.5483**	**0.7073**	**0.6963**	**0.6896**	**0.6353**	**0.6415**	**0.6551**	**0.5718**	**0.6432**
	XLM-RoBERTa$_{LARGE}$	0.3473	0.6475	0.6798	0.6330	0.6080	0.6171	0.6189	0.5762	0.5910
	InfoXLM$_{LARGE}$	0.3679	0.6775	0.6604	0.6346	0.6096	0.6659	0.6057	0.5800	0.6002
	LayoutXLM$_{LARGE}$	**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$_{BASE}$	0.667	0.4144	0.3023	0.3055	0.371	0.2767	0.3286	0.3936	0.3824
	InfoXLM$_{BASE}$	0.6852	0.4408	0.3603	0.3102	0.4021	0.2880	0.3587	0.4502	0.4119
	LayoutXLM$_{BASE}$	**0.794**	**0.6019**	**0.4715**	**0.4565**	**0.5757**	**0.4846**	**0.5252**	**0.539**	**0.5561**
	XLM-RoBERTa$_{LARGE}$	0.7074	0.5205	0.3939	0.3627	0.4672	0.3398	0.418	0.4997	0.4637
	InfoXLM$_{LARGE}$	0.7325	0.5536	0.4132	0.3689	0.4909	0.3598	0.4363	0.5126	0.4835
	LayoutXLM$_{LARGE}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
RE	XLM-RoBERTa$_{BASE}$	0.2659	0.1601	0.2611	0.2440	0.2240	0.2374	0.2288	0.	
	InfoXLM$_{BASE}$	0.2920	0.2405	0.2851	0.2481	0.2454	0.2193	0.2027	0.	
	LayoutXLM$_{BASE}$	**0.5483**	**0.4494**	**0.4408**	**0.4708**	**0.4416**	**0.4090**	**0.3820**	**0.**	
	XLM-RoBERTa$_{LARGE}$	0.3473	0.2421	0.3037	0.2843	0.2897	0.2496	0.2617	0.	
	InfoXLM$_{LARGE}$	0.3679	0.3156	0.3364	0.3185	0.3189	0.2720	0.2953	0.	
	LayoutXLM$_{LARGE}$	**0.6404**	**0.5531**	**0.5696**	**0.5780**	**0.5615**	**0.5184**	**0.4890**	**0.**	

1996	0.2276
2049	0.2423
3685	**0.4388**
2333	0.2765
2554	0.3100
4795	**0.5487**

on X), where

Table 3: Zero-shot transfer accuracy (F1) on the XFUND dataset (fine-tuning on FUNSD, testing "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$_{BASE}$	0.6633	0.883	0.7786	0.6223	0.7035	0.6814	0.7146	0.6726	0.7149
	InfoXLM$_{BASE}$	0.6538	0.8741	0.7855	0.5979	0.7057	0.6826	0.7055	0.6796	0.7106
	LayoutXLM$_{BASE}$	**0.7924**	**0.8973**	**0.7964**	**0.7798**	**0.8173**	**0.821**	**0.8322**	**0.8241**	**0.8201**
	XLM-RoBERTa$_{LARGE}$	0.7151	0.8967	0.7828	0.6615	0.7407	0.7165	0.7431	0.7449	0.7502
	InfoXLM$_{LARGE}$	0.7246	0.8919	0.7998	0.6702	0.7376	0.7180	0.7523	0.7332	0.7534
	LayoutXLM$_{LARGE}$	**0.8068**	**0.9155**	**0.8216**	**0.8055**	**0.8384**	**0.8372**	**0.853**	**0.8650**	**0.8429**
RE	XLM-RoBERTa$_{BASE}$	0.3638	0.6797	0.6829	0.6828	0.6727	0.6937	0.6887	0.6082	0.6341
	InfoXLM$_{BASE}$	0.3699	0.6493	0.6473	0.6828	0.6831	0.6690	0.6384	0.5763	0.6145
	LayoutXLM$_{BASE}$	**0.6671**	**0.8241**	**0.8142**	**0.8104**	**0.8221**	**0.8310**	**0.7854**	**0.7044**	**0.7823**
	XLM-RoBERTa$_{LARGE}$	0.4246	0.7316	0.7350	0.7513	0.7532	0.7520	0.7111	0.6582	0.6896
	InfoXLM$_{LARGE}$	0.4543	0.7311	0.7510	0.7644	0.7549	0.7504	0.7356	0.6875	0.7037
	LayoutXLM$_{LARGE}$	**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

UniLM AI

Pre-trained (foundation) models across tasks (understanding, generation and translation), languages (100+ languages), and modalities (language, image, audio, vision + language, audio + language, etc.)

The family of UniLM AI:

- UniLM (v1@NeurIPS'19 | v2@ICML'20 | v3@ACL'21): unified pre-training for language understanding and generation
- InfoXLM (v1@NAACL'21 | v2@ACL'21): multilingual/cross-lingual pre-trained models for 100+ languages
- DeltaLM (NEW): encoder-decoder pre-training for language generation and translation for 100+ languages
- MiniLM (v1@NeurIPS'20 | v2@ACL'21): small and fast pre-trained models for language understanding and generation
- AdaLM (v1@ACL'21): domain, language, and task adaptation of pre-trained models
- LayoutLM (v1@KDD'20 | v2@ACL'21): multimodal (text + layout/format + image) pre-training for Document AI (e.g. scanned documents, PDF, etc.)
- LayoutXLM (NEW): multimodal (text + layout/format + image) pre-training for multilingual document understanding
- LayoutReader (EMNLP'21): Pre-training of text and layout for reading order detection
- BEiT (NEW): BERT Pre-Training of Image Transformers
- UniSpeech (v1@ICML'21): Speech Pre-Training for ASR and TTS
- s2s-ft: sequence-to-sequence fine-tuning toolkit
- XLM-T (NEW): Multilingual NMT w/ pretrained cross-lingual encoders

https://github.com/microsoft/unilm