



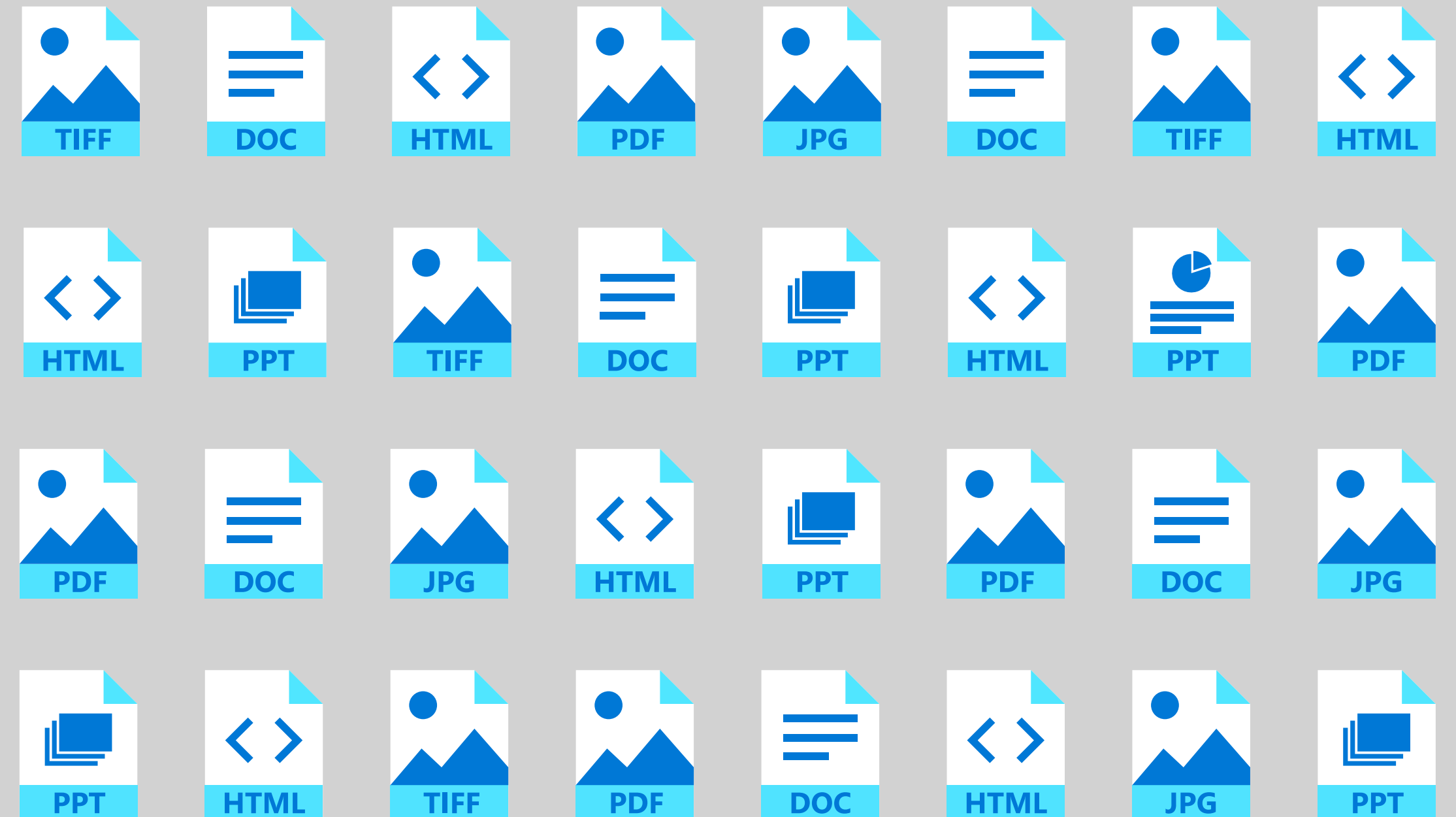
# ***Document AI: Benchmarks, Models and Applications***

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**)

DATE RECEIVED: **Unk.** TESTED: **12/28/78** REPORTED: **5/3/79** NO. **A4**

INVESTIGATOR(S): **H. S. Tong & M. S. Forte** (NOTEBOOK PAGE **BI014-23**)

SIGNATURE(S): *H.S. Tong* *M.S. Forte (by D. Poole)*

STRAIN OF MICE: **Swiss-Webster** MALE  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	DOSE (mg/kg BODY WEIGHT)	RESULTS (NO DEAD/NO 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 symptoms of respiratory depression, constriction of blood vessels, and in-activity. 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**

LORILLARD RESEARCH CENTER FORM 7 (5-80)

Form

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 6291458

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

Approval: 04702B

Amount: \$33.79  
+ Gratuity Not Inc: 6  
= Balance Due: 39.79

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

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

Guest Copy

Receipt

**UBS** Global Research 31 January 2018

**First Read**  
**Microsoft Corp.**  
**Azure Acceleration Impresses in Solid Q2**

**Equities**  
Americas Software Buy

12-month rating Buy

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

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

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 \$10.7.2B while FY19 goes from \$11.6.0B to \$11.7.3B. 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 EV/CY18 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.

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 (US\$)	28,172	27,188	29,331	32,724	38,812	45,469	52,923	61,117
Net earnings (US\$)	21,627	21,434	25,732	28,301	31,650	36,943	42,868	49,383
EPS (US\$, diluted)	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

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

Report

**Access** Information Protected. 10445 48th Avenue Denver, CO 80228

Page 1 of 1  
**Invoice**  
1 877 FileLine | InformationProtected.com

New Belgium Brewery Company  
Attn: Accounts Payable Manager  
520 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: \_\_\_\_\_

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
<b>Storage</b>						
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
			<b>TOTAL FOR Storage</b>			<b>522.69</b>
			<b>TAX</b>			<b>0.00</b>
<b>Service</b>						
		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 Refill	4.00	6.0000	N	24.00
		FileBRIDGE Records + AccessMETRICS	1.00	0.0000	N	0.00
			<b>TOTAL FOR Service</b>			<b>24.00</b>
			<b>TAX</b>			<b>0.00</b>
<b>Transportation</b>						
		Shred Rotation Transportation - Scheduled trip	2.00	0.0000	N	0.00
			<b>TOTAL FOR Transportation</b>			<b>0.00</b>
			<b>TAX</b>			<b>0.00</b>
			<b>SUB-TOTAL</b>			<b>546.69</b>
			<b>TAX</b>			<b>0.00</b>
			<b>INVOICE TOTAL</b>			<b>\$546.69</b>

PLEASE NOTE: To the extent you do not have a currently effective written contract for services with an Access or Retrieval company, by paying this invoice, you agree that the terms and conditions found on <http://www.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 Retrieval 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 Retrieval company, by paying this invoice, you agree that the terms and conditions found on [www.informationprotected.com/baa](http://www.informationprotected.com/baa) constitute a legally effective BAA or BASA, as applicable, between you and such Access or Retrieval 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 Retrieval are not binding.

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	



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/>

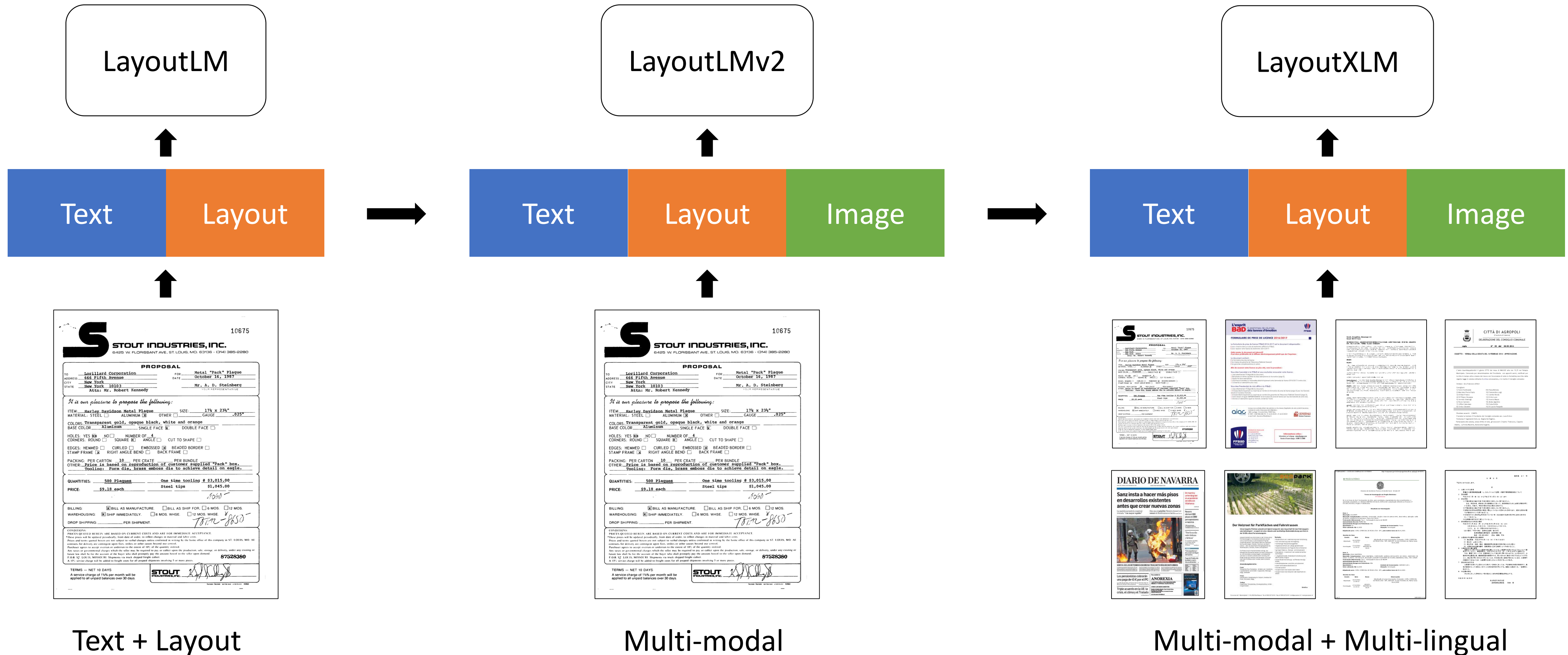


# LayoutLM Model Family

---



# LayoutLM -> LayoutLMv2 -> LayoutXLM

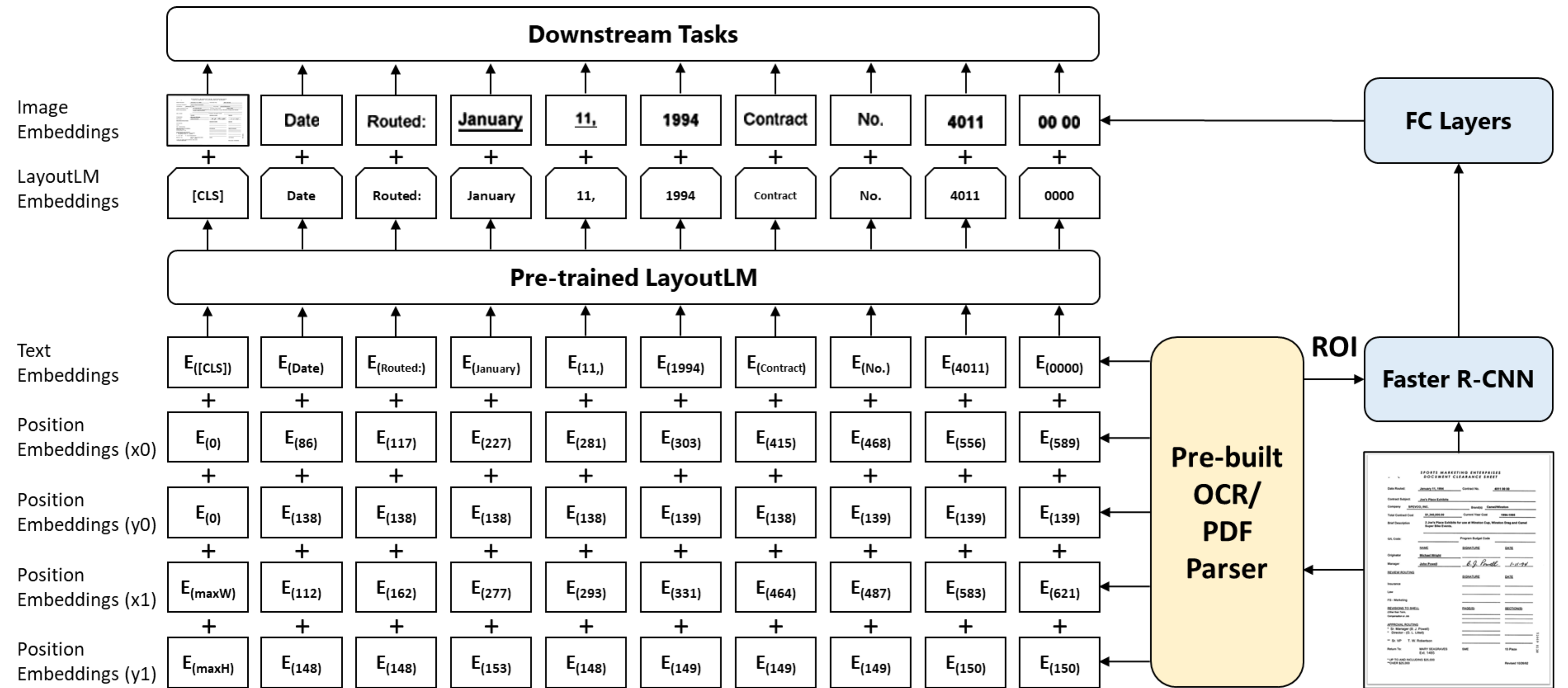


Text + Layout

Multi-modal

Multi-modal + Multi-lingual

# LayoutLM



\* Text embeddings initialized by BERT/UniLM



**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 \_\_\_\_\_

	<u>NAME</u>	<u>SIGNATURE</u>	<u>DATE</u>
Originator	<u>Michael Wright</u>	_____	_____
Manager	<u>John Powell</u>	<u>B. J. Powell</u>	<u>1-11-94</u>

**REVIEW ROUTING**

	<u>SIGNATURE</u>	<u>DATE</u>
Insurance	_____	_____
Law	_____	_____
FS - Marketing	_____	_____

**REVISIONS TO SHELL**  
(Other than Term, Compensation or Job)

	<u>PAGE(S)</u>	<u>SECTION(S)</u>
_____	_____	_____
_____	_____	_____

**APPROVAL ROUTING**

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

\*\* Sr. VP T. W. Robertson

Return To: MARY SEAGRAVES SME 13 Plaza  
Ext. 1485

\* 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
<u>January</u>	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

↓

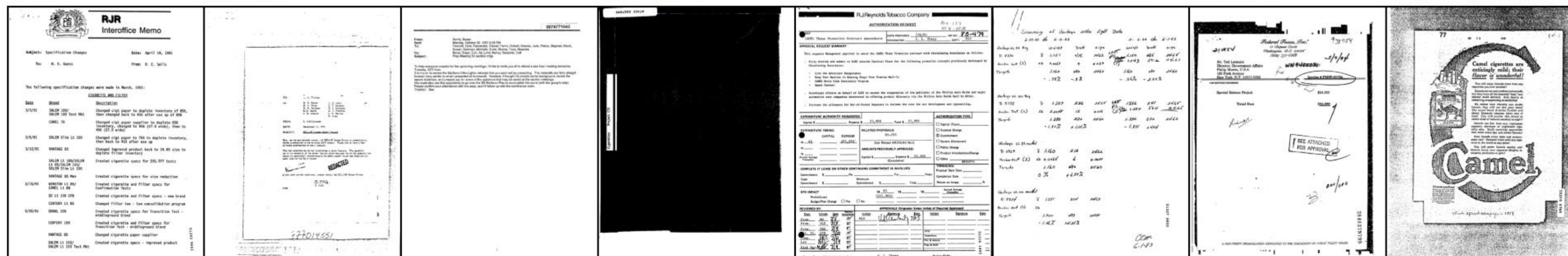
Input

Tok 0 Tok 1 Tok 2 Tok 3 Tok 4 Tok 5 Tok 6 Tok 7

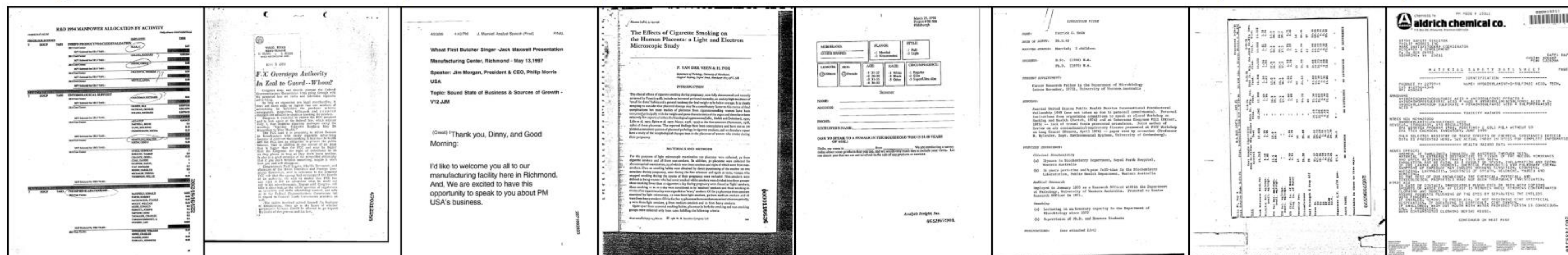
Token Embeddings	$E_{(Tok 0)}$	$E_{(Tok 1)}$	$E_{(Tok 2)}$	$E_{(Tok 3)}$	$E_{(Tok 4)}$	$E_{(Tok 5)}$	$E_{(Tok 6)}$	$E_{(Tok 7)}$
	+	+	+	+	+	+	+	+
Position Embeddings (x0)	$E_{(Tok 0)}$	$E_{(Tok 1)}$	$E_{(Tok 2)}$	$E_{(Tok 3)}$	$E_{(Tok 4)}$	$E_{(Tok 5)}$	$E_{(Tok 6)}$	$E_{(Tok 7)}$
	+	+	+	+	+	+	+	+
Position Embeddings (y0)	$E_{(Tok 0)}$	$E_{(Tok 1)}$	$E_{(Tok 2)}$	$E_{(Tok 3)}$	$E_{(Tok 4)}$	$E_{(Tok 5)}$	$E_{(Tok 6)}$	$E_{(Tok 7)}$
	+	+	+	+	+	+	+	+
Position Embeddings (x1)	$E_{(Tok 0)}$	$E_{(Tok 1)}$	$E_{(Tok 2)}$	$E_{(Tok 3)}$	$E_{(Tok 4)}$	$E_{(Tok 5)}$	$E_{(Tok 6)}$	$E_{(Tok 7)}$
	+	+	+	+	+	+	+	+
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

letter memo email filefolder form handwritten invoice advertisement



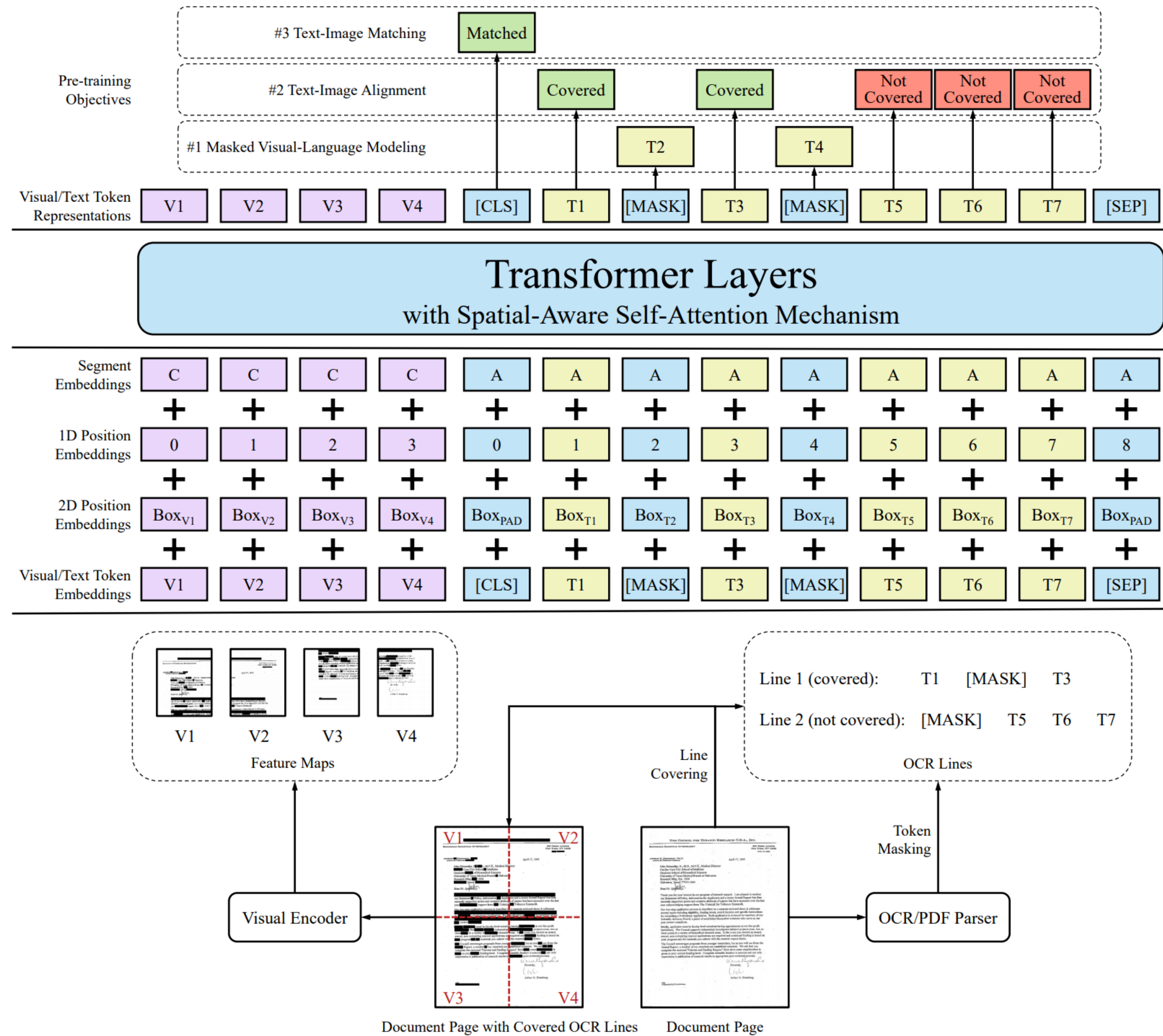
budget news article presentation scientific publication questionnaire resume scientific report specification



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

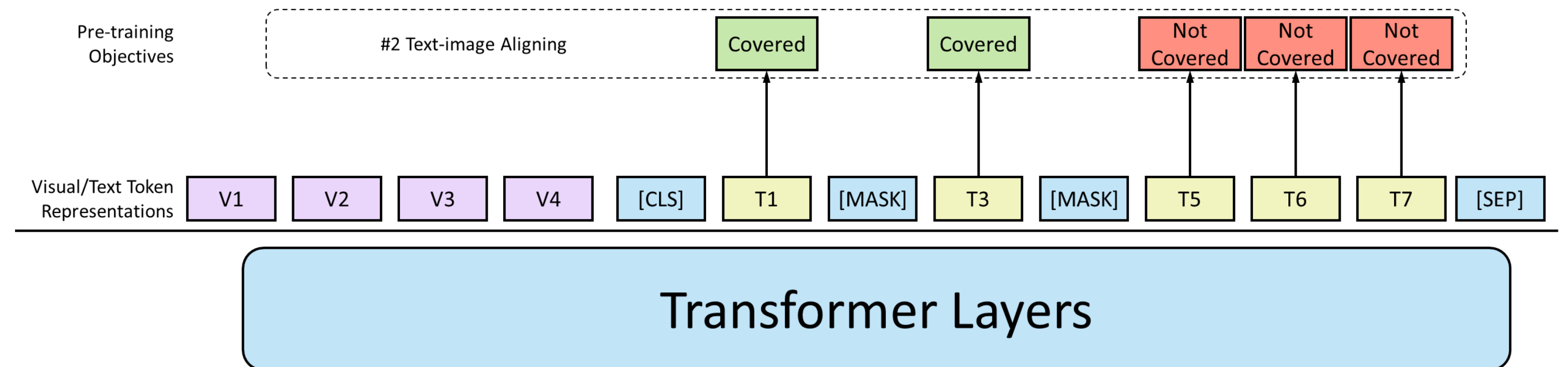
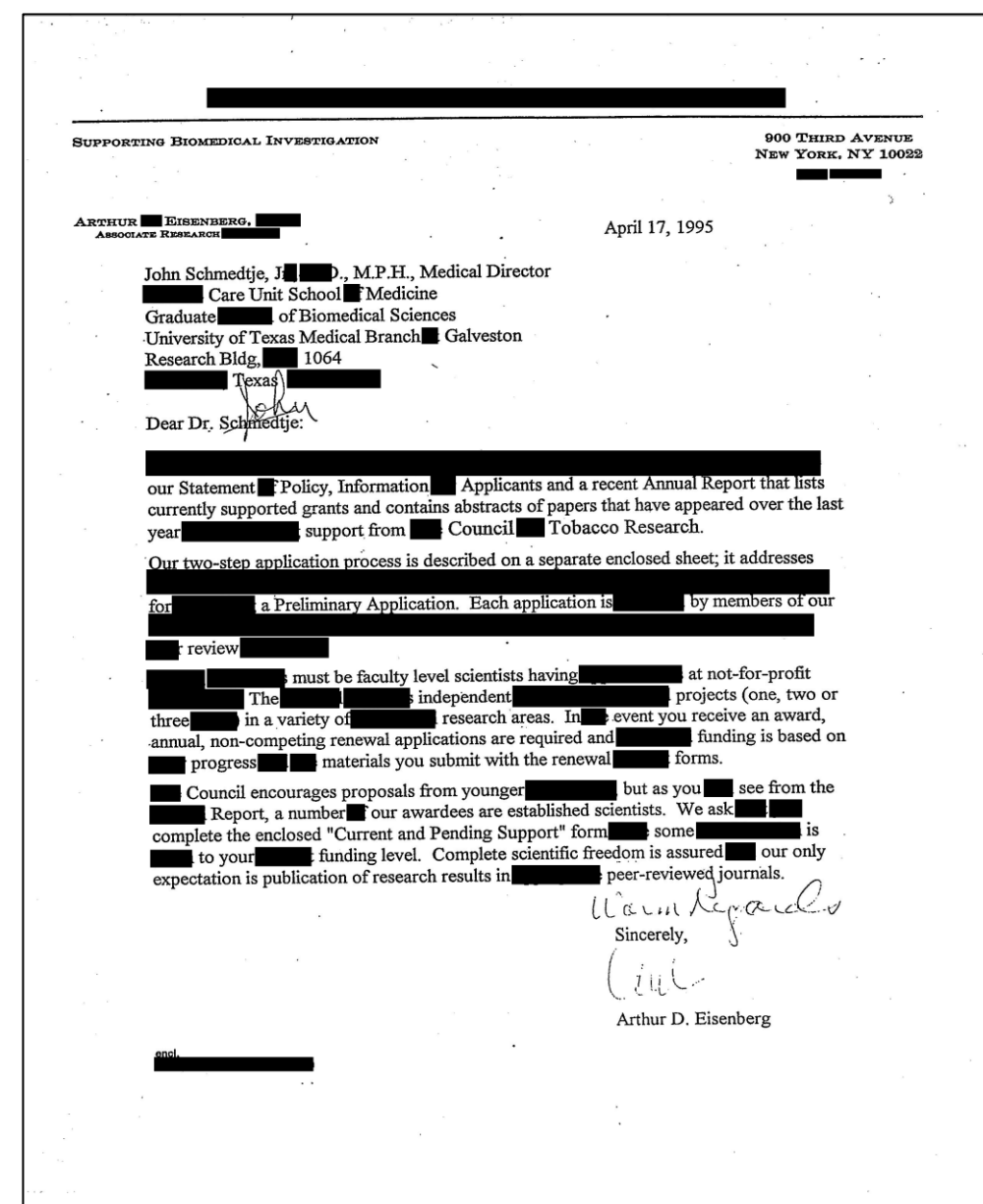
<https://ir.nist.gov/cdip/>

# LayoutLMv2

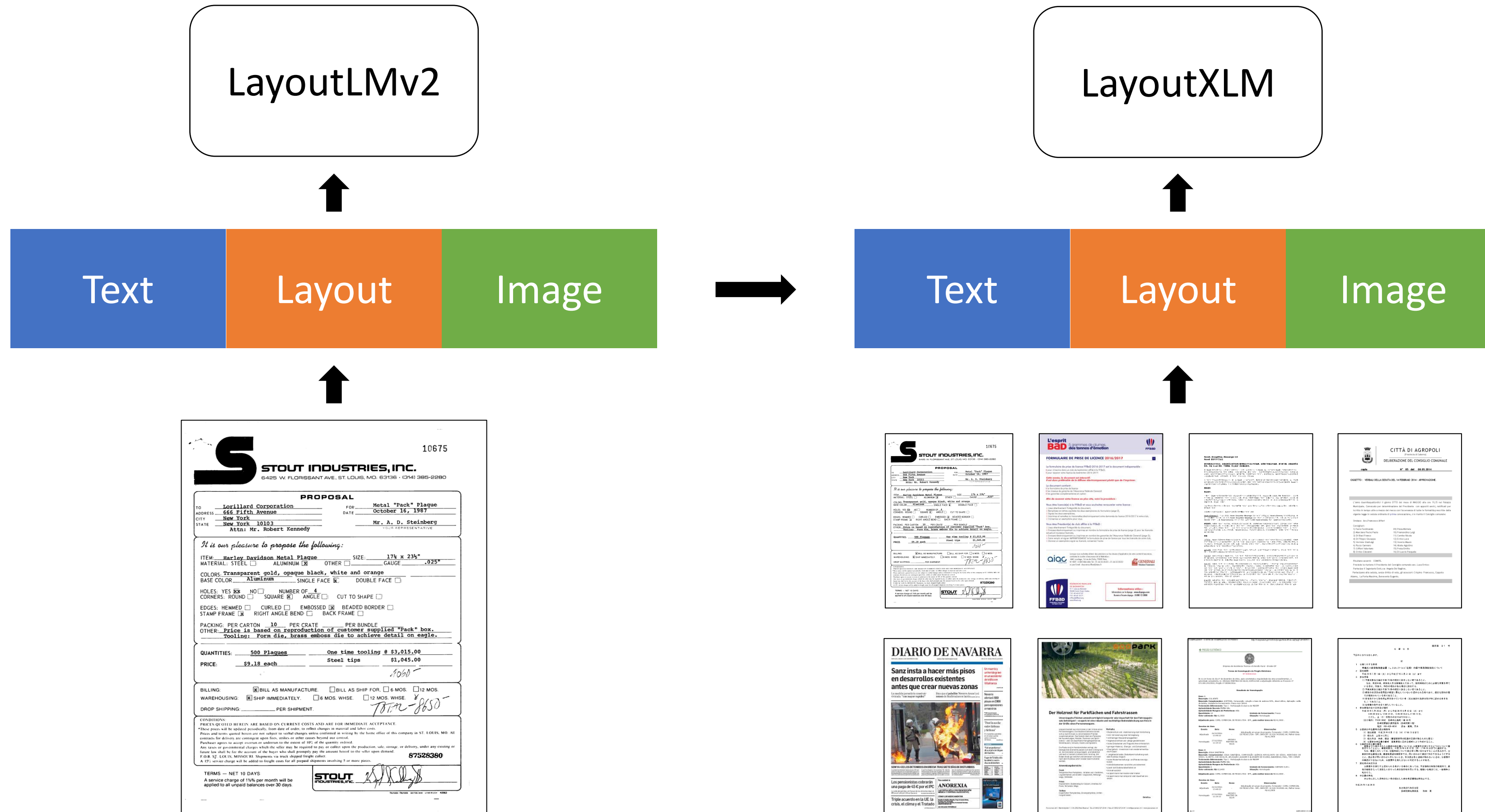


# Pre-training Tasks

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



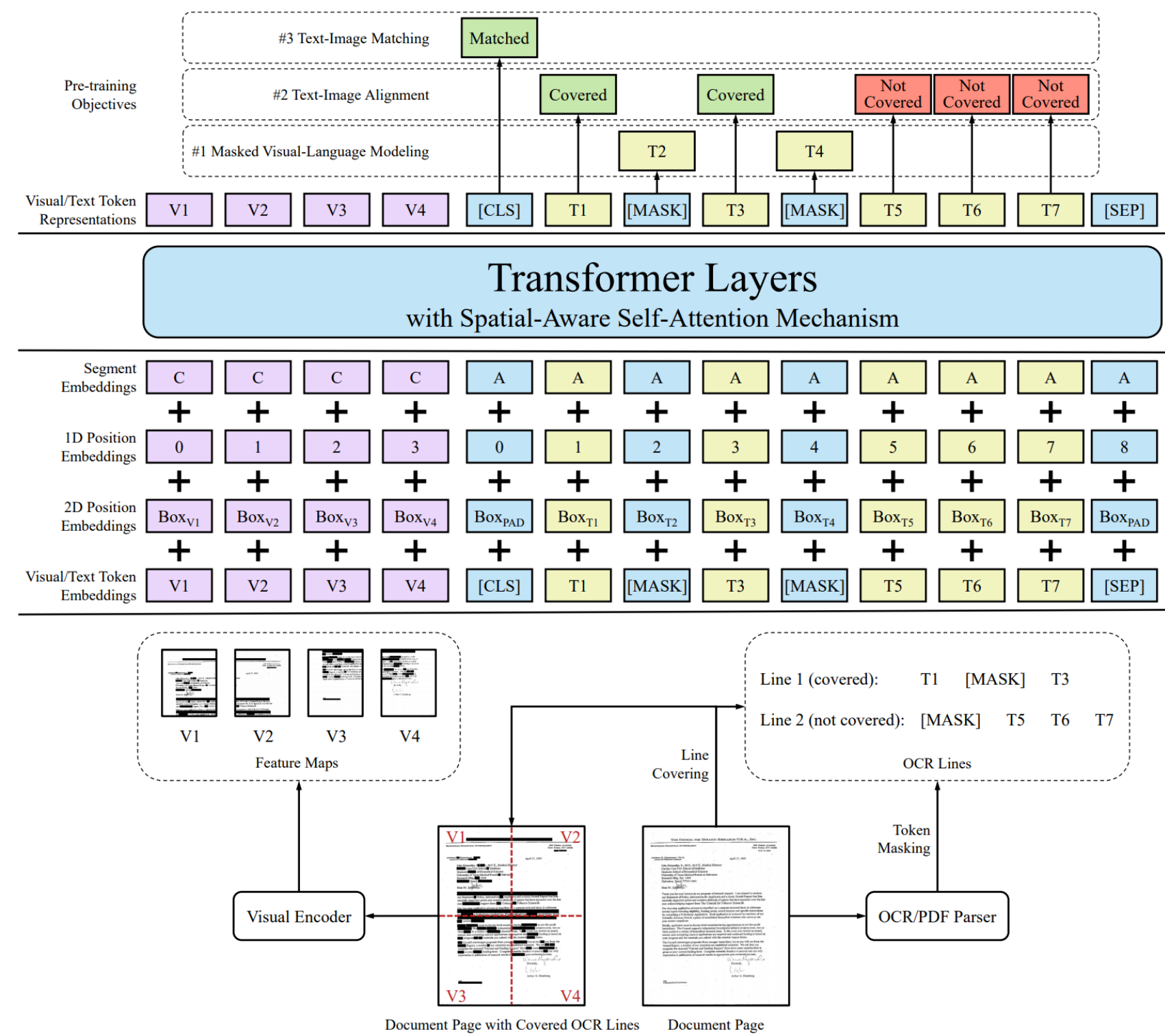
# LayoutLMv2 -> LayoutXLM



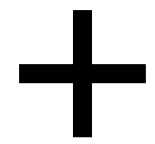
Multi-modal

Multi-modal + Multi-lingual

# LayoutXLM



## LayoutLMv2 + InfoXLM



**(a) English:** FISCAL ANALYST POSITION POSTING

**(b) Chinese:** CITTÀ DI AGROPOLI DELIBERAZIONE DEL CONSIGLIO COMUNALE

**(c) Japanese:** Der Holzrost für Parkflächen und Fahrradstraßen

**(d) Spanish:** UNIVERSIDAD DE GADALUQUÍA

**(e) French:** L'esprit BAD

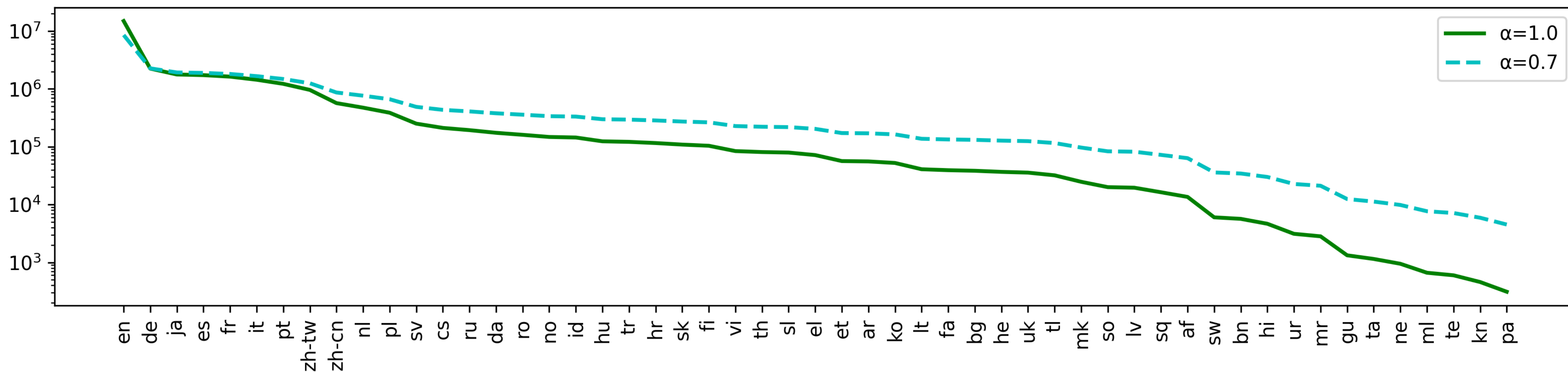
**(f) Italian:** CITTÀ DI AGROPOLI DELIBERAZIONE DEL CONSIGLIO COMUNALE

**(g) German:** Der Holzrost für Parkflächen und Fahrradstraßen

**(h) Portuguese:** UNIVERSIDADE DE ALGARVES

## Multi-lingual documents (50+ languages)

# 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/...](https://huggingface.co/spaces/nielsr/...)

Interactive demo: LayoutLMv2

Demo for Microsoft's LayoutLMv2, a Transformer for state-of-the-art document image understanding tasks. This particular model is fine-tuned on FUNSD, a dataset of manually annotated forms. It annotates the words appearing in the image as QUESTION/ANSWER/HEADER/OTHER. To use it, simply upload an image or use the example image below and click 'Submit'. Results will show up in a few seconds. If you want to make the output bigger, right-click on it and select 'Open image in new tab'.

image

annotated image

21K views | 0:06 / 0:09 | Clear | Submit | Screenshot | Flag

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

---





13°C and the natural ventilation reduced it further by 5°C. The indoor operative temperature will not rise above 26°C, which is the requirement for the building. Looking at the impact of the shading device coupled with natural ventilation, it can be assumed that this should solve most of the overheating problems, when the temperature rises over 24°C.

4.2.3 Visual control

The glare analysis was performed only for two rooms on two opposite facades, room 1 and room 6. The simulations were carried out from the point of view of a person staying in bed. Annual Daylight Glare Probability simulation checks every hour of the year and if the DGP was higher than 0.35, it was noted on the graph in grey scale. Darker colour means bigger glare issues, black colour means that the glare is intolerable and DGP is over 0.45. The figure below shows the results when there is no shading applied, also during shading season and when the person turns the head away from the window.

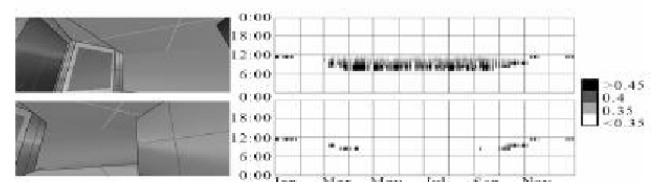


Figure 4-21 Results of annual Daylight Glare Probability for room 1 in case of normal view and view when head away from the window

On the left side of the figure 4-21 there are approximate fields of view from the point where the head of the observer is. First one is for the person looking at the window, second is when the person looks away. On the right side of the figure there are temporal maps of Daylight Glare Probability for both views. Turning the head away solves almost all problems but the remaining issues are in the range of intolerable glare. In this case it is 69 hours of intolerable glare. Applying the shading during summer and additionally lowering it during the hours when there is direct sunlight reduces the amount of hours with intolerable glare to 13h without looking away. However lowering the shading reduces the solar gains during the heating season and increases the heating from 9.4 to 10.9 kWh/(m²·year), thus by 16%. It will also influence the amount of light in the room, which will affect the electric lighting design. The table below shows the effects of keeping the head away and applying shading on heating demand in base case for Room 1 and in the other two cases for room 6.

Table 4-4 Results of annual DGP and corresponding heating demand when using chosen external shading device as a visual control device

Table 3: Results of the component classifiers on the Airbus using five weighting schemes

SEMESTER MAY 2015: ACADEMIC WRITING

GROUP: 7
TIME/VENUE: MONDAY 9 – 11; LR 5
LECTURER: DR ENA BHATTACHARYYA

Table 4: Results of the component classifiers on the HomeBuilders using five weighting schemes

Table 5: Results of the component classifiers on the HomeBuilders using five weighting schemes

Table 6: Results of the component classifiers on the HomeBuilders using five weighting schemes

Learning process [23, 29]. Therefore, for enhancing the algorithm's performance, we conduct feature reduction via matching all adjectives and adverbs against SWN. Since we are interested in positive and negative classes only polar features are considered, and the reduction is done by removing neutral terms because they do not carry the clustering characteristics of reviews. When applying feature reduction on Airbnos and HomeBuilders datasets, there are slight changes which are shown in Figure 6.

Sentiment scores. In Figure 7, the sentiment scores from SWN are added to all the matrices. The polarity score has a negative impact on accuracy which was anticipated because the sentiment score is the average score of the reviews to which a term belongs, and the context in which a term occurs, is not considered. However, the average score is likely to correctly indicate the term polarity, that is, whether it is positive, negative or neutral. This step doubles the number of the vector space models which improves the ensemble method by promoting the groups' identification.

Experiments on multi-domain datasets. In this section, we present the results of ACACCE on different domains datasets. After applying the contextual analysis and constructing the matrices, the last step is to feed the matrices into the ensemble method, in which a document will be classified as a positive/negative instance if the majority of the com-

Table 2: The results of LO and NLO fits to H1 & ZEUS data [31], with various lower cuts on Q² in the fit

Table 2: Matrix with columns for Q² (GeV²), A₀, A₁, (A₀/A₁)%, and χ²/ndof for various LO and NLO fits.

Table 3: Fit results for various Q² and s values, showing minimum values of Z².

Table 3: Fit results for various Q² and s values, showing minimum values of Z².

Table with 2 columns: Original and Traducción. Text about Gloria's coffee and Jay's roasting beans.

Table with 2 columns: Original and Traducción. Text about Claire's gift and Phil's response.

Table with 2 columns: Original and Traducción. Text about Missoum's visit and Amigo's comment.

Table with 2 columns: Original and Traducción. Text about Haley's visit to Utah and her conversation with Andy.

Branch: M.A. Political Science & Public Administration. Table of exam dates, papers, titles, and marks.

Branch: M.Sc., Mathematics. Table of exam dates, papers, titles, and marks.

Branch: M.A. Hindi. Table of exam dates, papers, titles, and marks.



Figure 5: Video-to-image translation results and their reconstructed results on the Celeb and PaSC datasets

Table with 4 columns: FaceNet Score, Image-to-Video, Video-to-Image, and FaceNet Score. Rows for Model1, Model2, and Model3.

Table 1: FaceNet score measured by L2 distance of the FaceNet [31] face embeddings of image-to-video and video-to-image translation for the three proposed models on the Celeb and PaSC datasets.

In the video-to-image translation experiment (Fig. 5 (a)-(b)-(d)-(f)), on the Celeb face dataset, Model1 can often translate plausible faces between images and videos, but the identity information is not kept during translation like the fourth and the fifth samples. While Model2 can address the problem to some extent, it still fails in some cases like the 5-th facial sample whose mouth is occluded by a moustache. Thanks to imposing the facial identity loss, Model3 can maintain the identity better in the process of face translation. Similar observations can be achieved on the PaSC face dataset by comparing the translated facial samples especially the first and the fourth samples that are shown in (a)-(b)-(d)-(f).

As we can see from some failure cases shown in Fig. 7, the non-neutral and quick moving faces are the most difficult to translate for all of the three proposed models. In addition, the visual quality of the Model3's generated samples are not always the best which suggests the identity preserving term would be very strong to prevent the model from getting more visually plausible results.

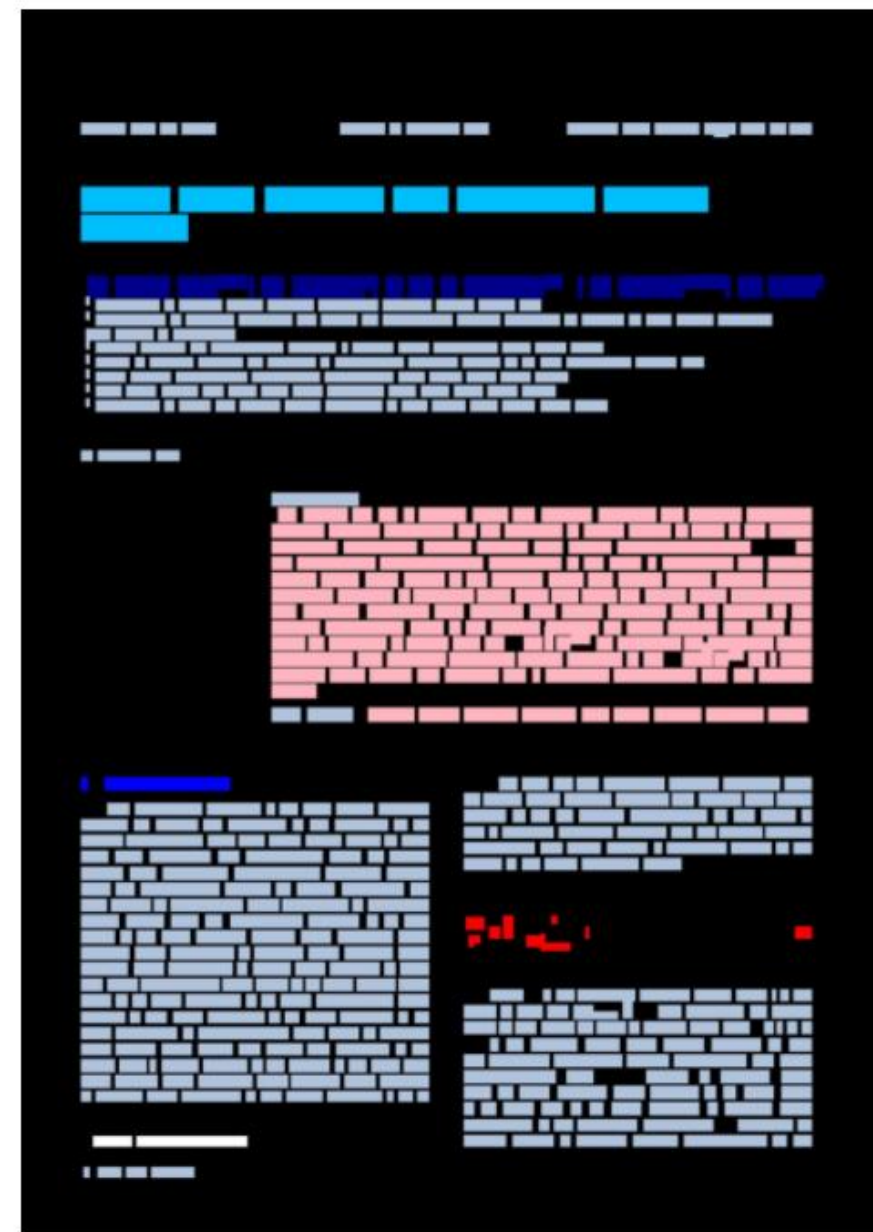
4.3. Quantitative Evaluation
We use verification scores of the state-of-the-art face verification network FaceNet [31] to measure the quantitative results of the generated 1000 samples using the three different translation models. As reported in Table 1, we can find that the proposed Model3 can achieve the lowest verification score which means it has the strongest ability to preserve the identity during the image-to-video and video-to-image face translation task.

Table with 4 columns: Final states, Benchmark, Backgrounds, and Z². Rows for various k and m values.

Table 4: The number of events for various final states at 1000 fb⁻¹ of luminosity at the LHC for center of mass energy (COM) of 14 TeV.

negligible, due to the mostly-singlet nature of the η₁. We have also checked for other triplet gauge boson contributions to this final state, but they are all either zero or negligible. To reduce further the SM backgrounds we apply a veto on the mass peak of the 2 boson, by requiring that |m₄₃ - m₂₂| ≥ 5 GeV and |m₃₃ - m₂₂| > 10 GeV respectively. As one may deduce from Table [11] and Table [12] the application of these two cuts, though reduces the SM background quite drastically, does not affect the signal, which remains unchanged. Finally, we apply the constraint |m₃₃ - m₄₃| < 125 GeV to ensure the search for hidden scalars, i.e., m₄₃ < 125 GeV, which causes an even larger suppression of the background. At this level the signal significances are still below 3σ at 13 TeV and reach 3.20σ only in the case of the benchmark point BP2, at 14 TeV.

Next we apply the constraint |m₃₃ - m₄₃| ≤ 10 GeV to favour the search for a possible mass peak of the pseudoscalar and this enhances the signal significance to 3.40σ, 1.70σ and 2.64σ respectively for BP1, BP2 and BP3 at 13 TeV. At 14 TeV these numbers are 2.47σ, 2.51σ and 3.27σ respectively. Similar peaks are noted in p, parton-mass distribution, i.e. with p₃ |m₃₃ - m₄₃| ≤ 5 GeV, give signal significances of 2.63σ and 2.65σ for BP1 and BP2, at a center of mass energy of 13 TeV. BP3 in this case runs out of statistics. At 14 TeV the signal significances are 2.05σ, 2.82σ and 2.04σ respectively. The leptonic modes thus need higher luminosities ∼ 2000 fb⁻¹ in order to reach the discover limit for a light pseudoscalar.



MNRAS 000, 1–7 (2000) Preprint 23 November 2019 Compiled using MNRAS L<sup>A</sup>T<sub>E</sub>X style file v3.0

### Cosmic String Detection with Tree-Based Machine Learning

A. Vafaei Sadr<sup>1,2,3</sup>, M. Farhang<sup>1</sup>, S. M. S. Movahed<sup>1,4</sup>\*, B. Bassett<sup>3,5,6,7</sup>, M. Kunz<sup>2</sup>

<sup>1</sup> Department of Physics, Shahid Beheshti University, Velenjani, Tehran 19378, Iran  
<sup>2</sup> Department of Physics, Theoretical and Cosmic for Astrophysics, University of Geneva, 12 Quai Ernest Ansermet, 1211 Geneva 4, Switzerland  
<sup>3</sup> Abruzzo Institute for Mathematical Sciences, 6 Molise Road, Montebelluna 36101, South Africa  
<sup>4</sup> School of Physics, Institute for Research in Fundamental Sciences (IPM), P. O. Box 19395-5511, Tehran, Iran  
<sup>5</sup> South African Astronomical Observatory, Observatory, Cape Town, 7925, South Africa  
<sup>6</sup> SCA, South Africa, The Park, Park Road, Pietermaritzburg, Cape Town 6001, South Africa  
<sup>7</sup> Department of Maths and Applied Maths, University of Cape Town, Cape Town, South Africa

23 November 2019

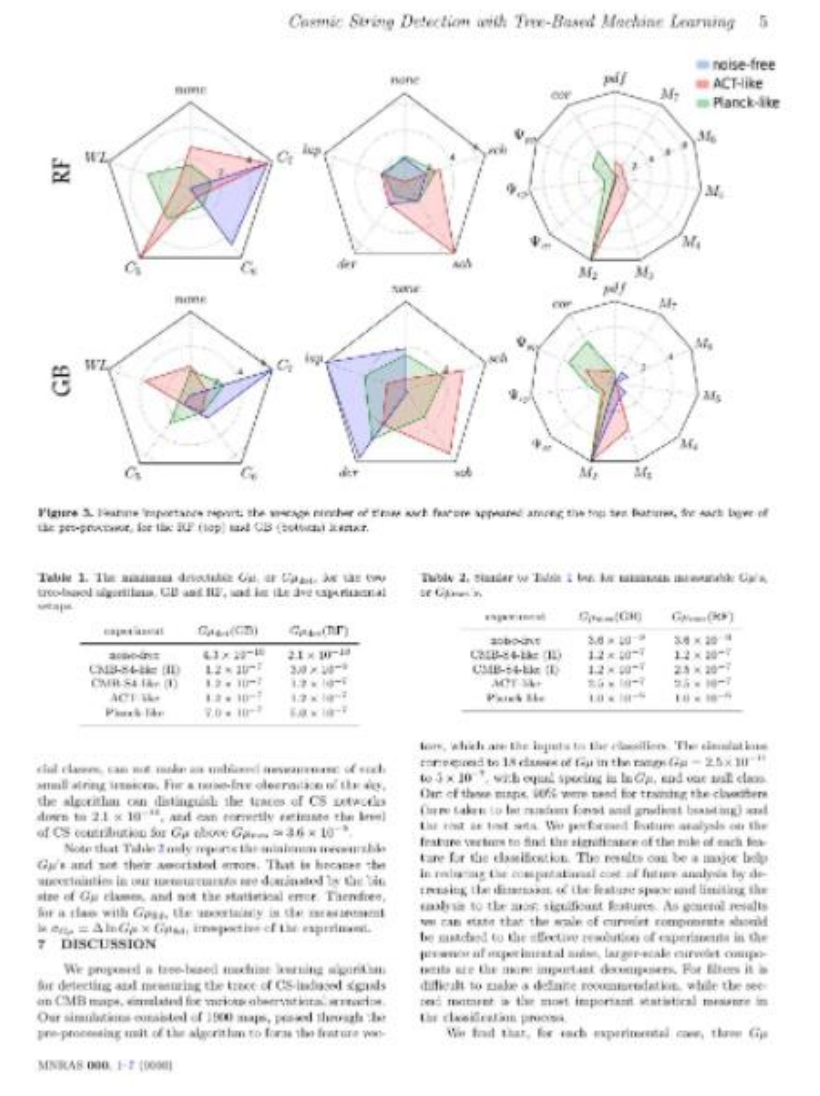
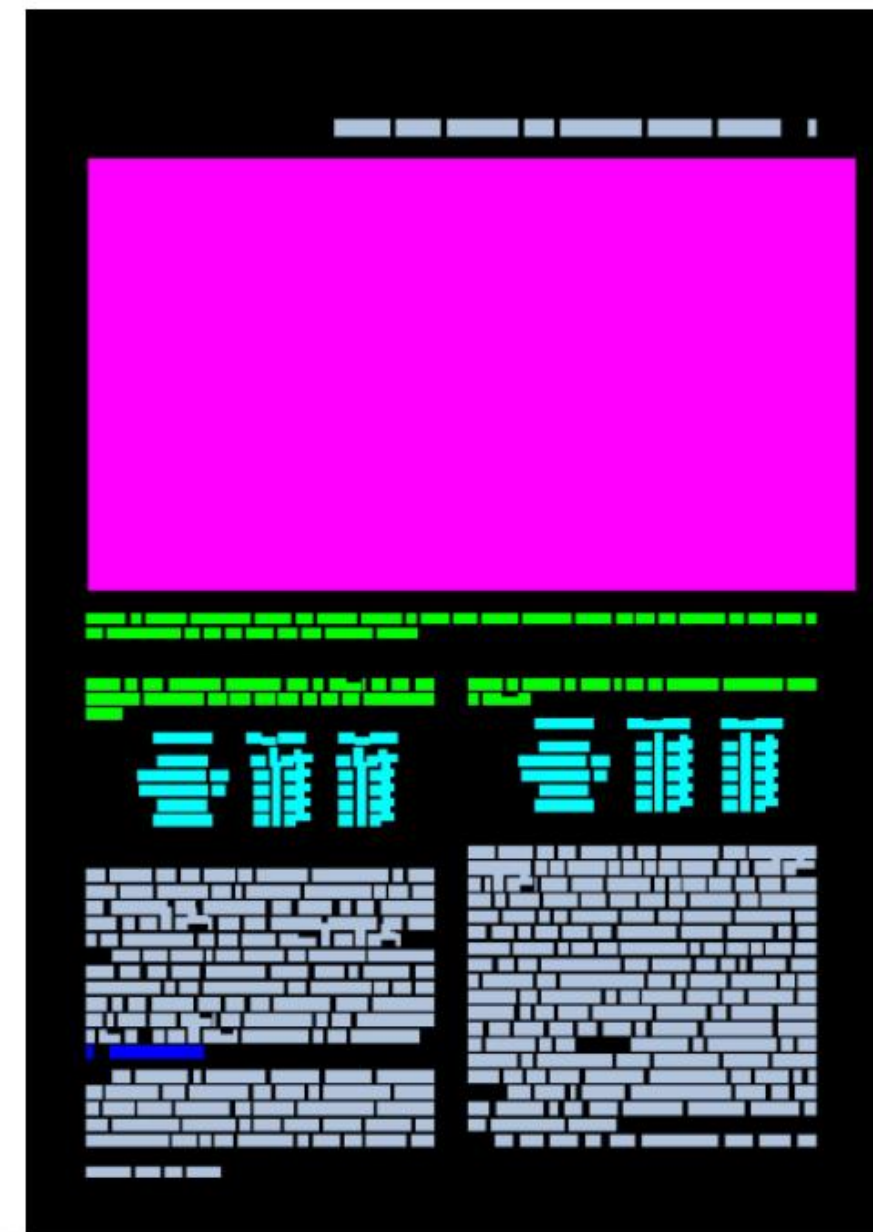
**ABSTRACT**  
 We explore the use of random forest and gradient boosting, two powerful tree-based machine learning algorithms, for the detection of cosmic strings in maps of the cosmic microwave background (CMB), through their unique Gini-Kaiser-Selwyn effect on the temperature anisotropies. The information 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 3 $\sigma$  detection of strings with  $G\mu \geq 2.1 \times 10^{-7}$  for noise-free,  $\mathcal{N}(\mathcal{P})$  modulation CMB observations. The minimum detectable tension increases to  $G\mu \geq 3.0 \times 10^{-8}$  for a more realistic, CMB S4-like (H) strategy, still a significant improvement over the previous results.

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

**1 INTRODUCTION**  
 The inflationary paradigm is the most widely accepted scenario for seeding the structure in the Universe, as far as our observational tests with flying colors. There is, however, both theoretical and observational reasons for contributions from alternative, well-motivated scenarios. Among these are perturbations seeded by cosmic topological defects formed at cosmological phase transitions. In particular, cosmic strings (CS) are theoretically expected to be produced in the early Universe (Kibble 1976; Zeldovich 1968; Vilenkin 1981; Vachaspati & Vilenkin 1984; Vilenkin 1985; Shellard 1987; Hochstadt & Kibble 1995; Vilenkin & Shellard 2000; Subramanian 2007; Davis et al. 2008; Degen 2009; Davis et al. 2010; Copeland et al. 1984; Subramanian 1997; Sarangi & Tye 2002; Copeland et al. 2004; Pogosian et al. 2004; Majumdar & Christian-Davies 2002; Dvali & Vilenkin 2004; Kibble 2004; Hory Tye 2008). The detection of CS would open a unique window to the physics of the early Universe (Kibble 1976; Zeldovich 1968; Vilenkin 1981; Vilenkin & Shellard 2000; Finocchia & Tye 2005). Therefore a lot of effort has been put into developing powerful statistical tools for cosmic string network detection and putting tight upper bounds on the CS tension, parametrized by  $G\mu$ , where  $G$  and  $\mu$  represent Newton's constant and the string's tension, respectively. The string tension is intimately related to the energy of the phase transition epoch,

$$\frac{G\mu}{c^2} = \nu \left( \frac{m^2}{M_{\text{Planck}}^2} \right) \quad (1)$$

where  $\nu$  is the symmetry breaking energy scale,  $c$  is the speed of light and  $M_{\text{Planck}} = \sqrt{\hbar c/8\pi G}$  represents the Planck mass. In this paper we work in natural units with  $\hbar = c = 1$ . A CS network would leave unique imprints on cosmic microwave background (CMB) anisotropies. The Gini-Kaiser-Selwyn (KS) effect (Kaiser & Selwyn 1984; Gini 1931 1985; Selwyn 1988; Baecher et al. 1988; Allen et al. 1992; Fox et al. 1997; Singard & Baecher 2012) corresponds to the integrated Sachs-Wolfe effect caused by moving strings. It produces line-like discontinuities on the



The colors of semantic structure labels are:

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

# ReadingBank

**Children and young people with a clear mental health diagnosis**  
OSCA will offer care to children and young people who have already been specialist CAMHS criteria, i.e. have been diagnosed with a serious mental health disorder where:

- The child/young person is at risk of placement breakdown and failed to engage with or disengaged from specialist CAMHS services
- Where the intensity of an intervention required to support a child in placement is greater than the resources available within specialist CAMHS (more than one visit per week required), and there is a history of the child and young person failing to engage with these services on a regular basis

**Where a mental health diagnosis is less clear**  
OSCA will care coordinate complex cases that meets at least one of the following criteria:

- The child/young person is looked after, adopted or under a child protection plan
- The young person is significantly involved in the criminal justice system / or has major substance misuse issues
- The young person has an Education, Health and Care Plan (EHCP), and is educated within specialist educational provision

**And where:**

- Significant emotional, behavioural, or mental health concerns that have been identified through the Youth Offending Service (YOS) team assessment, Education, Health and Care Plan (EHCP), Family Star assessment, or Department of Health Framework for Assessment

**And where a minimum of two of the following criteria apply:**

- The child or young person is at risk of placement breakdown (either home or a care placement)
- The child and young person's needs cannot be met by the range of professionals currently involved with the case
- A standard primary mental health intervention is CLEARLY not sufficient to meet the child's needs
- A range of other primary mental health interventions have already been tried and have proved unsuccessful or there is a history of failure to engage

**Consultation and support to frontline professionals**  
OSCA will provide named workers to support the following agencies:

- Looked after children's services
- Youth offending and substance misuse services
- Special schools for children with emotional difficulties

OSCA will provide support, advice and opportunities for frontline children services to ensure that children and young people are:

- Appropriately supported at the right level of care
- Ensure timely access into additional services where required

**Child Interventions**  
The OSCA team will provide Child support which involves intensive community-oriented treatment to children and young people in the acute / crisis phase of mental illness, which, in the absence of the Child Intervention, would be at risk of repeat admission.

OSCA team will

Page 3 of 8

(a)

**Children and young people with a clear mental health diagnosis**  
OSCA will offer care to children and young people who have already been specialist CAMHS criteria, i.e. have been diagnosed with a serious mental health disorder where:

- The child/young person is at risk of placement breakdown and failed to engage with or disengaged from specialist CAMHS services
- Where the intensity of an intervention required to support a child in placement is greater than the resources available within specialist CAMHS (more than one visit per week required), and there is a history of the child and young person failing to engage with these services on a regular basis

**Where a mental health diagnosis is less clear**  
OSCA will care coordinate complex cases that meets at least one of the following criteria:

- The child/young person is looked after, adopted or under a child protection plan
- The young person is significantly involved in the criminal justice system / or has major substance misuse issues
- The young person has an Education, Health and Care Plan (EHCP), and is educated within specialist educational provision

**And where:**

- Significant emotional, behavioural, or mental health concerns that have been identified through the Youth Offending Service (YOS) team assessment, Education, Health and Care Plan (EHCP), Family Star assessment, or Department of Health Framework for Assessment

**And where a minimum of two of the following criteria apply:**

- The child or young person is at risk of placement breakdown (either home or a care placement)
- The child and young person's needs cannot be met by the range of professionals currently involved with the case
- A standard primary mental health intervention is CLEARLY not sufficient to meet the child's needs
- A range of other primary mental health interventions have already been tried and have proved unsuccessful or there is a history of failure to engage

**Consultation and support to frontline professionals**  
OSCA will provide named workers to support the following agencies:

- Looked after children's services
- Youth offending and substance misuse services
- Special schools for children with emotional difficulties

OSCA will provide support, advice and opportunities for frontline children services to ensure that children and young people are:

- Appropriately supported at the right level of care
- Ensure timely access into additional services where required

**Child Interventions**  
The OSCA team will provide Child support which involves intensive community-oriented treatment to children and young people in the acute / crisis phase of mental illness, which, in the absence of the Child Intervention, would be at risk of repeat admission.

OSCA team will

Page 3 of 8

(b)

**Table 5.8.32.A – Glasgow Point south neighbourhood plan: material change of use**

Use	Category of development and assessment	Assessment benchmarks
<b>If in the neighbourhood plan area</b>		
SDCA, if suitable development where not listed in this table	No change	Glasgow Point south neighbourhood plan code
<b>If in the Alford use zone</b>		
Centre activities (activity group)	Accepted development, subject to compliance with identified requirements	Not applicable
	If existing on existing premises, where:	
	(a) gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre;	
	(b) complying with all applicable outcomes in section 8 of the Centre or mixed use code	
	<b>Accessible development – Code assessment:</b>	
	If existing on existing premises, where:	Centre or mixed use code – purpose, overall outcomes and section 8 outcomes only
	(a) gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre;	
	(b) not complying with all applicable outcomes in section 8 of the Centre or mixed use code	
	If existing in new premises or an existing premises with an increase in gross floor area, where gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre	Glasgow Point south neighbourhood plan code Centre or mixed use code Prescribed secondary code

Page 3 of 8

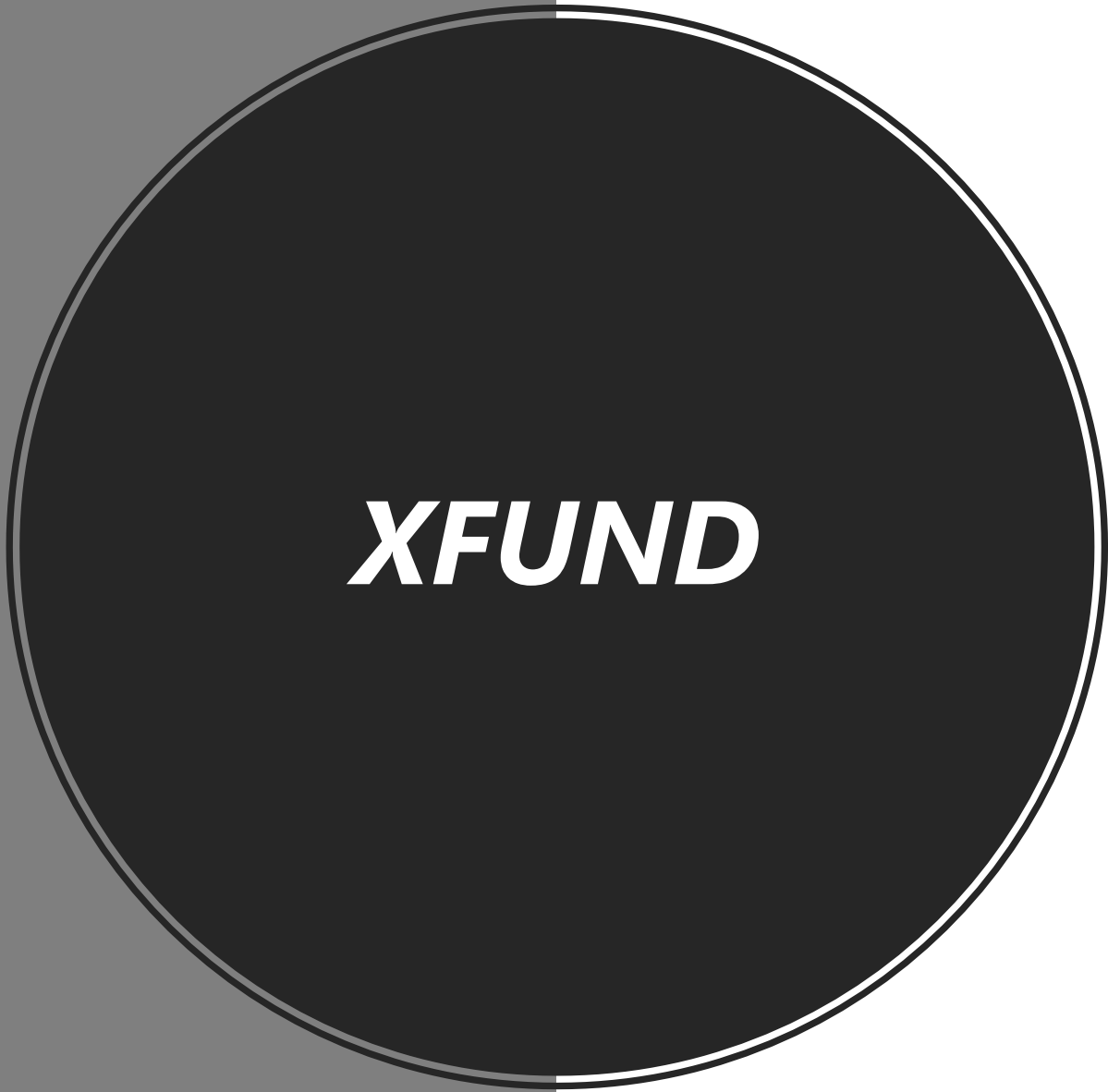
(c)

**Table 5.8.32.A – Glasgow Point south neighbourhood plan: material change of use**

Use	Category of development and assessment	Assessment benchmarks
<b>If in the neighbourhood plan area</b>		
SDCA, if suitable development where not listed in this table	No change	Glasgow Point south neighbourhood plan code
<b>If in the Alford use zone</b>		
Centre activities (activity group)	Accepted development, subject to compliance with identified requirements	Not applicable
	If existing on existing premises, where:	
	(a) gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre;	
	(b) complying with all applicable outcomes in section 8 of the Centre or mixed use code	
	<b>Accessible development – Code assessment:</b>	
	If existing on existing premises, where:	Centre or mixed use code – purpose, overall outcomes and section 8 outcomes only
	(a) gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre;	
	(b) not complying with all applicable outcomes in section 8 of the Centre or mixed use code	
	If existing in new premises or an existing premises with an increase in gross floor area, where gross floor area is no greater than 1,500sqm for any individual tenancy where shop or shop component of a shopping centre	Glasgow Point south neighbourhood plan code Centre or mixed use code Prescribed secondary code

Page 3 of 8

(d)



Form in Chinese: 开放式基金财产业务申请表. Includes fields for applicant information, company details, and a barcode at the bottom.

Chinese

Form in English: RESPECTED BROWN & WILKINSON INTERNATIONAL TOBACCO. Includes technical specifications for a cigarette and a barcode at the bottom.

English

Form in Italian: MODULO ADESIONE A POLIZZA RESPONSABILITÀ CIVILE. Includes personal and insurance details for ANNI LUCA.

Italian

Form in Portuguese: IDENTIFICAÇÃO DO COORDENADOR DO PROJETO. Includes details for Maria Jesus Pinto and the project location.

Portuguese

Form in German: Bewerbungsformular - Lehrstelle 2017. Includes applicant details for Hans Schmit and contact information for ANDRITZ HYDRO GmbH.

German

Form in Spanish: FORMULARIO DE SOLICITUD DE BECA. Includes personal information for Carmen Garcia and contact details for CEPB.

Spanish

Form in French: FORMULAIRE D'INSCRIPTION Programme Vendanges en France 2020. Includes registration details for a wine harvest program.

French

Form in Japanese: 登録用紙. Includes registration details for a company and contact information for home labo.

Japanese



# 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
BERT <sub>BASE</sub>	0.6026	0.8968	0.9099	0.7790
UniLMv2 <sub>BASE</sub>	0.6648	0.9092	0.9459	0.7950
BERT <sub>LARGE</sub>	0.6563	0.9025	0.9200	0.7910
UniLMv2 <sub>LARGE</sub>	0.7072	0.9205	0.9488	0.8180
LayoutLM <sub>BASE</sub>	0.7866	0.9472	0.9438	0.8270
LayoutLM <sub>LARGE</sub>	0.7895	0.9493	0.9524	0.8340
LayoutLMv2 <sub>BASE</sub>	0.8276	0.9495	0.9625	0.8330
LayoutLMv2 <sub>LARGE</sub>	<b>0.8420</b>	<b>0.9601</b>	<b>0.9781</b>	<b>0.8520</b>
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	–
RoBERTa <sub>BASE</sub> 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

<b>Model</b>	<b>Accuracy</b>	<b>#Parameters</b>
BERT <sub>BASE</sub>	89.81%	110M
UniLMv2 <sub>BASE</sub>	90.06%	125M
BERT <sub>LARGE</sub>	89.92%	340M
UniLMv2 <sub>LARGE</sub>	90.20%	355M
LayoutLM <sub>BASE</sub> (w/ image)	94.42%	160M
LayoutLM <sub>LARGE</sub> (w/ image)	94.43%	390M
LayoutLMv2 <sub>BASE</sub>	95.25%	200M
LayoutLMv2 <sub>LARGE</sub>	<b>95.64%</b>	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 <sup>6</sup> (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
BERT <sub>BASE</sub>	train	0.6354	110M
UniLMv2 <sub>BASE</sub>	train	0.7134	125M
BERT <sub>LARGE</sub>	train	0.6768	340M
UniLMv2 <sub>LARGE</sub>	train	0.7709	355M
LayoutLM <sub>BASE</sub>	train	0.6979	113M
LayoutLM <sub>LARGE</sub>	train	0.7259	343M
LayoutLMv2 <sub>BASE</sub>	train	0.7808	200M
LayoutLMv2 <sub>LARGE</sub>	train	0.8348	426M
LayoutLMv2 <sub>LARGE</sub>	train + dev	0.8529	426M
LayoutLMv2 <sub>LARGE</sub> + QG	train + dev	<b>0.8672</b>	426M
Top-1 on DocVQA Leaderboard (30 models ensemble) <sup>7</sup>	-	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	<b>0.9166</b>	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	<b>0.9810</b>	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	<b>0.9311</b>

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)	<b>0.9658</b>	<b>0.9594</b>	<b>0.9625</b>
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 <sub>BASE</sub>	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 <sub>BASE</sub>	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 <sub>BASE</sub>	<b>0.9816</b>	0.8595	0.9597	0.8947	1.0000	0.8957	0.8948	0.9788	0.9338	0.9598	0.8633	<b>0.9579</b>	0.9316
BERT <sub>LARGE</sub>	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 <sub>LARGE</sub>	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 <sub>LARGE</sub>	0.9784	0.8783	0.9556	0.8974	<b>1.0000</b>	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 <sub>BASE</sub>	0.9815	0.8907	<b>0.9669</b>	0.9430	0.9990	0.9292	<b>0.9300</b>	0.9843	<b>0.9437</b>	0.9664	0.8818	0.9575	0.9478
X101+LayoutLM <sub>LARGE</sub>	0.9802	<b>0.8964</b>	0.9666	<b>0.9440</b>	0.9994	<b>0.9352</b>	0.9293	<b>0.9844</b>	<b>0.9430</b>	<b>0.9670</b>	<b>0.8875</b>	0.9531	<b>0.9488</b>

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 $\uparrow$	ARD $\downarrow$
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	<b>0.9819</b>	<b>1.75</b>

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 $\uparrow$	ARD $\downarrow$
Heuristic Method	0.3391	13.61
Tesseract OCR	0.7532	1.42
LayoutReader	<b>0.9360</b>	<b>0.27</b>

Table 5: Adaption to text lines of Tesseract OCR

Method	Avg. Page-level BLEU $\uparrow$	ARD $\downarrow$
Heuristic Method	0.3752	10.17
The commercial OCR	0.8530	2.40
LayoutReader	<b>0.9430</b>	<b>0.59</b>

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

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

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

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



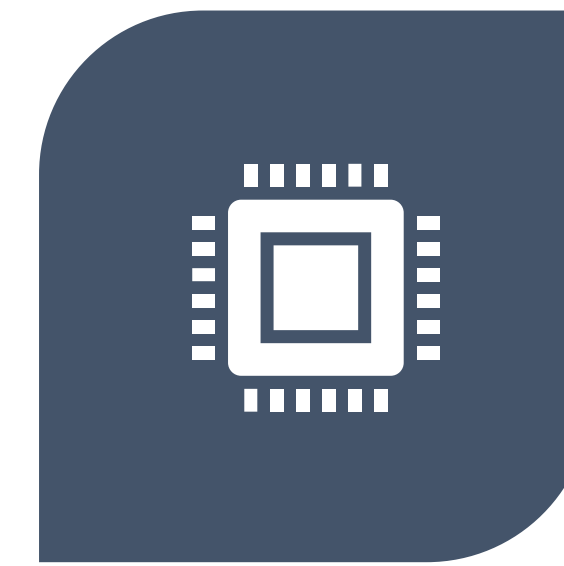
MODEL LIMITATIONS



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  
<https://github.com/microsoft/unilm>

**XLM-T** ( NEW ): Multilingual NMT w/ pretrained cross-lingual encoders