CS 772: Deep learning for natural language processing

Information Extraction

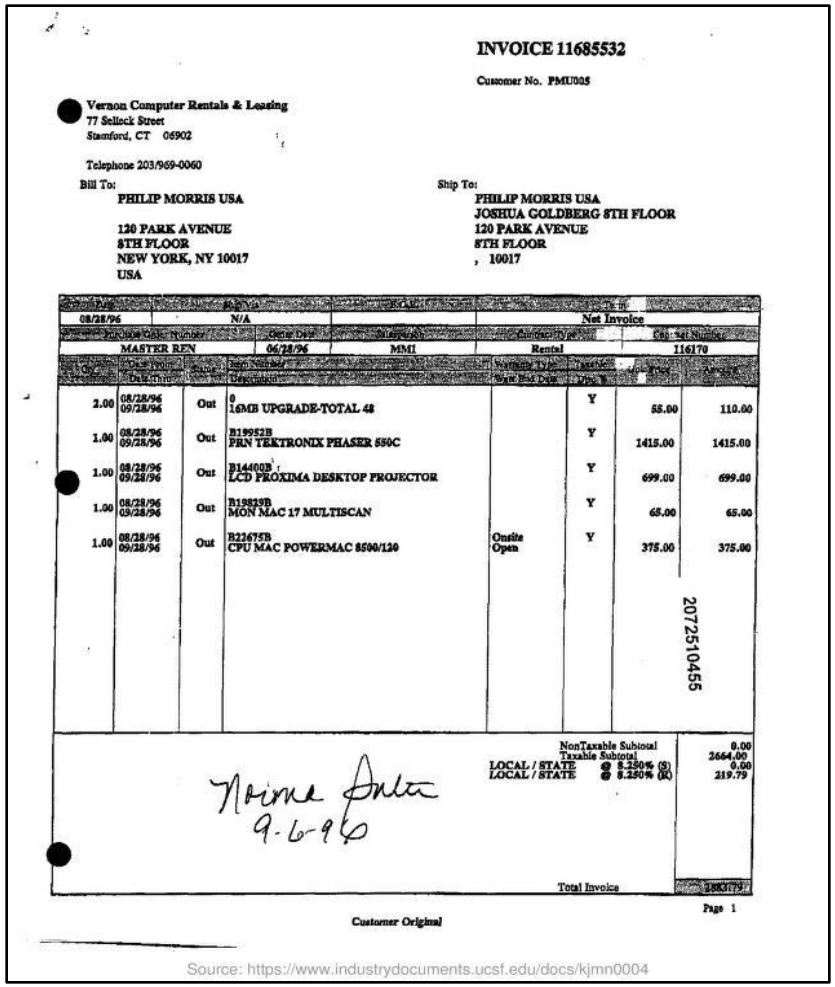
Saiful Haq
4th Year PhD Candidate at IIT Bombay
Director of AI and Staff Research Engineer, Hyperbots Inc

What is the plan for today

- Section 1: Motivation and Overview
- Section 2: Core Concepts
- Section 3: Modern Approaches to Information Extraction
- Section 4: QA

Motivation and Overview

Introduction



INVOICE 11685532

Customer No. PMU005

Vernon Computer Rentals & Leasing
77 Selleck Street
Stamford, CT 06902

Telephone 203/969-0060

Bill To:
PHILIP MORRIS USA

120 PARK AVENUE
8TH FLOOR
NEW YORK, NY 10017
USA

Ship To:
PHILIP MORRIS USA
JOSHUA GOLDBERG 8TH FLOOR
120 PARK AVENUE
8TH FLOOR
, 10017

Date	P.O. #			Net Invoice
08/28/96	N/A			
Invoice Date / Number	Order Date	Salesman	Currency Type	Cust. Id Number
MASTER REN 06/28/96	MMI	Rental		116170

Qty	Date Out / Date In	Status	Item Number / Description	Warranty Type / Warranty End Date	Taxable	Unit Price	Amount
2.00	08/28/96 09/28/96	Out	0 16MB UPGRADE-TOTAL 48		Y	55.00	110.00
1.00	08/28/96 09/28/96	Out	B19952B PRN TEKTRONIX PHASER 550C		Y	1415.00	1415.00
1.00	08/28/96 09/28/96	Out	B14400B LCD PROXIMA DESKTOP PROJECTOR		Y	699.00	699.00
1.00	08/28/96 09/28/96	Out	B19819B MON MAC 17 MULTISCAN		Y	65.00	65.00
1.00	08/28/96 09/28/96	Out	B22675B CPU MAC POWERMAC 8500/120	Onsite Open	Y	375.00	375.00

2072510455

Norma Dalton 9-6-96

NonTaxable Subtotal	0.00
Taxable Subtotal	2664.00
LOCAL / STATE @ 8.250% (S)	0.00
LOCAL / STATE @ 8.250% (R)	219.79
Total Invoice	2883.79

Page 1

Customer Original

Source: https://www.industrydocuments.ucsf.edu/docs/kjmn0004

Scanned Invoice Document

- Majority of business-to-business communications involve interchange of financial documents (e.g. *invoices, receipts, purchase orders*, etc.) through email, postal mail, or fax.
  - Majority of enterprise data exists digitally as scanned PDFs or Images.
- **Where do we use this data?**
  - Multiple applications but today we will focus on procurement and accounting.

Procurement lifecycle in enterprises

Task: Procure 100 Chairs for the CSE Department, IIT Bombay

Stage	Document	Purpose
Send Purchase Order	Purchase Order (PO)	It's created by the **procurement team of IIT Bombay** and sent to the **Vendor**. "We want 100 chairs at ₹500 each"
Receive Orders	Goods Receipt Note (GRN)	Issued by the Procurement Team; confirms what was actually received. "We received 100 chairs"
Receive Invoice	Invoice	Sent by the Vendor; requests payment. "Please pay ₹50,000 for 100 chairs at ₹500 each"
3-way matching		PO, GRN and Invoice are ***compared*** by a human (Analyst) or a software and payment is made.

Procurement lifecycle in enterprises

We want 100 chairs at ₹500 each

Purchase Order

We received 100 chairs

Good Receipt Note

Please pay ₹50,000 for 100 chairs at ₹500 each

Invoice

Do the documents match?

Note: Fields which we need to compare include quantity, unit price and total amount.

Procurement lifecycle in enterprises

Do the documents match?

Note: Fields which we need to compare include quantity, unit price and total amount.

Procurement lifecycle in enterprises

Do the documents match?

Note: Fields which we need to compare include quantity, unit price and total amount.

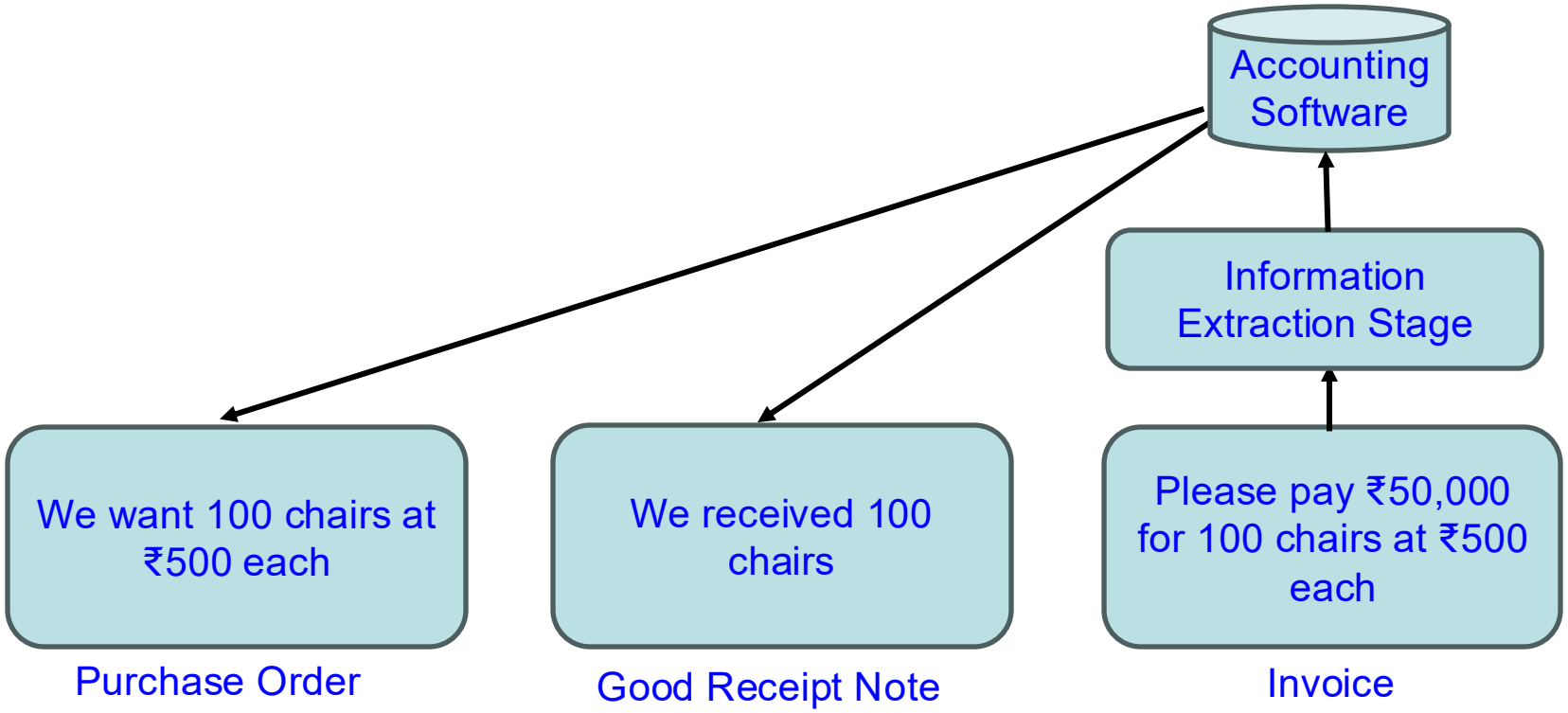
Challenges in 3-way matching

- The **process of 3-way matching**, which involves verifying consistency between a *purchase order (PO)*, a *goods receipt note (GRN)*, and the *invoice*, is traditionally performed by human analysts. This manual verification process is highly **error-prone and time-consuming** due to the following reasons:
  - Large enterprises receive **1,000–10,000 invoices per day**, making manual review infeasible at scale.
  - Invoices can be **long and complex**, sometimes spanning **hundreds of pages**, especially in sectors such as manufacturing or logistics.
  - Invoices often arrive in **varied formats (**PDFs, scans, images, or even handwritten notes), making consistency checks difficult.

Challenges in 3-way matching

- Errors or delays in 3-way matching can lead to significant operational and financial consequences:
  - **Overpayments:** Human oversight can result in paying more than the actual amount due.
  - **Late payment penalties:** Delays in validation and approval may lead to fines or interest charges.
  - **Missed early payment discounts:** Many invoices offer discounts for early settlement, which are lost if processing is delayed.
  - **Vendor relationship damage:** Persistent delays or payment disputes can harm the company's reputation and vendor trust, potentially disrupting future supply chains.

How do modern enterprises mitigate this challenge?

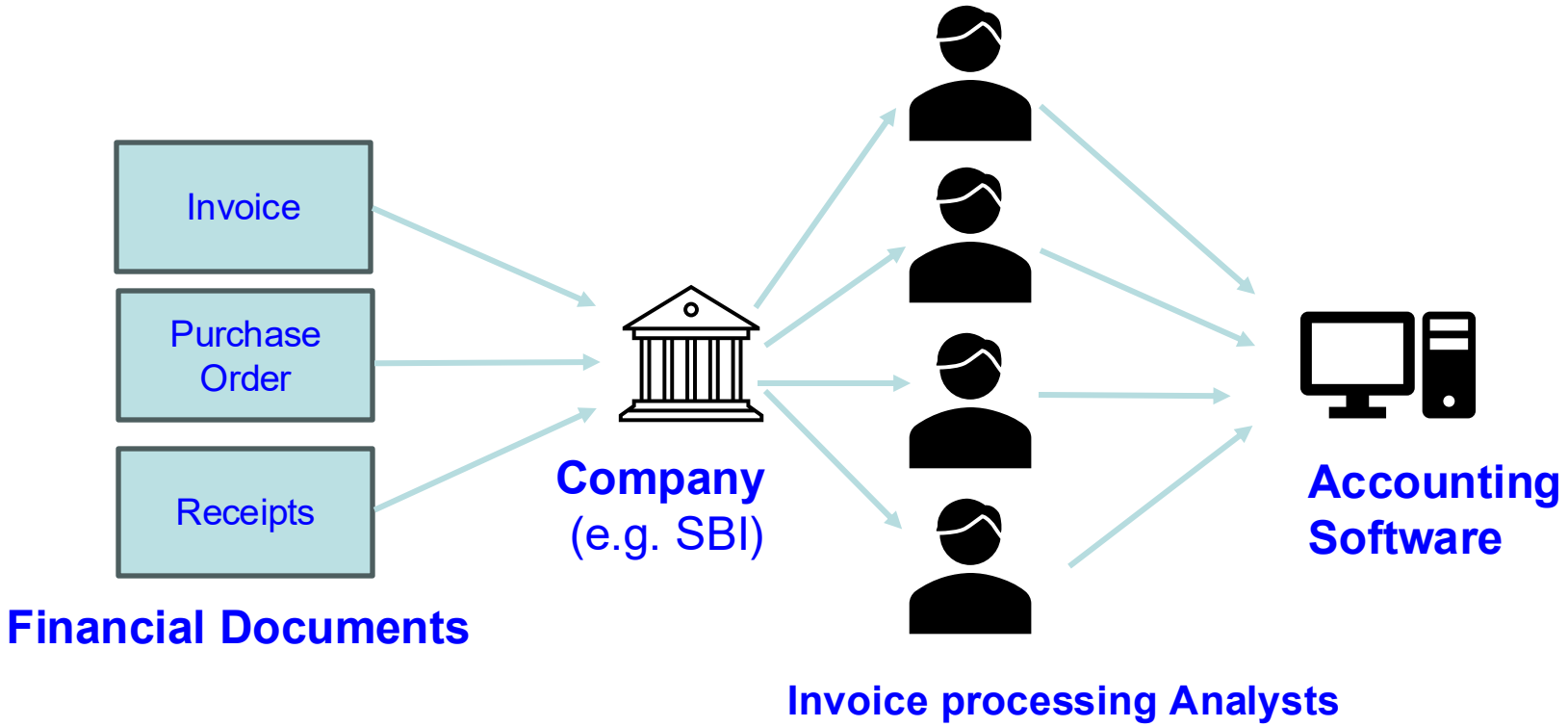


Documents like Purchase Orders (POs) and Goods Receipt Notes (GRNs) are generated *within* the company's accounting software, so the software already knows their structured fields. But the Invoice comes from an external vendor, often as a PDF or scanned copy. To perform the '3-way match' i.e comparing PO, GRN, and Invoice, we need to **extract the invoice fields** from that semi-structured document. That extraction process, turning document image into structured fields like 'vendor_name,' 'invoice_number,' 'total_amount,' etc. is the core job of an **Information Extraction system.**"

Challenges in 3-way matching

- Historically, enterprises have relied on ***manual data entry*** for invoice documents, which is both ***time-consuming*** and ***costly.*** Why?
  - While 3-way matching can be automated using downstream enterprise software, these systems depend on the availability of structured, machine-readable data (e.g., JSON, CSV, or XML). However, Invoice documents typically exist in semi-structured formats **(**PDFs, scans, or images) requiring **Information extraction** to convert them into structured representations.
  - This **information extraction stage** goes far beyond conventional Optical Character Recognition (OCR). It requires an understanding of **semantics, layout, and contextual relationships** within the document, as well as **domain-specific reasoning**. Consequently, automating this step is a major bottleneck.

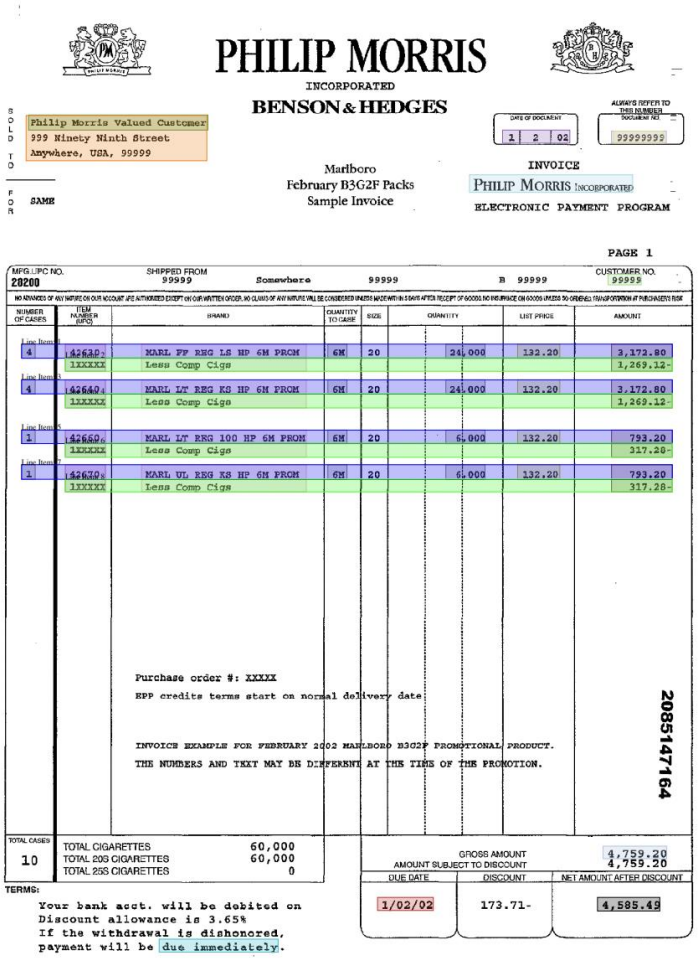
Impact of automating the information extraction stage



- Average **salary** of an Invoice Processing Analyst in India is **3.4 Lakhs per annum**. Their task is to manually enter information from Invoices into the Accounting software.
- AI models can be used to **automate this task**. This **reduces** the employee **head count** which further **reduces** both ***costs*** and ***time*** associated with Information extraction.

Core Concepts

Problem Statement



PHILIP MORRIS
INCORPORATED
BENSON & HEDGES

Philip Morris Valued Customer
999 Ninety Ninth Street
Anywhere, USA, 99999

SAME

Marlboro
February B3G2F Packs
Sample Invoice

INVOICE
PHILIP MORRIS INCORPORATED
ELECTRONIC PAYMENT PROGRAM

PAGE 1

NUMBER OF CASES	ITEM NUMBER (UPC)	BRAND	QUANTITY TO CASE	SIZE	QUANTITY	LIST PRICE	AMOUNT
4	42630	MARL FF REG LS HP 6M PROM	6M	20	24,000	132.20	3,172.80
	XXXXX	Less Comp Cigs					1,269.12-
4	42640	MARL LT REG KS HP 6M PROM	6M	20	24,000	132.20	3,172.80
	XXXXX	Less Comp Cigs					1,269.12-
1	42650	MARL LT REG 100 HP 6M PROM	6M	20	6,000	132.20	793.20
	XXXXX	Less Comp Cigs					317.28-
1	42670	MARL UL REG KS HP 6M PROM	6M	20	6,000	132.20	793.20
	XXXXX	Less Comp Cigs					317.28-

Purchase order #: XXXXX
EPP credits terms start on normal delivery date

INVOICE EXAMPLE FOR FEBRUARY 2002 MARLBORO B3G2F PROMOTIONAL PRODUCT.
THE NUMBERS AND TEXT MAY BE DIFFERENT AT THIS TIME OF THE PROMOTION.

2085147164

TOTAL CASES		
10	TOTAL CIGARETTES	60,000
	TOTAL 20S CIGARETTES	60,000
	TOTAL 25S CIGARETTES	0

GROSS AMOUNT 4,759.20
AMOUNT SUBJECT TO DISCOUNT 4,759.20

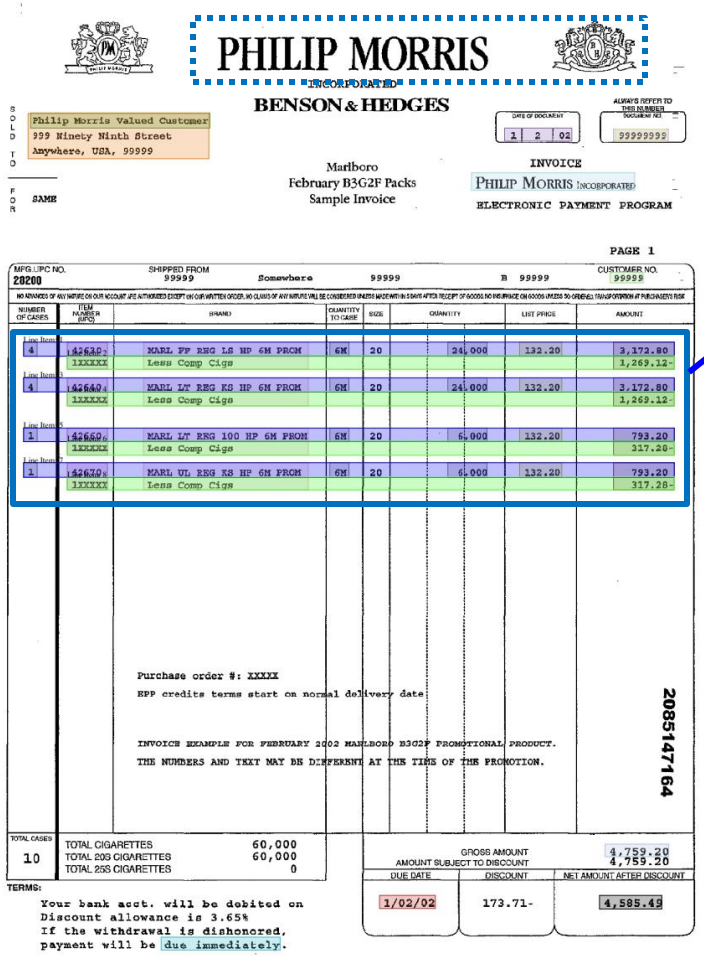
DUE DATE	DISCOUNT	NET AMOUNT AFTER DISCOUNT
1/02/02	173.71-	4,585.49

TERMS:
Your bank acct. will be debited on
Discount allowance is 3.65%
If the withdrawal is dishonored,
payment will be due immediately.

Problem statement: Develop a state-of-the art invoice information extraction model.

1. **Input:** Invoice in JPEG/PNG Format
2. **Output:** Key-value pairs, where key is a field in the invoice (Invoice ID, Invoice Date, Gross Amount, etc.) and the value is the text corresponding to that field. *e.g. {"invoice_id": None, "invoice_date": "01/02/2002", "gross_amount": "4,585.49"}*
3. Model should understand the ***semantics***, ***layout*** and ***context*** of the information within the document and should be also able to perform **reasoning**.

Problem Statement



PHILIP MORRIS INCORPORATED

BENSON & HEDGES

Philip Morris Valued Customer
999 Ninety Ninth Street
Anywhere, USA, 99999

SOLD TO

FOR SAME

Marlboro
February B3G2F Packs
Sample Invoice

INVOICE

PHILIP MORRIS INCORPORATED

ELECTRONIC PAYMENT PROGRAM

DATE OF DOCUMENT 1 2 02

ALWAYS REFER TO THIS NUMBER DOCUMENT NO. 99999999

PAGE 1

MFG. LIPC NO. 20200 | SHIPPED FROM 99999 Somewhere 99999 B 99999 | CUSTOMER NO. 99999

NUMBER OF CASES	ITEM NUMBER (UPC)	BRAND	QUANTITY TO CASE	SIZE	QUANTITY	LIST PRICE	AMOUNT
Line Item 1							
4	42630	MARL FF REG LS HP 6M PROM	6M	20	24.000	132.20	3,172.80
	1XXXXX	Less Comp Cigs					1,269.12-
Line Item 2							
4	42640	MARL LT REG KS HP 6M PROM	6M	20	24.000	132.20	3,172.80
	1XXXXX	Less Comp Cigs					1,269.12-
Line Item 3							
1	42660	MARL LT REG 100 HP 6M PROM	6M	20	6.000	132.20	793.20
	1XXXXX	Less Comp Cigs					317.28-
Line Item 4							
1	42670	MARL UL REG KS HP 6M PROM	6M	20	6.000	132.20	793.20
	1XXXXX	Less Comp Cigs					317.28-

Purchase order #: XXXXX
RPP credits terms start on normal delivery date

INVOICE EXAMPLE FOR FEBRUARY 2002 MARLBORO B3G2F PROMOTIONAL PRODUCT.
THE NUMBERS AND TEXT MAY BE DIFFERENT AT THE TIME OF THE PROMOTION.

2085147164

TOTAL CASES		
10	TOTAL CIGARETTES	60,000
	TOTAL 20S CIGARETTES	60,000
	TOTAL 25S CIGARETTES	0

GROSS AMOUNT	4,759.20
AMOUNT SUBJECT TO DISCOUNT	4,759.20

DUE DATE	DISCOUNT	NET AMOUNT AFTER DISCOUNT
1/02/02	173.71-	4,585.49

TERMS:
Your bank acct. will be debited on
Discount allowance is 3.65%
If the withdrawal is dishonored,
payment will be due immediately.

Types of fields in an Invoice:

- **Line-items:** List of product and services. Mostly exists as table but can also occur as free text.
- **Entities:** All non-line-item fields such as invoice id, receiver address, total amount, etc.

Evaluation Metric – Micro Averaged F1 Score

Document	Field	GT (ground truth)	Model Extracted	Correct? (text match?)
Document 1	Invoice ID	“1234”	“1234”	Yes
Document 1	Gross Amount	“100”	“1000”	No
Document 2	Invoice ID	“456”	“456”	Yes
Document 2	Gross Amount	“200”	None	No

We emphasize text matching meaning the extracted text must *exactly* match the ground truth. If not, it’s treated as incorrect (false positive or false negative as appropriate)

From this table we can compute for **all field instances** across both documents:

- True Positives (TP): number of fields where model extracted exactly correct text = 2
- False Positives (FP): number of fields predicted by model but incorrect text match = 1
- False Negatives (FN): number of ground-truth fields not correctly extracted = 2
- Micro Averaged Precision AP = TP / (TP + FP) = 2 / (2 + 1) = 2/3 = 0.66
- Micro Averaged Recall AR = TP / (TP + FN) = 2 / (2 + 2) = 2/4 = 0.50
- Micro Averaged F1 = 2 × (AP× AR) / (AP+ AR) = 2 ×(0.66×0.50)/(0.66+0.50) = 0.57

Evaluation Metric – Micro Averaged F1 Score

Document	Field	GT (ground truth)	Model Extracted	Correct? (text match?)
Document 1	Line item 1:Description	“Microsoft”	“Microsoft”	Yes
Document 1	Line item 1:Amount	“100”	“1000”	No
Document 2	Line item 1:Description	“Macbook”	“Macbook”	Yes
Document 2	Line item 1:Amount	“200”	None	No

We emphasise text matching meaning the extracted text must *exactly* match the ground truth. If not, it’s treated as incorrect (false positive or false negative as appropriate)

From this table we can compute for **all field instances** across both documents:

- True Positives (TP): number of fields where model extracted exactly correct text = 2
- False Positives (FP): number of fields predicted by model but incorrect text match = 1
- False Negatives (FN): number of ground-truth fields not correctly extracted = 2
- Micro Averaged Precision AP = TP / (TP + FP) = 2 / (2 + 1) = 2/3 = 0.66
- Micro Averaged Recall AR = TP / (TP + FN) = 2 / (2 + 2) = 2/4 = 0.50
- Micro Averaged F1 = 2 × (AP× AR) / (AP+ AR) = 2 ×(0.66×0.50)/(0.66+0.50) = 0.57

Evaluation Metric – Macro averaged F1 score

- **F1 Score** = 2* Precision * Recall/ (Precision + Recall)
- **EM Score**
  - Example:
    - Reference tokens = DL NLP
    - Generated tokens = DL
    - Precision = 1, Recall = 0.5, F1 Score = 0.66, EM Score = 0
- **Field Level:**
  - $Precision_F$ = (RG_F / TG_F)
  - $Recall_F$ = (RG_F / TR_F)
  - **$F1_F$ = 2*($Precision_F$* $Recall_F$) / ($Precision_F$ + $Recall_F$)**
  - **Exact Match (EM_F): 1 if RG_F = TR_F else 0**
- **Overall:**
  - $F1_{overall}$ = A / B = ($\sum F1\_x\__F$) / (Total Number of Fields in x)
  - $EM_{overall}$: replace, F1 by EM in the above formula.
- **Nomenclature:**
  - RG_F = No. of Relevant Words in a Field “F” of Generated
  - TG_F = Total No. of Words in a Field “F” of Generated
  - TR_F = Total No. of Words in a Field “F” of Human Annotated

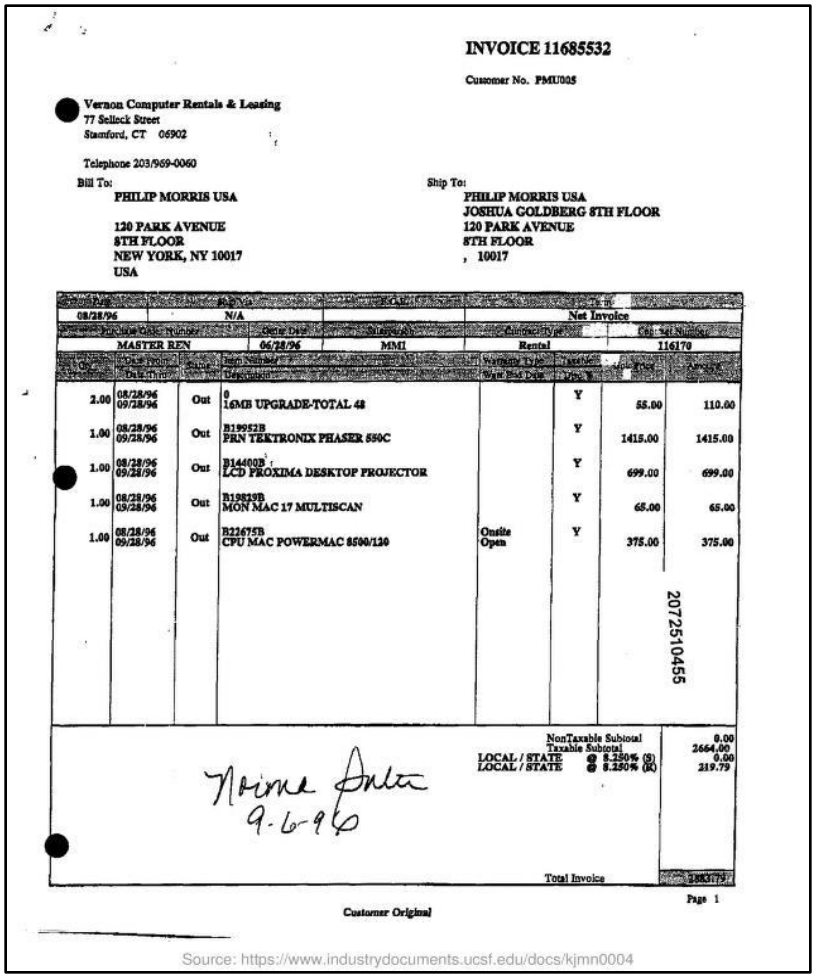
Evaluation Metric – Average Normalised Levenshtein Distance

Given two strings A and B, the Levenshtein distance d(A,B) is the *minimum number of single-character edits* (insertions, deletions, or substitutions) required to change A into B.

Ground Truth	**Output**	**Levenshtein Distance**	**Normalized Distance**
"Invoice"	"Invioce"	2	2 / 7 = 0.286
"Amount"	"Amunt"	1	1 / 6 = 0.167
"Total"	"Totla"	2	2 / 5 = 0.4

ANLD = 0.284

Why are financial documents semi-structured in nature?



INVOICE 11685532

Customer No. PMU005

Vernon Computer Rentals & Leasing
77 Selleck Street
Stamford, CT 06902

Telephone 203/969-0060

Bill To:
PHILIP MORRIS USA

120 PARK AVENUE
8TH FLOOR
NEW YORK, NY 10017
USA

Ship To:
PHILIP MORRIS USA
JOSHUA GOLDBERG 8TH FLOOR
120 PARK AVENUE
8TH FLOOR
, 10017

08/28/96		N/A			Net Invoice			
MASTER REN		06/28/96	MMI		Rental		116170	
2.00	08/28/96 09/28/96	Out	0 16MB UPGRADE-TOTAL 4S			Y	55.00	110.00
1.00	08/28/96 09/28/96	Out	B19952B PRN TEKTRONIX PHASER 550C			Y	1415.00	1415.00
1.00	08/28/96 09/28/96	Out	B14400B LCD PROXIMA DESKTOP PROJECTOR			Y	699.00	699.00
1.00	08/28/96 09/28/96	Out	B19819B MON MAC 17 MULTISCAN			Y	65.00	65.00
1.00	08/28/96 09/28/96	Out	B22675B CPU MAC POWERMAC 8500/120		Onsite Open	Y	375.00	375.00

2072510455

NonTaxable Subtotal 0.00
Taxable Subtotal 2664.00
LOCAL / STATE @ 8.250% (S) 0.00
LOCAL / STATE @ 8.250% (K) 219.79

Total Invoice 2883.79

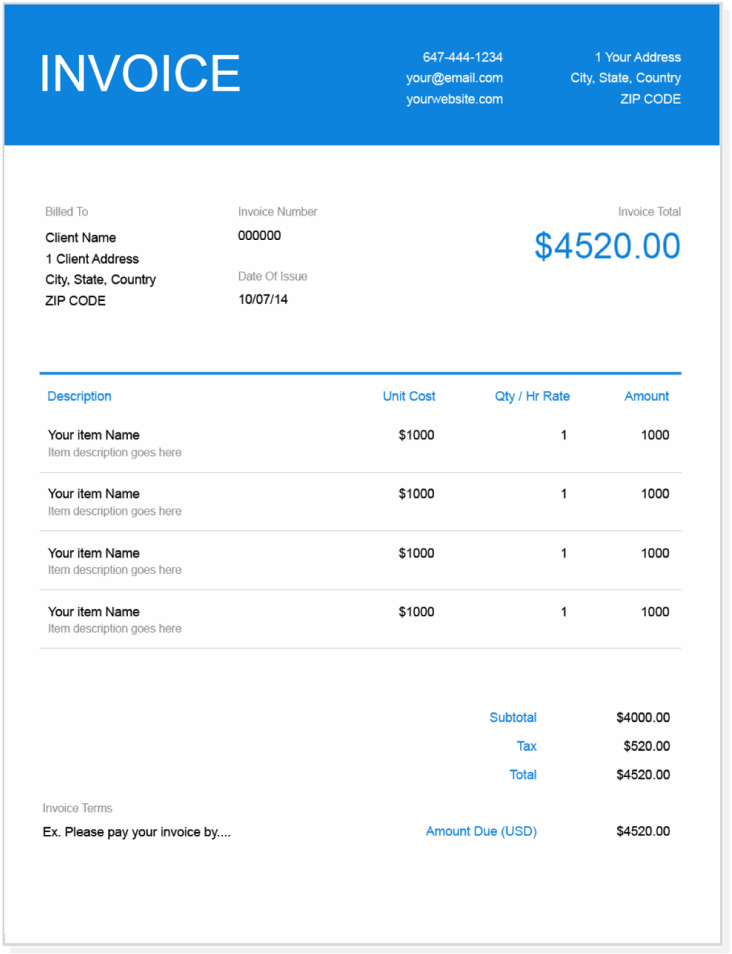
Page 1

Customer Original

Source: https://www.industrydocuments.ucsf.edu/docs/kjmn0004

- **Structured content** (e.g., table, graphs, etc.)
- **Unstructured** free text
- Their layouts are designed primarily for **human readability**

Role of Semantics and Reasoning in Information extraction



INVOICE

647-444-1234
your@email.com
yourwebsite.com

1 Your Address
City, State, Country
ZIP CODE

Billed To
Client Name
1 Client Address
City, State, Country
ZIP CODE

Invoice Number
000000

Date Of Issue
10/07/14

Invoice Total
$4520.00

Description	Unit Cost	Qty / Hr Rate	Amount
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000

Subtotal	$4000.00
Tax	$520.00
Total	$4520.00

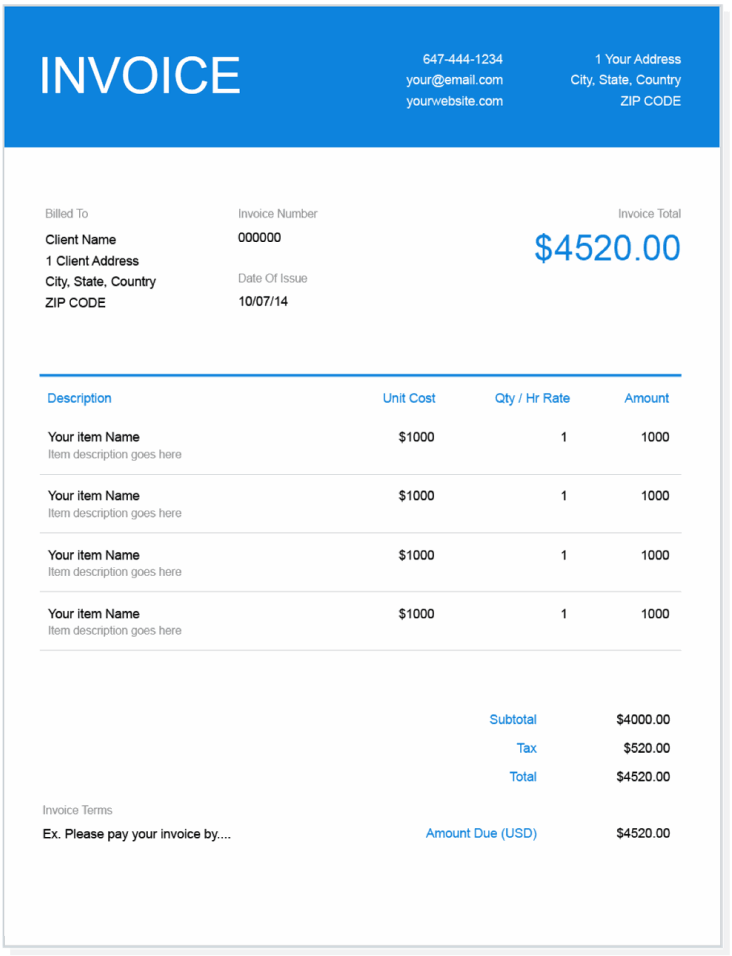
Invoice Terms
Ex. Please pay your invoice by....

Amount Due (USD) $4520.00

- **Semantics is used for Field Standardization:**
  - "invoice date" is "date of issue"
  - "net amount" is "subtotal"
  - "gross amount" is "total"
  - "payment terms" is "invoice terms"

- **Semantics and reasoning for relationship inference:**
  - Given "invoice date" and "payment terms", we can infer the "due date".
    - If payment terms is "Please pay your invoice by end of next month" and Invoice date is 28th May 2025, then due date will be 30th June 2025.

Why is context important?



INVOICE

647-444-1234
your@email.com
yourwebsite.com

1 Your Address
City, State, Country
ZIP CODE

Billed To
Client Name
1 Client Address
City, State, Country
ZIP CODE

Invoice Number
000000

Date Of Issue
10/07/14

Invoice Total
$4520.00

Description	Unit Cost	Qty / Hr Rate	Amount
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000
Your item Name Item description goes here	$1000	1	1000

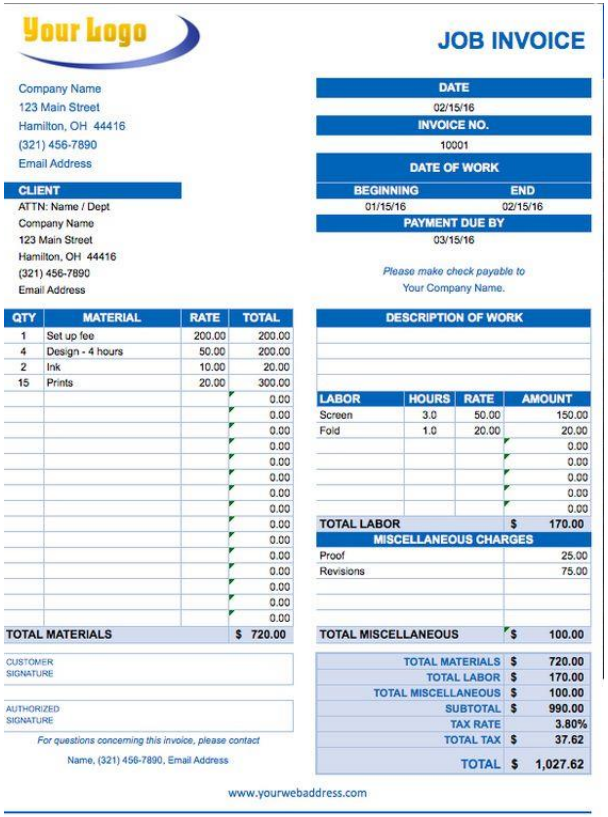
Subtotal	$4000.00
Tax	$520.00
Total	$4520.00

Invoice Terms
Ex. Please pay your invoice by....

Amount Due (USD) $4520.00

- **Context for localization and reducing hallucination:**
  - "bill to" sub-fields like bill to name and bill to address occur in the same position of the document.
  - "vendor" sub-fields like vendor name, vendor email and vendor website occur in the same position of the document.
  - If the model uses context information for prediction, it reduces misclassification.

Diversity in layouts



Your Logo

JOB INVOICE

Company Name
123 Main Street
Hamilton, OH 44416
(321) 456-7890
Email Address

DATE: 02/15/16
INVOICE NO.: 10001
DATE OF WORK: BEGINNING 01/15/16 — END 02/15/16
PAYMENT DUE BY: 03/15/16

Please make check payable to
Your Company Name.

CLIENT
ATTN: Name / Dept
Company Name
123 Main Street
Hamilton, OH 44416
(321) 456-7890
Email Address

QTY	MATERIAL	RATE	TOTAL
1	Set up fee	200.00	200.00
4	Design - 4 hours	50.00	200.00
2	Ink	10.00	20.00
15	Prints	20.00	300.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
			0.00
TOTAL MATERIALS			$ 720.00

DESCRIPTION OF WORK

LABOR	HOURS	RATE	AMOUNT
Screen	3.0	50.00	150.00
Fold	1.0	20.00	20.00
			0.00
			0.00
			0.00
			0.00
			0.00
TOTAL LABOR			$ 170.00

MISCELLANEOUS CHARGES

Proof	25.00
Revisions	75.00
TOTAL MISCELLANEOUS	$ 100.00

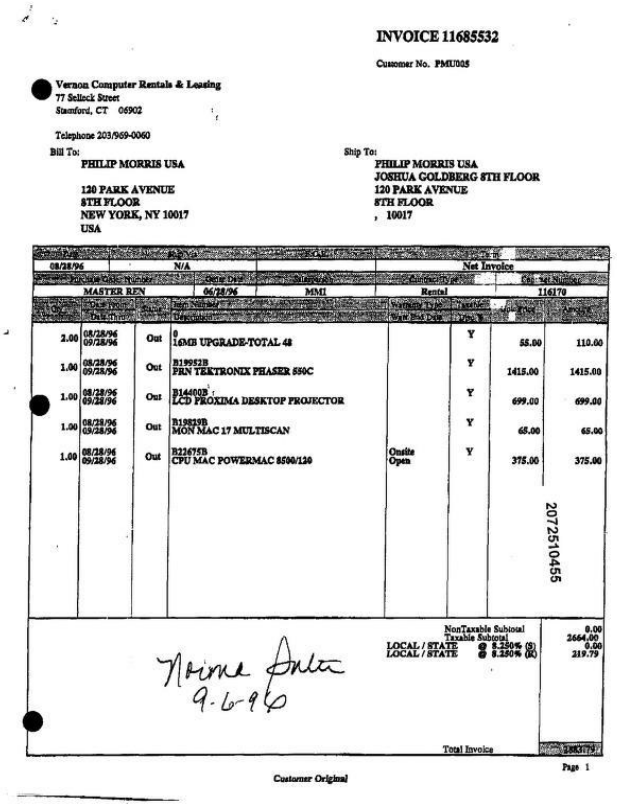
CUSTOMER SIGNATURE

AUTHORIZED SIGNATURE

For questions concerning this invoice, please contact
Name, (321) 456-7890, Email Address

TOTAL MATERIALS	$ 720.00
TOTAL LABOR	$ 170.00
TOTAL MISCELLANEOUS	$ 100.00
SUBTOTAL	$ 990.00
TAX RATE	3.80%
TOTAL TAX	$ 37.62
TOTAL	$ 1,027.62

www.yourwebaddress.com



INVOICE 11685532

Customer No. PMU005

Vernon Computer Rentals & Leasing
77 Selleck Street
Stamford, CT 06902

Telephone 203/969-0060

Bill To:
PHILIP MORRIS USA
120 PARK AVENUE
8TH FLOOR
NEW YORK, NY 10017
USA

Ship To:
PHILIP MORRIS USA
JOSHUA GOLDBERG 8TH FLOOR
120 PARK AVENUE
8TH FLOOR
, 10017

08/28/96 | N/A | Net Invoice
MASTER REN | 06/28/96 | MMI | Rental | 116170

2.00	08/28/96 09/28/96	Out	0 16MB UPGRADE-TOTAL 48		Y	55.00	110.00
1.00	08/28/96 09/28/96	Out	B19952B PRN TEKTRONIX PHASER 550C		Y	1415.00	1415.00
1.00	08/28/96 09/28/96	Out	B14400B LCD PROXIMA DESKTOP PROJECTOR		Y	699.00	699.00
1.00	08/28/96 09/28/96	Out	B19829B MON MAC 17 MULTISCAN		Y	65.00	65.00
1.00	08/28/96 09/28/96	Out	B22675B CPU MAC POWERMAC 8500/120	Onsite Open	Y	375.00	375.00

20725110455

NonTaxable Subtotal 0.00
Taxable Subtotal 2664.00
LOCAL / STATE @ 8.250% (S) 0.00
LOCAL / STATE @ 8.250% (M) 219.79

Total Invoice

Customer Original

Page 1

Source: https://www.industrydocuments.ucsf.edu/docs/kjmn0004

- Documents ***can have different layouts*** (e.g. receiver address is present at the top-left in the left invoice, while it is present at the top-right in the right invoice)
- Documents can have ***more than one tables.***
- ***Fields can have sub-fields*** (e.g. "Date of Work" field has "Beginning" and "End" subfields in the left invoice).
- **Junk Data** is also present from an end user perspective (e.g. The number "20725110455" in the left invoice).

Design Requirements for an Invoice Information Extraction Model

1. Model should extract information with a **Micro Averaged F1 score of 1.**
2. Invoices can have different layouts (e.g., the supplier address can appear at the bottom of the invoice instead of top). **Model should be able to adapt to unseen layouts at test time.**
3. Invoices can have multiple pages (up to 100 pages). **Model with should be able to extract information from multipage invoices.**
4. Invoices can have multiple tables. M**odel should be able to understand structured data and extract tables accurately.**
5. Invoices can have fields which contain sub-fields. **Model should extract all subfields accurately.**
6. Model should not hallucinate.
   - If a field is not present in the input invoice, it should be returned "None" in the extracted output.

Design Requirements for an Invoice Information Extraction Model

7. Financial institutions cannot send sensitive customer data to powerful commercial LLMs such as GPT-5, Gemini, or Claude, as these models typically operate outside of the private cloud infrastructure of the enterprise. **Model should be locally deployed.**

8. Customer data cannot be used for training models, and publicly available datasets often fails to reflect the complex layouts and formats found in real-world financial documents. Creating high-quality synthetic data that accurately represents customer information is both resource-intensive and time-consuming. **There needs to be a synthetic data generation pipeline which can mimic the real-world invoices to train the model.**

9. While large open source LLMs offer promising performance, they require significant computational resources, often demanding multiple A100 GPUs to achieve high accuracy with low latency. Unfortunately, such hardware is often unavailable in financial firms due to cost or infrastructure limitations. **Model should not use many GPU-hours.**

*Addressing these challenges requires innovative methods to balance **privacy, accuracy, and computational efficiency***

What is the difference between Information Extraction and Information Retrieval?

Feature	**Information Retrieval**	**Information Extraction**
Input	A query + a large document or collection of documents	One or more documents (unstructured or semi-structured)
Output	A set of documents (or parts thereof) ranked by relevance	Structured information (e.g., key-value pairs, entities, relations) extracted from documents
Goal	Retrieve the *right document(s)*	Extract the *right facts or information* from documents

Modern Approaches to Information Extraction

Literature Survey of Models

Model	ANLS score	Efficiency	Parameter Size	Training Data	Inference Speed	Real-World Performance
Donut	67-72% (with fine-tuning)	Moderate, OCR-free, faster than OCR-based methods	200M (base)	Synthetic (SynthDoG) + DocVQA fine-tuning	~809ms per page	Good for simple documents, limited by complex layouts
UDOP	84-85%	Efficient use of image-text token integration, requires OCR	794M (UDOP-base)	1.1M documents (unlabeled) + DocVQA and other doc datasets	Moderate, requires OCR step	Strong for multi-modal tasks (forms, invoices)
LayoutLM v1/v2/v3	72-**87%**	Moderate, requires OCR	113M (v1), 200M (v2), 368M (v3)	IIT-CDIP (11M doc images), synthetic text-image pairs	Fast post-OCR	**Widely used in enterprise (invoice, contract parsing)**
DocLLM	82.8% (7B version)	Very efficient with OCR input; uses layout encoding	1B (DocLLM-1B), 7B (DocLLM-7B)	11M document pages (IIT-CDIP), instruction-tuned on 16 datasets	Fast (1-2 seconds for 7B version)	High performance, versatile across document types
DocOwl	80.7-82.2% (1.5/2.0)	High efficiency, especially in DocOwl 2.0 with compression	8B	Two-stage learning: pre-training on documents + multi-task fine-tuning	Fast due to visual compression	Optimized for real-time document QA (multi-page, visual references)
InternVL **OCR+ GPT-4**	91.6% (InternVL 2.0, after fine-tuning); **GPT-4 ~82.8% (zero-shot)**	Computationally heavy, fast on GPUs but slower on large models	InternVL: 8B; QwenVL: 7B-14B; GPT-4 >1T	General internet data, fine-tuned on document tasks like DocVQA	Varies from seconds (smaller models) to slower with larger models	**Top performance for complex documents, suitable for enterprise use**

LayoutLM: Pre-training Text and Layout for Document Image Understanding

- LayoutLM *is a pretraining method of text and layout* for document image understanding tasks.
- Till LayoutLM, most of the document understanding (information extraction) methods confronted two limitations:
  - They relied on a few human-labelled training samples without fully exploring the possibility of using large-scale unlabelled training samples.
  - They usually leveraged either pre-trained Computer vision models or NLP models but *did not consider a joint training of textual and layout information*. Therefore, it was important to investigate how self-supervised pre-training of text and layout may help in this area.

Ref: *Xu, Yiheng, et al. "Layoutlm: Pre-training of text and layout for document image understanding." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.*

LayoutLM: Pre-training Text and Layout for Document Image Understanding

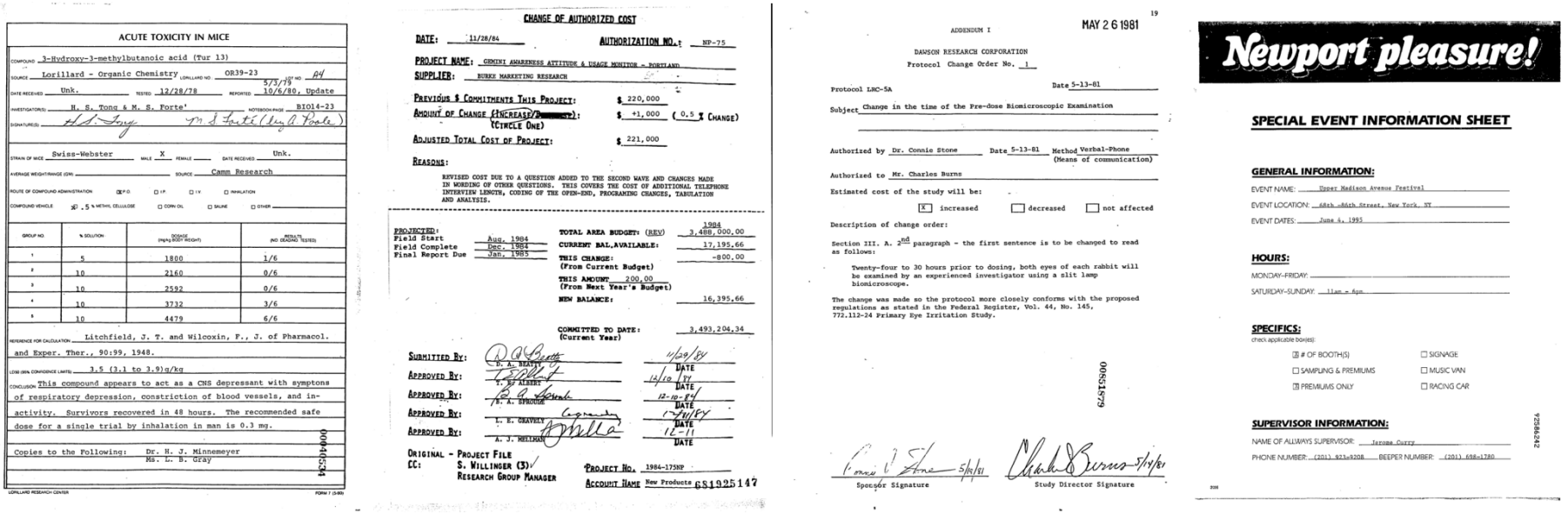
BERT (Bidirectional encoder representations from transformers):

- Introduced by Google in 2018, where *they claimed its conceptually simple and empirically powerful.*
- Architecture Overview
  - Encoder-only Transformer (no decoder) and processes input bidirectionally.
  - BASE = 12 encoder layers, hidden size 768, 12 attention heads.
- Input Design (for two-sentence modelling)
  - Sequence format: `[CLS]` Sentence A `[SEP]` Sentence B `[SEP]`.
  - Two segments allow "Next Sentence Prediction".
- Tokeniser: WordPiece; up to 512 tokens.
- Pre-training Objectives
  - Masked Language Modelling (MLM) – randomly mask ~15 % of tokens in sequence and predict them using both left and right context.
  - Next Sentence Prediction (NSP) – given two input sentences, predict whether the second follows the first in the source corpus.
- Why it matters
  - Helps the model learn deep contextual representations (not just left-to-right or right-to-left), both sides of context at once.
  - Once pre-trained, BERT can be *fine-tuned* for downstream tasks (classification, QA, NER) by adding a task-specific head.

LayoutLM: Pre-training Text and Layout forDocument Image Understanding

Does a masked language model converge faster than a left to right language model?

LayoutLM: Pre-training Text and Layout for Document Image Understanding



ACUTE TOXICITY IN MICE

CHANGE OF AUTHORIZED COST

ADDENDUM I

Newport pleasure!

- Business/legal/academic documents are visually rich i.e. apart from text, they have layout, tables and forms.
- Although BERT Like models have become SOTA in many NLP tasks, they take text input and ignore layout information.

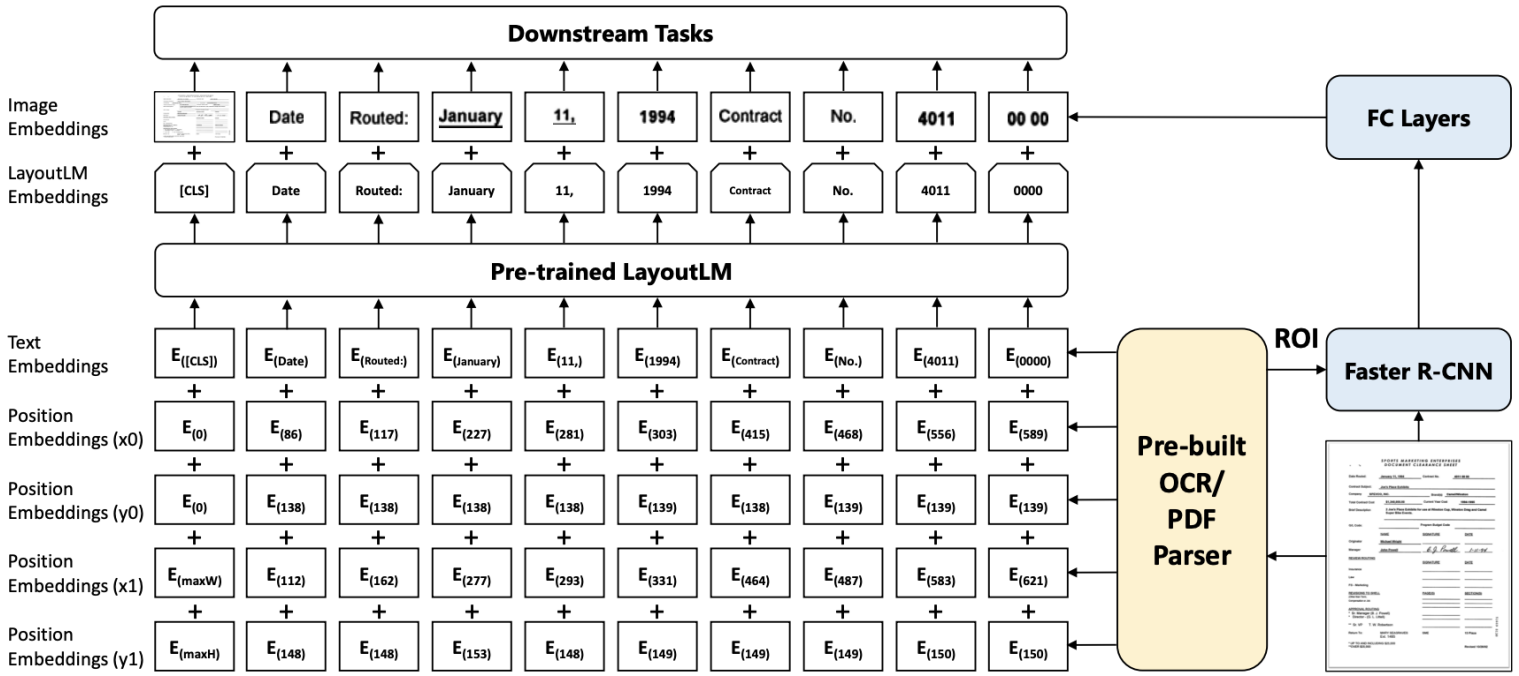
LayoutLM: Pre-training Text and Layout for Document Image Understanding

- Features which substantially improve language representations in document:
  - Document layout information: Relative 2-D position of words contribute a lot to semantic representation.
    E.g. for the key "Passport Number: ",
    its corresponding value will appear ***at the bottom*** *or* ***right****.*
  - Visual information:
    - Document Image: Indicates layout which helps in classification.
    - Word level: Colour, style, etc. which provides hints in sequence labelling tasks.

LayoutLM: Pre-training Text and Layout forDocument Image Understanding

- Pretraining dataset used in LayoutLM
  - IIT CDIP Test Collection 1.0:
    - Contains more than 6 million scanned documents with 11 million scanned document images. The scanned documents are in a variety of categories, including letter, memo, email, file folder, form, handwritten, invoice, advertisement, budget, news articles, presentation, scientific publication, questionnaire, resume, scientific report, specification, and many others.
    - Each document has its corresponding text and metadata stored in XML files.
    - Tesseract, an open-source OCR engine, is used to obtain the tokens and their 2-D positions.

LayoutLM: Pre-training Text + Layout for Document Image Understanding



LayoutLM is initialised from BERT BASE and adds two types of embeddings to the text embedding:

1. 2D position embedding: Denotes relative ***spatial*** position of a token in the document
2. Image Embedding: Vector representation of scanned token images to capture font directions, types, and colour.

LayoutLM: Pre-training Text + Layout for Document Image Understanding

- 2D position embedding
  - Document page is a coordinate system with the top left origin.
  - In this setting, the bounding box can be precisely defined by (x0, y0, x1, y1), where (x0, y0) corresponds to the position of the upper left in the bounding box, and (x1, y1) represents the position of the lower right.
  - 4 position embedding layers are added with two embedding tables, where the embedding layers representing the same dimension share the same embedding table.

LayoutLM: Pre-training Text + Layout for Document Image Understanding

- Image Embeddings:
  - Image is split into several pieces using the bounding box of each word from OCR results, and they have a one-to-one correspondence with the words.
  - Token image embeddings are generated with these pieces of images using the Faster R-CNN model.
  - [CLS] token embeddings is generated with whole document image using the Faster R-CNN model.

LayoutLM: Pre-training Text + Layout for Document Image Understanding

- Pretraining Strategies:
  - Masked Visual-Language Model
    - How it works?
      - Randomly **mask** 15% of input tokens.
      - Keep their 2-D position embeddings (x_0, y_0, x_1, y_1) and image features (if any).
      - Model predicts the masked tokens using context + relative spatial position of the masked token on the page.
      - **Objective:** minimize cross-entropy loss between predicted and true tokens.
    - Effect of keeping 2-D positional embeddings unmasked**:**
      - The model learns correlations such as *"text near the top-right corner is often an invoice number"* or *"numbers aligned in a column are probably totals"*. This bridges the gap between **visual layout** and **language context**, enabling the model to reason spatially, not just linguistically.

LayoutLM: Pre-training Text + Layout for Document Image Understanding

- Pretraining Strategies:
  - Multi-label Document Classification
    - Motivation
      - Many documents carry multiple semantic tags (e.g., “invoice”, “form”, “letter”).
      - The model should learn a **document-level representation** that captures content and structure for classification.
    - How it works?
      - Use IIT-CDIP dataset with document tags.
      - Feed entire document into LayoutLM; use [CLS] token embedding as global representation.
      - Apply multi-label classification head (sigmoid activation) → predict multiple tags per document.
      - **Objective**: Binary Cross-Entropy (BCE) per label, summed across tags
    - Outcomes?
      - Encourages model to learn document-level knowledge clusters.
      - **Produces stronger [CLS] embeddings useful for downstream tasks.** Optional, used only when labelled tags are available

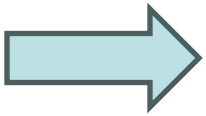
LayoutLM: Pre-training Text + Layout for Document Image Understanding

- The pre-trained LayoutLM model is fine-tuned on three document image understanding tasks including:
  - Form understanding task
  - Receipt understanding task
  - Document image classification task.
- For the form and receipt understanding tasks, LayoutLM predicts {B, I, E, S, O} tags for each token and uses sequential labelling to detect each type of entity in the dataset.
- For the document image classification task, LayoutLM predicts the class labels using the representation of the [CLS] token.

LayoutLM: Pre-training Text +Layout for Document Image Understanding

Total Amount: $25.30

Date: 2021-08-01



Token	Tag
Total	B-TOTAL
Amount	E-TOTAL
:	O
$	B-AMOUNT
25	I-AMOUNT
.30	E-AMOUNT
Date	B-DATE
:	O
2021-08-01	S-DATE

B = Beginning of an Entity

E = End of an Entity

O = Outside any Entity

I = Inside an Entity

S = Single token Entity

LayoutLM: Pre-training Text and Layout for Document Image Understanding

Finetuning datasets used in LayoutLM:

- **FUNSD Dataset:** includes 199 real, fully annotated, scanned forms with 9,707 semantic entities and 31,485 words. These forms are organized as a list of semantic entities that are interlinked. Each semantic entity comprises a unique identifier, a label (i.e., question, answer, header, or other), a bounding box, a list of links with other entities, and a list of words. The dataset is split into 149 training samples and 50 testing samples. Word-level F1 score is adopted as the evaluation metric.
- **SROIE Dataset:** Dataset contains 626 receipts for training and 347 receipts for testing. Each receipt is organized as a list of text lines with bounding boxes. Each receipt is labeled with four types of entities which are {company, date, address, total}. The evaluation metric is the exact match of the entity recognition results in the F1 score.
- **RVL-CDIP Dataset:** Dataset consists of 400,000 grayscale images in 16 classes, with 25,000 images per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images. The images are resized, so their largest dimension does not exceed 1,000 pixels. The 16 classes include {letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo}. The evaluation metric is the overall classification accuracy.

LayoutLM: Pre-training Text and Layout for Document Image Understanding

Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	0.5469	0.671	0.6026	110M
	$\text{RoBERTa}_{\text{BASE}}$	0.6349	0.6975	0.6648	125M
	$\text{BERT}_{\text{LARGE}}$	0.6113	0.7085	0.6563	340M
	$\text{RoBERTa}_{\text{LARGE}}$	0.678	0.7391	0.7072	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.665	0.7355	0.6985	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.7377	0.782	0.7592	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.7597	0.8155	0.7866	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7076	0.7695	0.7372	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.7194	0.7780	0.7475	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.7171	0.805	0.7585	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	0.7536	0.806	0.7789	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7101	0.7815	0.7441	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**0.7677**	**0.8195**	**0.7927**	160M

Table 1: Model accuracy (Precision, Recall, F1) on the FUNSD dataset

LayoutLM: Pre-training Text and Layout for Document Image Understanding

Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	0.9099	0.9099	0.9099	110M
	$\text{RoBERTa}_{\text{BASE}}$	0.9107	0.9107	0.9107	125M
	$\text{BERT}_{\text{LARGE}}$	0.9200	0.9200	0.9200	340M
	$\text{RoBERTa}_{\text{LARGE}}$	0.9280	0.9280	0.9280	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.9388	0.9388	0.9388	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.9380	0.9380	0.9380	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.9431	0.9431	0.9431	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.9438	0.9438	0.9438	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.9402	0.9402	0.9402	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.9460	0.9460	0.9460	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.9416	0.9416	0.9416	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	**0.9524**	**0.9524**	**0.9524**	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.9416	0.9416	0.9416	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.9467	0.9467	0.9467	160M
Baseline	Ranking 1^{st} in SROIE	0.9402	0.9402	0.9402	-

Table 4: Model accuracy (Precision, Recall, F1) on the SROIE dataset

LayoutLM: Pre-training Text and Layout for Document Image Understanding

Modality	Model	Accuracy	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	89.81%	110M
	$\text{RoBERTa}_{\text{BASE}}$	90.06%	125M
	$\text{BERT}_{\text{LARGE}}$	89.92%	340M
	$\text{RoBERTa}_{\text{LARGE}}$	90.11%	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	91.25%	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	91.48%	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	91.65%	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	91.78%	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	91.74%	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	91.78%	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	91.88%	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	91.90%	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	94.31%	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**94.42%**	160M
Baselines	VGG-16 [1]	90.97%	-
	Stacked CNN Single [2]	91.11%	-
	Stacked CNN Ensemble [2]	92.21%	-
	InceptionResNetV2 [25]	92.63%	-
	LadderNet [20]	92.77%	-
	Multimodal Single [3]	93.03%	-
	Multimodal Ensemble [3]	93.07%	-

Table 5: Classification accuracy on the RVL-CDIP dataset

LayoutLM: Pre-training Text and Layout for Document Image Understanding

Does LayoutLM use image embeddings during Pretraining stage?

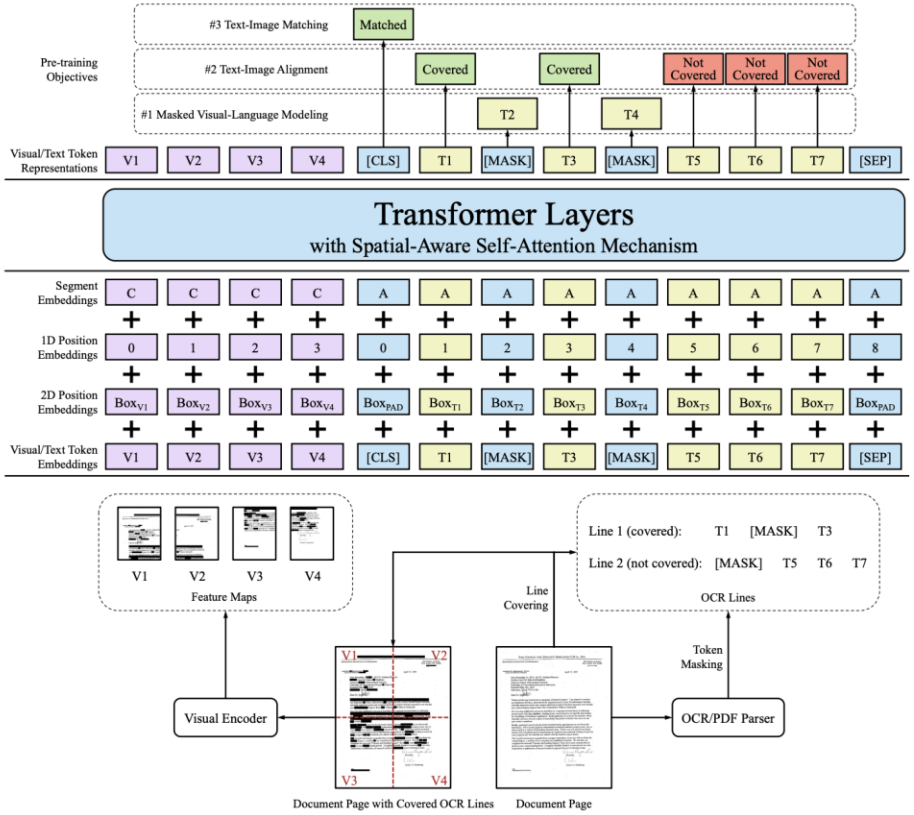
LayoutLM: Pre-training Text and LayoutforDocument Image Understanding

Does LayoutLM use image embeddings during finetuning stage for Form understanding and Receipt Task?

LayoutLM: Pre-training Text + Layout for Document Image Understanding

- How to use LayoutLM for information extraction (IE)?
  - Document image → OCR → tokens + bounding-boxes
  - Each token gets:
    - token/text embedding
    - 2-D spatial (bounding-box) embedding
  - Transformer processes combined embeddings → contextualised token vectors
  - Token classification head (linear + softmax) predicts field labels for IE
  - Even when a semantic-labelled entity spans **multiple tokens**, the model uses token-level classification:
    - Use a schema like IOB (B = begin, I = inside) → e.g. “B-VENDOR_NAME”, “I-VENDOR_NAME”
    - After prediction, adjacent tokens with same entity type are merged into the complete entity span
  - This lets us handle multi-token field values within the sequence-labelling framework

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding



Key differences from LayoutLM:

1. Uses Image Embeddings during Pretraining stage.
2. Spatial-aware self-attention mechanism which involves a 2-D relative position representation for token pairs.

Ref: *Xu, Yang, et al. "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding." arXiv preprint arXiv:2012.14740 (2020).*

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

- **Text Embedding ti** = TokEmb(wi) + PosEmb1D(i) + SegEmb(si)
- **Visual Embedding vi** = Proj(VisTokEmb(I)i) + PosEmb1D(i) + SegEmb([C])

Step	Operation	Input Shape	Output Shape	Purpose
1	CNN (ResNeXt-FPN)	[3, 224, 224]	[C, H′, W′]	Extract visual features
2	Average Pooling → Fixed Grid	[C, H′, W′]	[C, H, W]	Uniform visual grid
3	Flatten	[C, H, W]	[H×W, C]	Sequence of visual tokens
4	Linear Projection	[H×W, C]	[H×W, d_model]	Match text embedding dim

- **Layout Embedding li** = Concat(PosEmb2Dx(xmin, xmax, width), PosEmb2Dy(ymin, ymax, height))
- **Note:** Layout embedding is projected down from 6x768to 768 using a Linear Layer.

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

Why do we perform Average Pooling while calculating Visual token Embedding?

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

- X(0) = {v0, …, VWH-1, t0, …, tL} + {L0, ……, LWH-1+L}
- Qi =xi WQ, Ki =xi WK , Vi=xi WV
- Vanilla attention αij = QiKjT / sqrt(dhead)
- Why is this not enough?
  - The model only knows the *index difference* (j − i), not the *spatial distance* between tokens.
  - So it can't tell whether one token is *above*, *below*, or *beside* another.

Bias	**Meaning**	**Indexing**
(b{(1D)})	1D relative position bias (like in T5)	depends on (j - i)
(b{(2D_x)})	horizontal (x-axis) relative bias	depends on (x_j - x_i)
(b{(2D_y)})	vertical (y-axis) relative bias	depends on (y_j - y_i)

- αij′ =αij +b(j−i)(1D) +b(xj −xi)(2Dx) +b(yj −yi)(2Dy)
- After adding the spatial biases, we normalize and apply the usual softmax attention.

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

The **spatial bias** modifies the *attention mechanism itself*, i.e. it changes *how much one token attends to another* depending on their **relative** 2D distance.

- Biases control *who attends to whom* (interaction strength).
- Embeddings control *what the token's representation initially encodes* (positional features).

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

- Pretraining tasks:
  - Masked Visual-Language Modelling:
    - Randomly mask some text tokens and ask the model to recover the masked tokens. Meanwhile, the layout information remains unchanged, which means the model knows each masked token's location on the page.
    - **To avoid visual clue leakage, image regions corresponding to masked tokens are masked on the raw page image input before feeding it into the visual encoder.**
    - The output representations of masked tokens from the encoder are fed into a classifier over the whole vocabulary, driven by a cross-entropy loss.

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

- Pretraining tasks:
  - Text-Image Alignment (TIA): Teach the model to link *specific text tokens* to their *exact regions* on the document image.
    - Randomly pick some **text lines**. Why?
      OCR word boxes can be small and noisy; covering whole lines is more stable visually.
    - **Cover (occlude)** their corresponding regions on the input page image (like hiding them).
    - For each token, the model must classify:
      [Covered] or **[Not Covered]**.
    - This is done using a classifier on top of encoder outputs → binary cross-entropy loss
    - **Note:** If a token is *also masked* in MVLM (Masked Visual Language Modelling), its TIA loss is ignored, so the model doesn't trivially learn "if [MASK] → Covered.
    - In TIA, 15% of the lines are covered.

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

- Pretraining tasks:
  - Text–Image Matching (TIM): Teach the model to understand if the *whole image* matches the *text sequence* — global alignment.
  - **How it works:**
    - Feed the [CLS] embedding into a classifier → predicts **Same document page? (Yes/No)**
    - Positive = text & image from the same page.
    - Negative = image replaced with another page or dropped entirely.
  - In TIM, 15% images are replaced, and 5% are dropped

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

Dataset	# of keys or categories	# of examples (train/dev/test)
IIT-CDIP	–	11M/0/0
FUNSD	4	149/0/50
CORD	30	800/100/100
SROIE	4	626/0/347
Kleister-NDA	4	254/83/203
RVL-CDIP	16	320K/4K/4K
DocVQA	–	39K/5K/5K

Pretraining

Finetuning

Table 1: Statistics of datasets

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

Model	FUNSD	CORD	SROIE	Kleister-NDA
$\text{BERT}_{\text{BASE}}$	0.6026	0.8968	0.9099	0.7790
$\text{UniLMv2}_{\text{BASE}}$	0.6890	0.9092	0.9459	0.7950
$\text{BERT}_{\text{LARGE}}$	0.6563	0.9025	0.9200	0.7910
$\text{UniLMv2}_{\text{LARGE}}$	0.7257	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

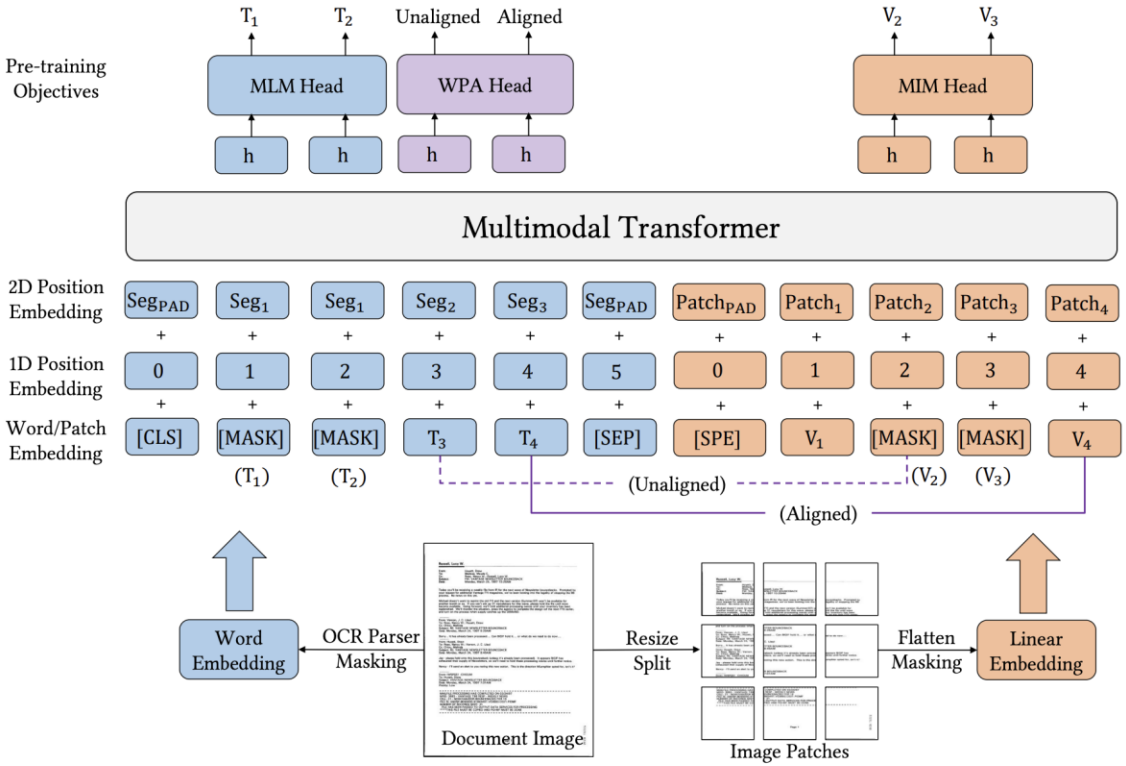
Table 2: Entity-level F1 scores of the four entity extraction tasks: FUNSD, CORD, SROIE and Kleister-NDA. Detailed per-task results are in the Appendix.

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

#	Model Architecture	Initialization	SASAM	MVLM	TIA	TIM	ANLS
1	$\text{LayoutLM}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$		✓			0.6841
2a	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓			0.6915
2b	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓	✓		0.7061
2c	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓		✓	0.6955
2d	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓	✓	✓	0.7124
3	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN	✓	✓	✓	✓	0.7217
4	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{UniLMv2}_{\text{BASE}}$ + X101-FPN	✓	✓	✓	✓	0.7421

Table 5: Ablation study on the DocVQA dataset, where ANLS scores on the validation set are reported. "SASAM" means the spatial-aware self-attention mechanism. "MVLM", "TIA" and "TIM" are the three pre-training tasks. All the models are trained using the whole pre-training dataset for one epoch with the BASE model size.

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking



LayoutLMv3 is a unified multimodal Transformer

Ref: *Huang, Yupan, et al. "Layoutlmv3: Pre-training for document ai with unified text and image masking." Proceedings of the 30th ACM international conference on multimedia. 2022.*

LayoutLMv3: Pre-training for Document AI withUnified Text and Image Masking

- TextEmbi = WordEmbi +Pos1DEmbi +Layout2DEmbi
  - WordEmbi is Initialized from **RoBERTa's** pretrained word embedding matrix.
  - LayoutLMv3 uses segment-level layout embeddings (LayoutLMv1/v2 used word level embeddings):
    - Words in the same text line or logical block share the same 2D layout position.
    - This reduces noise and redundancy, since words in a line typically form one semantic unit

LayoutLMv3: Pre-training for Document AIwithUnified Text and Image Masking

- Vision Embedding:
  - **Resize** the document image to a fixed resolution H×W
  - **Split** it into patches of size P×P.
    Number of patches = HxW/PxP
  - **Linear Projection:** Each patch is flattened and linearly projected into a vector of dimension D (same as Transformer hidden size).
  - **Flatten** all patches → form a sequence of MMM image tokens.
  - **Add learnable 1D position embeddings**

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

LayoutLMv3 has the same spatial-aware attention from LayoutLMv2.

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

Pretraining objectives:

Task	**What is masked**	**Model must reconstruct**	**Purpose**
MLM (Masked Language Modeling)	Text tokens	Masked words	Learn textual semantics
MIM (Masked Image Modeling)	Image patches	Masked visual patches	Learn visual perception
WPA (Word–Patch Alignment)	None (uses existing features)	Predict if a text token matches its visual patch	Learn cross-modal alignment

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

- Masked Language Modeling (MLM)
  - Randomly mask **30 %** of text tokens (word pieces).
  - Use *span masking*: mask contiguous runs of words; span lengths follow a **Poisson(λ = 3)** distribution. Why 3?
    - λ = 3 was empirically found to work well in **SpanBERT**, which inspired LayoutLMv3's masking.
  - Keep layout coordinates unchanged.
  - Input: corrupted sequence of text + image tokens
  - Output: predict the original masked words.
  - *The model learns rich textual context and how text meaning relates to its 2D location and surrounding image region.*

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

- Masked Image Modeling (MIM)
  - Randomly mask **≈ 40 %** of image patches using *blockwise* masking (whole contiguous regions).
  - Model reconstructs the discrete "visual tokens" of those masked patches.
  - Each image patch is first quantized by a **pretrained image tokenizer** (like a VQ-VAE or dVAE from **DiT**) into a vocabulary of **8192 discrete tokens**.
  - The model predicts those discrete token IDs, not raw pixels.
  - Encourages LayoutLMv3 to capture *high-level layout and structural cues* (tables, lines, blocks), not just pixel details.
- Word–Patch Alignment (WPA):
  - For each unmasked text token, Check if itscorresponding image patch(es) are **also unmasked.**
    - If Yes→ label = **Aligned (1)**
    - If No → label = **Unaligned (0)**

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

Table 1: Comparison with existing published models on the CORD [39], FUNSD [20], RVL-CDIP [16], and DocVQA [38] datasets. "T/L/I" denotes "text/layout/image" modality. "R/G/P" denotes "region/grid/patch" image embedding. We multiply all values by a hundred for better readability. †In the UDoc paper [14], the CORD splits are 626/247 receipts for training/test instead of the official 800/100 training/test receipts adopted by other works. Thus the score† is not directly comparable to other scores. Models denoted with ‡ use more data to train DocVQA and are expected to score higher. For example, TILT introduces one more supervised training stage on more QA datasets [40]. StructuralLM additionally uses the validation set in training [28].

Model	Parameters	Modality	Image Embedding	FUNSD F1↑	CORD F1↑	RVL-CDIP Accuracy↑	DocVQA ANLS↑
$\text{BERT}_{\text{BASE}}$ [9]	110M	T	None	60.26	89.68	89.81	63.72
$\text{RoBERTa}_{\text{BASE}}$ [36]	125M	T	None	66.48	93.54	90.06	66.42
$\text{BROS}_{\text{BASE}}$ [17]	110M	T+L	None	83.05	95.73	-	-
$\text{LiLT}_{\text{BASE}}$ [50]	-	T+L	None	88.41	96.07	95.68*	-
$\text{LayoutLM}_{\text{BASE}}$ [54]	160M	T+L+I (R)	ResNet-101 (fine-tune)	79.27	-	94.42	-
SelfDoc [31]	-	T+L+I (R)	ResNeXt-101	83.36	-	92.81	-
UDoc [14]	272M	T+L+I (R)	ResNet-50	87.93	98.94†	95.05	-
$\text{TILT}_{\text{BASE}}$ [40]	230M	T+L+I (R)	U-Net	-	95.11	95.25	83.92‡
$\text{XYLayoutLM}_{\text{BASE}}$ [15]	-	T+L+I (G)	ResNeXt-101	83.35	-	-	-
$\text{LayoutLMv2}_{\text{BASE}}$ [56]	200M	T+L+I (G)	ResNeXt101-FPN	82.76	94.95	95.25	78.08
$\text{DocFormer}_{\text{BASE}}$ [2]	183M	T+L+I (G)	ResNet-50	83.34	96.33	**96.17**	-
$\text{LayoutLMv3}_{\text{BASE}}$ (Ours)	133M	T+L+I (P)	Linear	**90.29**	**96.56**	95.44	**78.76**
$\text{BERT}_{\text{LARGE}}$ [9]	340M	T	None	65.63	90.25	89.92	67.45
$\text{RoBERTa}_{\text{LARGE}}$ [36]	355M	T	None	70.72	93.80	90.11	69.52
$\text{LayoutLM}_{\text{LARGE}}$ [54]	343M	T+L	None	77.89	-	91.90	-
$\text{BROS}_{\text{LARGE}}$ [17]	340M	T+L	None	84.52	97.40	-	-
$\text{StructuralLM}_{\text{LARGE}}$ [28]	355M	T+L	None	85.14	-	**96.08**	83.94‡
FormNet [25]	217M	T+L	None	84.69	-	-	-
FormNet [25]	345M	T+L	None	-	97.28	-	-
$\text{TILT}_{\text{LARGE}}$ [40]	780M	T+L+I (R)	U-Net	-	96.33	95.52	87.05‡
$\text{LayoutLMv2}_{\text{LARGE}}$ [56]	426M	T+L+I (G)	ResNeXt101-FPN	84.20	96.01	95.64	**83.48**
$\text{DocFormer}_{\text{LARGE}}$ [2]	536M	T+L+I (G)	ResNet-50	84.55	96.99	95.50	-
$\text{LayoutLMv3}_{\text{LARGE}}$ (Ours)	368M	T+L+I (P)	Linear	**92.08**	**97.46**	95.93	83.37

* LiLT uses image features with ResNeXt101-FPN backbone in fine-tuning RVL-CDIP.

Thank you