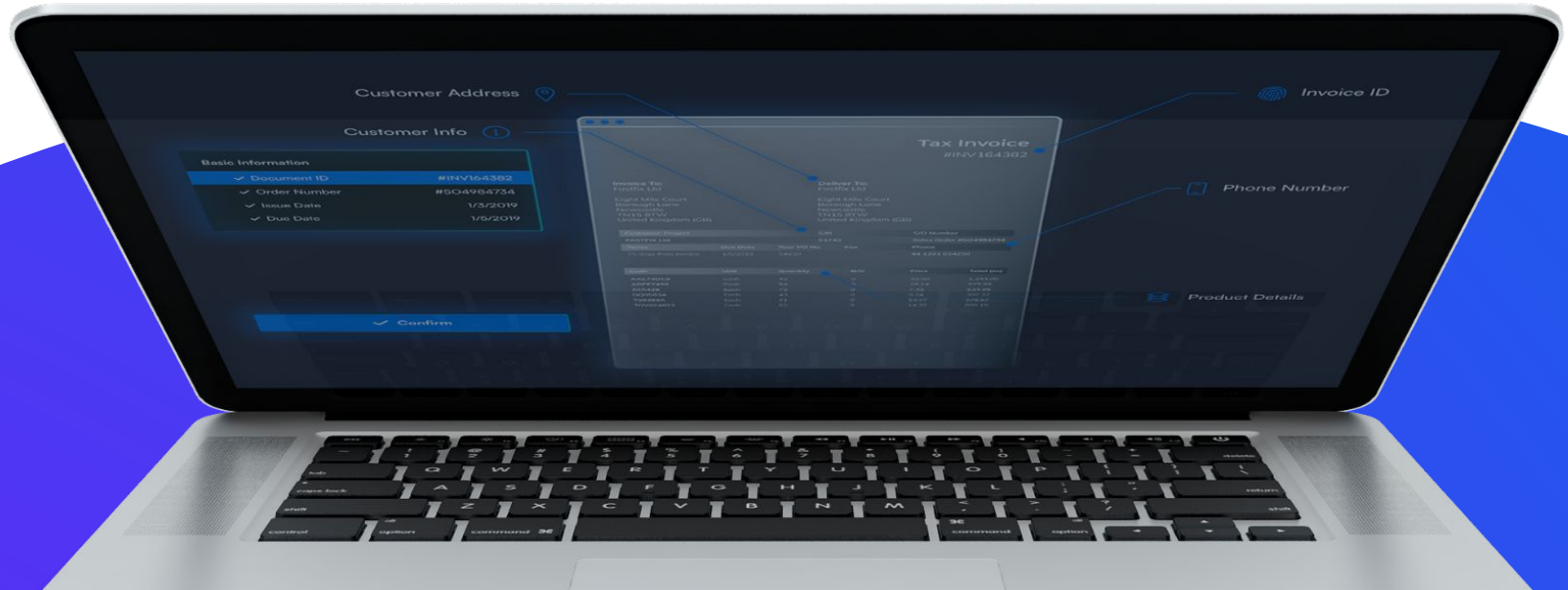


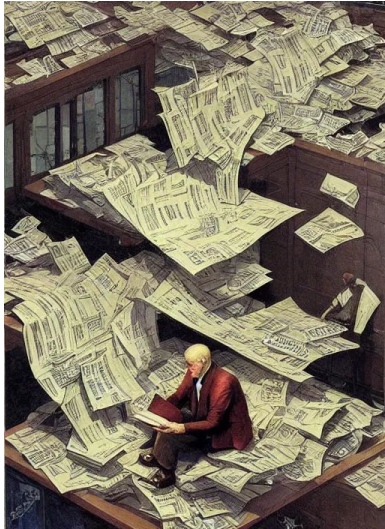
Teaching LLMs to Understand Business Documents

AI Days 2024 - 25/01/2024 - Tomáš Tunys (tomas.tunys@rossum.ai)

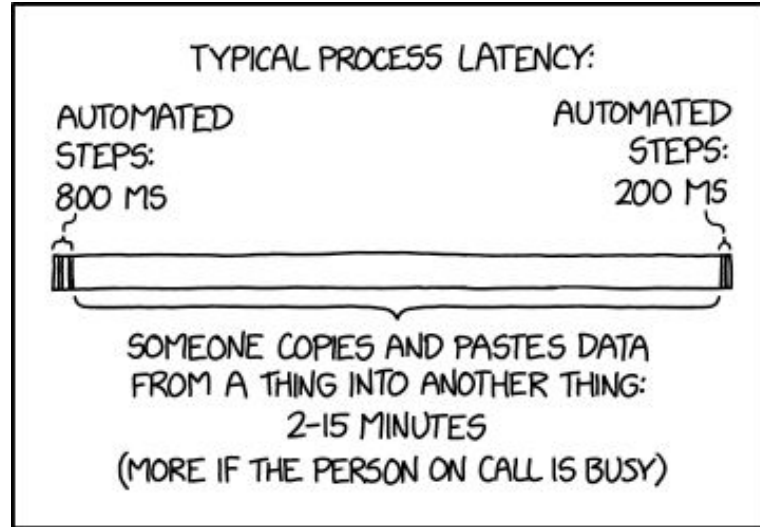


Context: Rossum Mission

A person to handle **a million transactions** in **a year**.



* Disclaimer: AI may or may not have been used to generate this image.



Source: <https://xkcd.com/2565>

Problem: Key Information Localisation & Extraction

INVOICE



VENDOR DETAILS

Lacté

60 Route de Luzinay

38540 Paris

lacte.incredible.bureau@gmail.com

VAT ID FR5684583

BILL TO

Dairy Products Ltd

12-13 Waterloo Rd

London NW2 7UH, UK

+44 7911 123456

Due date

SHIP TO

Dairy Products Ltd

Harbour House 12

Marine Parade, Dover CT17

+44 7911 123471

DATE

10/1/2021

INVOICE NO.

435321

10/6/2021

vendor address

Lacté

60 Route de Luzinay

38540 Paris

vendor tax ID

FR5684583

date issue

10/1/2021

document ID

435321

date due

10/6/2021

customer delivery address

Dairy Products Ltd

Harbour House 12

Marine Parade, Dover CT17

customer billing address

Dairy Products Ltd

12-13 Waterloo Rd

London NW2 7UH, UK

Problem: Key Information Localisation & Extraction

The screenshot displays a software interface for reviewing an invoice. The main area contains a table of items with columns for Code, Description, Unit, Quantity, B.I.O., Unit Price, and Total (ex). Annotations with arrows point to specific data points in the table, such as 'Unit', 'Quantity', 'B.I.O.', 'Unit Price', and 'Total (ex)'. A 'Review' sidebar on the left lists various address and customer details. A 'Line Items' section at the bottom has a button labeled 'Extract complex line items'. A 'Notes' section is also visible. The interface is titled 'AU Marketing Invoice.pdf' and 'Page 1 / 2'. A footer note states 'AUSTRALIAN OWNED AND MADE PIPE FLASHINGS'.

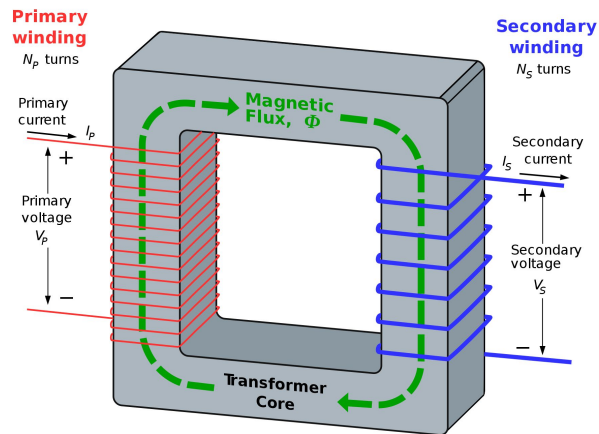
Code	Description	Unit	Quantity	B.I.O.	Unit Price	Total (ex)
AAE10B540-730	Aquadapt #10 EPDM Black 400-730mm	Each	42	0	32.00	1,344.00
	Aquadapt #10 EPDM Black 400-730mm	Each	54	0	18.14	979.56
AATFG75-325	Aquadapt #4 EPDM Grey 75-135mm	Each	72	0	7.43	534.96
AFE4190-120	FlashRite #4 EPDM Grey 90-150mm	Each	43	0	9.24	397.32
AAS4H75-154	Aquadapt #4 Silicone Grey 75-135mm	Each	21	0	13.27	278.67
AFT6H150-232	FlashRite #6 Silicone Red 150-230mm	Each	62	0	14.39	892.18
ASE1RS-37	Special EPDM #1 Black 5-45mm	Each	300	0	3.54	1,062
VETF4B-S	#4 Alluri EPDM Black 50-170mm 5 pk	5pk	3	0	43.67	131.01

Approach: Our ML Pipeline to Solve KILE

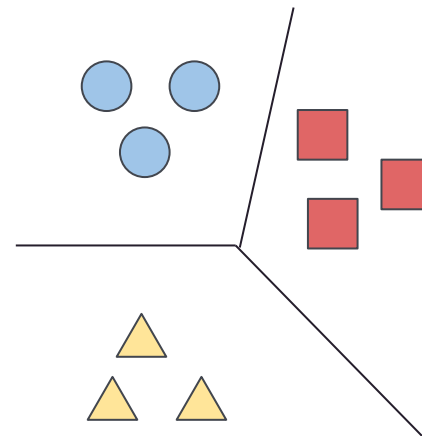
Step 1: OCR



Step 2: Embedding



Step 3: Classification

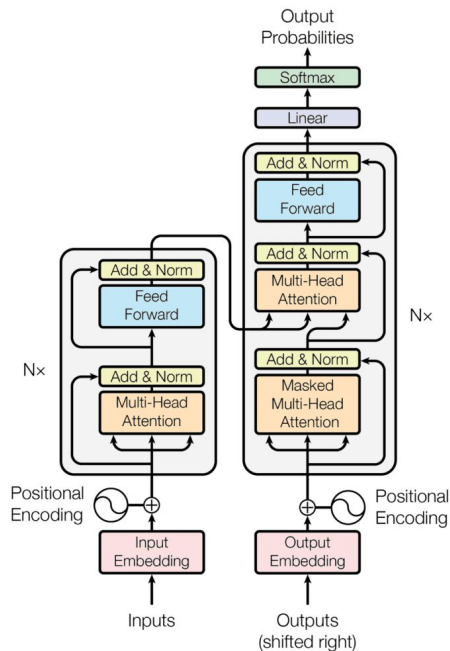


Approach: Our ML Pipeline to Solve KILE

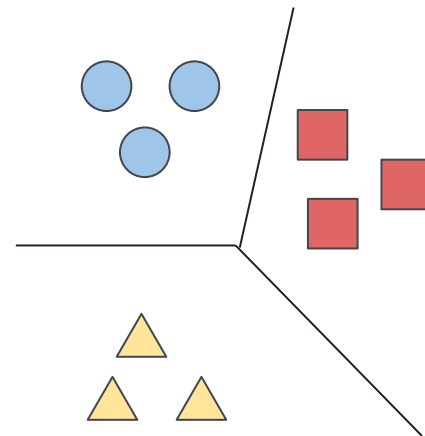
Step 1: OCR



Step 2: Embedding



Step 3: Classification



Step 1: OCR

“That’s 40-year old problem that is solved”
– our beginnings at StartupYard (2016)

Yann LeCun @ylercun
Nice.
This used to be really, really hard.

Vik Paruchuri @VikParuchuri · Jan 12
Announcing surya - a multilingual text line detection model for documents. It gives you accurate line-level bboxes and column breaks.
Find it here - github.com/VikParuchuri/s...

And Restricting Trade
U.N. Pals Russia From Rights Council
(Amid New Evidence of Atrocities)

First Black Woman to Serve as Justice
— Vote Is 53-47

By CARL BUELL
and NARI KARAN

INVOICE



VENDOR DETAILS
NAME: Lacté
ADDRESS: 60 Route de Luzinay
CITY: 93540 Paris
EMAIL: lacte.incredible.bureau@gmail.com
VAT ID: FR5684583

DATE: 10/1/2021
INVOICE NO.: 435321

BILL TO
Dairy Products Ltd
12-13 Waterloo Rd
London NW2 7UH, UK
+44 7911 123456

SHIP TO
Dairy Products Ltd
Harbour House 12
Marine Parade, Dover CT17 9BU
+44 7911 123471

ITEM	DESCRIPTION	QTY	UNIT PRICE	TOTAL
001	SKU-582372 Crème liquide semi épaisse légère 4% Mat.Gr. Country of Origin: France	90 ltr	19.20	1 728.00
002	SKU-989834 Beurre Demi-sel Moulé de Bretagne Allergens: Lactose Country of Origin: France	100 pcs	5.26	526.00
003	SKU-649293 Beurre Doux Tendre Country of Origin: France	70 pcs	4.80	336.00

Step 1: OCR

Text Localisation: variant of DBNet
([github:open-mmlab/mmlab](https://github.com/open-mmlab/mmlab))

Text Detection: variant of OCKRE
([github:rossumai/OCKRE](https://github.com/rossumai/OCKRE))

Training Data: Use PDFs + PDF parser
([pypi:python-poppler](https://pypi.org/project/python-poppler/))

Rossum history remark

- We used to crawl the Internet and scrape Uloz.to for PDF documents to train our first version of OCKRE.



The image shows a scanned invoice for Dairy Products Ltd. The invoice is titled 'INVOICE' and includes a barcode icon in the top right corner. The document is divided into several sections: Vendor Details, Bill To, Ship To, and a table of items. The Vendor Details section includes the name 'Lacté', address '60 Route de Luzinay, 39540 Paris', email 'lacte.incredible.bureau@gmail.com', and VAT ID 'FR5684583'. The Bill To section lists 'Dairy Products Ltd' at '12-13 Waterloo Rd, London NW2 7UH, UK'. The Ship To section lists 'Dairy Products Ltd' at 'Harbour House 12, Marine Parade, Dover CT17 9BU'. The items table contains three entries: 'Crème liquide semi épaisse légère 4% Mat.Gr.' (90 ltr, 19.20/unit, 1728.00 total), 'Beurre Demi-sel Moulé de Bretagne' (100 pcs, 5.26/unit, 526.00 total), and 'Beurre Doux Tendre' (70 pcs, 4.80/unit, 336.00 total). The invoice date is 10/1/2021 and the due date is 10/6/2021.

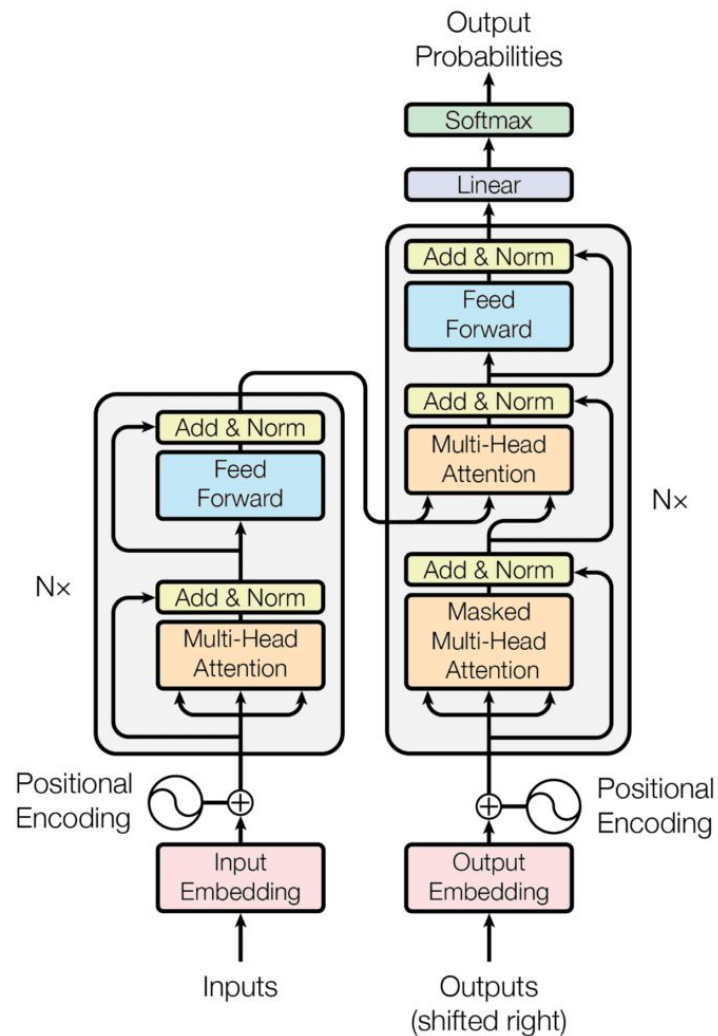
ITEM	DESCRIPTION	QTY	UNIT PRICE	TOTAL
001	SKU-592372 Crème liquide semi épaisse légère 4% Mat.Gr. Country of Origin: France	90 ltr	19.20	1 728.00
002	SKU-989834 Beurre Demi-sel Moulé de Bretagne Allergens: Lactose Country of Origin: France	100 pcs	5.26	526.00
003	SKU-649293 Beurre Doux Tendre Country of Origin: France	70 pcs	4.80	336.00

Step 2: Embedding

Turning Layout and Text information into vectors encoding semantic meaning.

Variant of LLM inspired by models like [hf:BERT/RoBERTa](#), [hf:LayoutXLM](#), [hf:UDOP](#).

- Encoder: “embedding quality control”
 - MLM pre-training *in distribution* documents.
 - Use *layout* (bounding boxes) during training.
- Decoder: “enhancing embeddings”
 - Its training is optional.
 - MVM - filling the “blanks” in the covered parts of the document.



Step 2: LLM Pre-training via M(V)LM

INVOICE



VENDOR DETAILS

Lacté
60 Route de Luzinay
38540 Paris
lacte.incredible.bureau@gmail.com
VAT ID - FR5684583

BILL TO

Dairy Products Ltd
12-13 Waterloo Rd
London NW2 7UH, UK
+44 7911 123456

DATE

10/1/2021

INVOICE NO.

435322

Due date

10/3/2021

SHIP TO

Dairy Products Ltd
Harbour House 12
Marine Parade, Dover CT17
9BU
+44 7911 123471

DATASET

```
(“INVOICE”, [476, 295, 618, 326]),  
 (“VENDOR”, [475, 420, 533, 433]),  
 (“DETAILS”, [534, 418, 591, 434]),  
 (“DATE”, [1134, 456, 1175, 471]),  
 ...
```



Step 2: LLM Pre-training via M(V)LM



[MASKED]



VENDOR DETAILS

Lacté
60 Route de Luzinay
38540 Paris
lacte.incredible.bureau@gmail.com
VAT ID - FR5684583

BILL TO

Dairy Products Ltd
12-13 Waterloo Rd
London NW2 7UH, UK
+44 7911 123456

Due date



10/1/2021

INVOICE NO.

435322

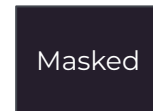
10/3/2021

SHIP TO

Dairy Products Ltd
Harbour House 12
Marine Parade, Dover CT17
9BU
+44 7911 123471

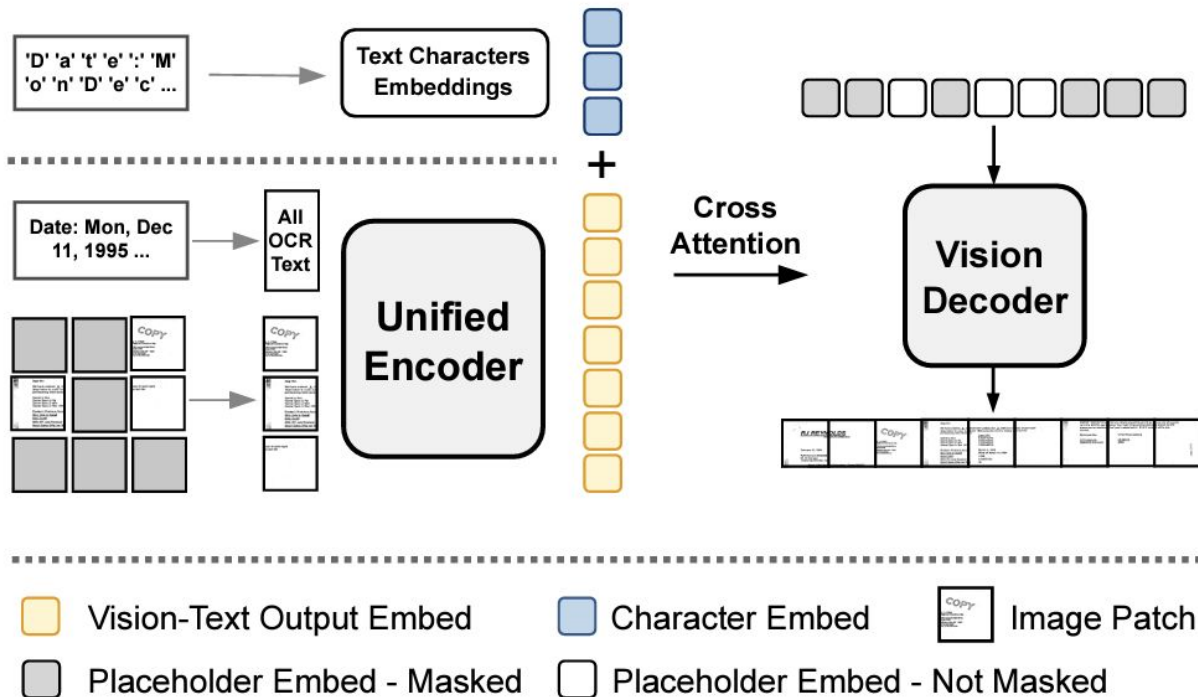
DATASET - TRAINING BATCH

```
([MASKED], [476, 295, 618, 326]),  
("VENDOR", [475, 420, 533, 433]),  
("DETAILS", [???, ???, ???, ???]),  
("DATE", [1134, 456, 1175, 471]),  
...
```



TRAINING OBJECTIVE: "Fill in the blanks"

Step 2: LLM Pre-training via M(V)LM



Step 3: Classification

LLM turns (localised) text into vectors encoding semantic meaning.

Here we can lean on vast collection of classic ML models: SVM, LogisticRegression, etc.

- One of the go-to libraries in Python: [pypi:scikit-learn](https://pypi.org/project/scikit-learn/)

Dataset: human-annotated documents - e.g. Rossum UI ([Slide 4](#))

- Data points are (field type, location, text).
- These are turned using LLMs into “classical” dataset of 2-tuples:
 - (embedding vector, class/number representing the field type)

Pre-training and Classification can be trained jointly → no need for separate model... has also other advantages.

Training: Self-supervised vs Supervised


Foundation Transformer


120 ltr 1.80 216.00

Ship Date: Jan 10. 2021

SUBTOTAL → 585.2 €
VAT (20%) → 585.2 €
TOTAL → 3511.2 €

Self-supervised training
Train to predict missing data on millions of pages of (unlabelled) documents.

 detected & extracted text




 *Embedding* = Transformer internal representation of the detected text.

Task-specific model

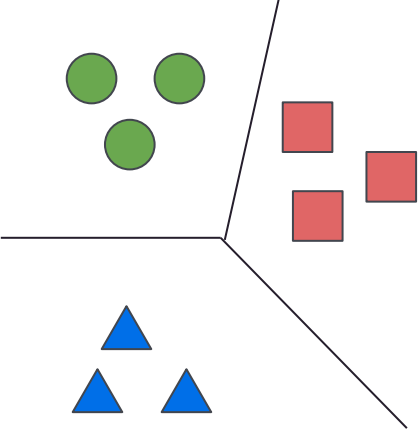
120 ltr 1.80 216.00

Ship Date: Jan 10. 2021

SUBTOTAL → 2 926.00 €
VAT (20%) → 585.2 €
TOTAL → 3511.2 €

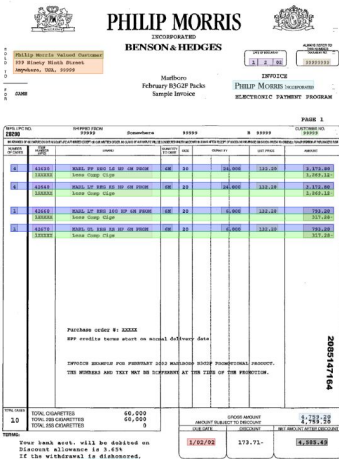
 → amount_subtotal
 → amount_tax
 → amount_total

Supervised training
Training across many datasets of annotated documents: Rossum datasets, customers annotated documents.



Reflection: Discriminative vs Generative AI

Reading an invoice is discriminative task:



PHILIP MORRIS INCORPORATED
BENSON & HEDGES
197 Kinston Road, Raleigh, NC 27601
February B&HGF Packs Sample Invoice

LINE	DESCRIPTION	QUANTITY	UNIT	PRICE	AMOUNT	TAX	TOTAL
10	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
11	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
12	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
13	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
14	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
15	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
16	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
17	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
18	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
19	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
20	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
21	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
22	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
23	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
24	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
25	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
26	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
27	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
28	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
29	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
30	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
31	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
32	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
33	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
34	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
35	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
36	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
37	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
38	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
39	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
40	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
41	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
42	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
43	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
44	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
45	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
46	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
47	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
48	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
49	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
50	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00

TOTAL DEDUCTIBLE: \$6,500
TOTAL BALANCE DUE: \$6,500

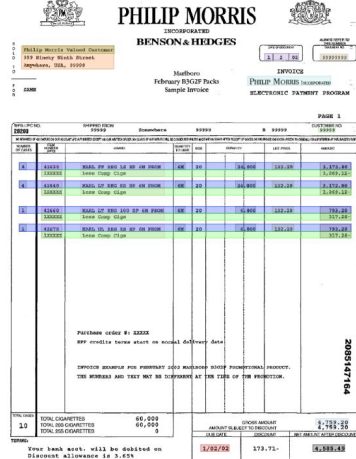
AMOUNT DUE TO PAYEE: \$6,500
TOTAL BALANCE DUE: \$6,500

2025/07/14

“Specialist”

- Fixed to a particular task (e.g. set of fields)
- Efficient & effective
- Less prone to simple mistakes

Reading an invoice as generative task:



PHILIP MORRIS INCORPORATED
BENSON & HEDGES
197 Kinston Road, Raleigh, NC 27601
February B&HGF Packs Sample Invoice

LINE	DESCRIPTION	QUANTITY	UNIT	PRICE	AMOUNT	TAX	TOTAL
10	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
11	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
12	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
13	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
14	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
15	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
16	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
17	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
18	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
19	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
20	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
21	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
22	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
23	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
24	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
25	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
26	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
27	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
28	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
29	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
30	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
31	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
32	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
33	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
34	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
35	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
36	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
37	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
38	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
39	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
40	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
41	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
42	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
43	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
44	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
45	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
46	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
47	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
48	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
49	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00
50	PHILIP MORRIS B&HGF PACKS	100	P	1.99	199.00	0.00	199.00

TOTAL DEDUCTIBLE: \$6,500
TOTAL BALANCE DUE: \$6,500

AMOUNT DUE TO PAYEE: \$6,500
TOTAL BALANCE DUE: \$6,500

2025/07/14

```
{  
  "Invid": "99999999",  
  "Due": "$4,505.49",  
  "Rcpt": "Philip Morris",  
}
```

- > Who is the recipient?
> Philip Morris Valued Customer
- > How much is due?
> Balance due is \$4,505.49

“Generalist”

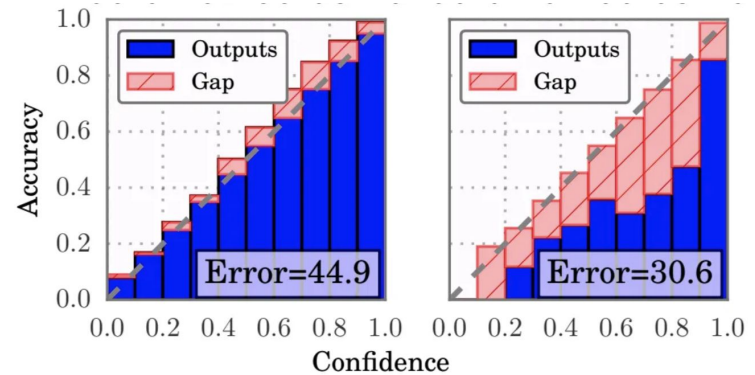
- Creative (but can invent nonsense)
- Versatile, handles the unexpected
- Deeper understanding of reality

Ups... That's not All for the End-to-End Automation.

For practical purposes of any AI – the model needs to be able to express the level of confidence in its predictions.

One way to tackle this is to make the final model predictions calibrated.

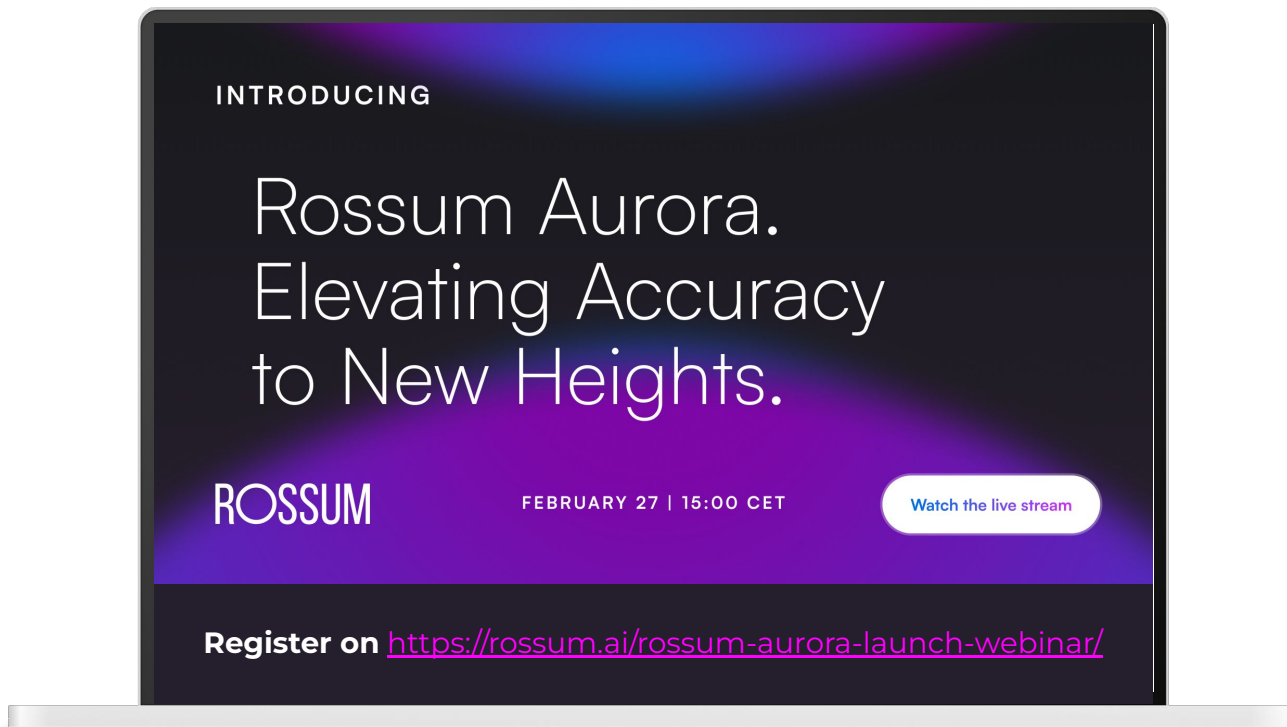
Luckily, this can be reduced to a classical AI problem (see [Probability calibration @scikit-learn](#))



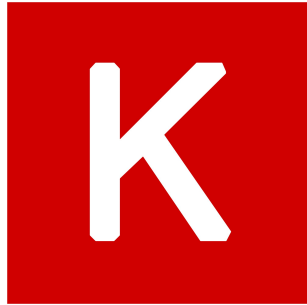
What are the Results / Impact of Rossum LLMs?



Coming Soon...



Software & Hardware



Software & Hardware

2017 - We Launch Rossum

“7GForce” - Our GPU “cluster”:

- 4x GeForce GTX 1080Ti
- 3x GeForce GTX 1070
- 40 CPUs / 128 GB RAM

All technology up to the “LLM era” Rossum could have been built using this.

- Prioritisation of experiments would be the challenge.
- True to be told, we’ve been also using AWS a lot.



Software & Hardware

2018 - We grow and so does our compute.

Adding 4 “Ry-cka” to our “cluster”:

- 12x GeForce GTX 1070
- 4x GeForce GTX 1080Ti
- 16 CPUs / 64 GB (each)

= 23 GPUs in total.

Has been enough for us until “recently” → LLMs started.



Software & Hardware

2022 - 2023 - Latest upgrades to our cluster for the “LLM era”:

Adding “Helena”:

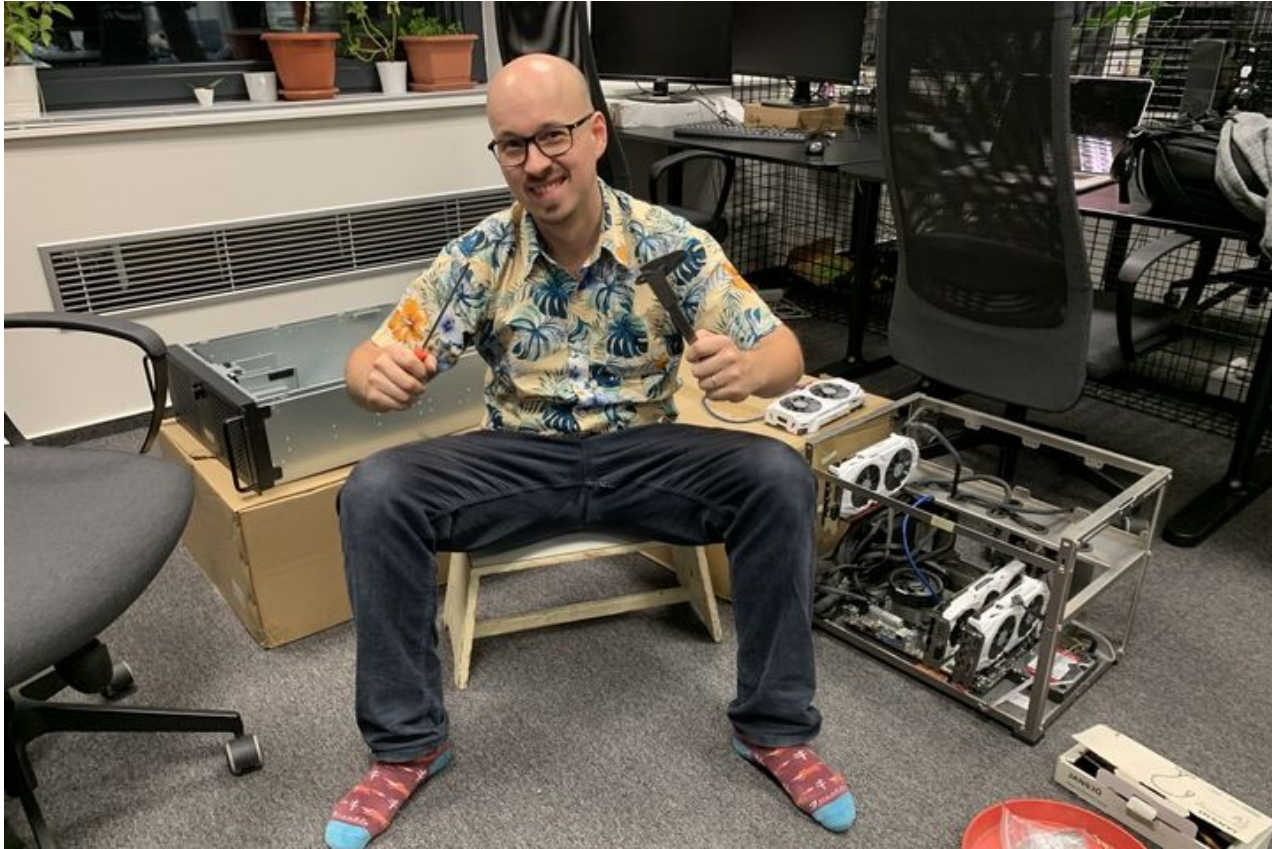
- 8x A100 GPUs
- 128 CPUs / 2T RAM

Adding “Sulla”:

- 8x RTX 6000 Ada
- 128 CPUs / 2T RAM



Software & **Hardware** - when things go wrong :/



Takeaways

ML Pipeline for Key Information Localisation and Extraction

- 4 Stages: OCR → LLM → Classification → Calibration
- Can be built on top of open-source software and models:
 - Data can be a challenge for fine-tuning, not for pre-training of LLMs.
 - Compute can be a challenge:
 - But you can go far with “relatively” little.
 - Modern fine-tuning techniques exist (leverage the open-source!)
- How LLMs are trained and how it works.
- What are pros and cons of Discriminative vs Generative approaches.

What is possible when this is put in practice?

- Come to see us at our [Rossum Aurora Product Launch](#)

Thank you