E2DIEVS: End-to-End Document Information Extraction and Verification System

Final Project Report

Yuvraj Bhadoriya
202411002

Priyanshi Patel
202411009

Manish Prajapati
202411012

1 Abstract

As we witnessed rapid digitalization in our country, India, there grows the need for automatic document verification systems. Earlier, the KYC (know your customer) methods were done manually, as opposed to today, we can now perform KYC anywhere on Earth. Traditionally, OCR-based pipelines are getting used extract text, and after applying several rule-based methods, we extract information from it. These methods failed miserably on Indian Identity documents such as Aadhar cards, and Pan cards. LayoutLM is a pre-trained model, known for extracting textual information from scanned documents. We employ that model, fine-tune it on the datasets which only consists of scanned form images and receipts with primary information, that eventually enhanced the performance. But, it faces limitations on scanning and understanding textual information from the Aadhar cards and Pan cards due to unrecognized formats. We overcome this limitation by using YoloV11, pre-trained on the dataset of aadhar cards and pan cards, which gives us text along with its category.

2 Introduction

Extracting information from scanned documents such as forms, invoices, receipts and identity cards has been emerged critically in AI. Microsoft Research introduced LayoutLM (Xu et al., 2019) which brought revolution in this domain. It incorporates spatial awareness in to the transformer which led to the state-of-the-art results on benchmarks such as FUNSD (Jaume et al., 2019) and SROIE (Huang et al., 2019). But it has a drawback of heavily relying on the accurate OCR and bounding box inputs during inference. That makes its multi-modal pre-training computationally expensive and less effective.

To address limitations faced by the LayoutLM, Donut (Kim et al., 2022) proposed an OCR-free approach. It uses a vision-transformer encoder which is combined with a text decoder to directly generate structured JSON output from raw document images without intermediate text recognition. The drawbacks of Donut are: training on diverse document types, and underperforms on densely formatted or mutli-lingual documents, such as Aadhar cards, and Pan cards.

The accurate field extraction from Aadhar cards and Pan cards remains particularly challenging due to multi-lingual content and inconsistent layouts across states. We came across an advanced object detection model, such as YoloV11 (Logasanjeev, 2024), which is fine-tuned on Indian ID datasets. It is more capable than LayoutLM and Donut for extracting the information from Indian IDs along with their categories. Our project contains the following novel ideas:

- LayoutLM fine-tuned on FUNSD and SROIE datasets showing improved performance.
- A novel pipeline that can extract information from the printed digital forms as well as Aadhar cards or Pan cards, and verify with the ground truth data.

3 Baseline Models

Our baseline model is *layoutlm-base-uncased*. LayoutLM (Xu et al., 2019) is a pre-trained multi-modal transformer which learns textual as well as spatial layout information from images of documents. It is pre-trained on IIT-CDIP dataset (Lewis et al., 2006), which consists of the following different documents:

- Letter
- Memo
- Email
- File-folder
- Form
- Handwritten Notes on page
- Invoices
- Advertisement
- Budget
- And many more.

LayoutLM was proposed by Microsoft Research Asia. A snippet of dataset it was pre-trained on can be seen in Figure 1. There are 400,000 grayscale images in 16 classes, each class accounting 25,000 images. After train, test and validation split, there are 320,000 images in the training set, 40,000 images in the validation set and 40,000 images in the test set. The images are resized such that the largest dimension do not exceed 1000 pixels.

Table 1: Model Specifications of LayoutLM.

Component	Details
Architecture	BERT-like Transformer
Input Modalities	1. Text (words) 2. 2D Layout
Pre-training Corpus	IIT-CDIP
Tokenizer	WordPiece (uncased)
Max Sequence Length	512
Parameters	∼110M

The model specification of the pre-trained LayoutLM can be seen in Table 1. The embeddings of the input modalities consists of the following information:

1. Adding 2D spatial embeddings to each token:
   - x-position-embeddings
   - y-position-embeddings
   - h-position-embeddings (height)
   - w-position-embeddings (width)
2. Pre-training on real scanned documents (not synthetic text), preserving:
   - Font styles
   - Table structures
   - Form fields
   - Spatial alignments

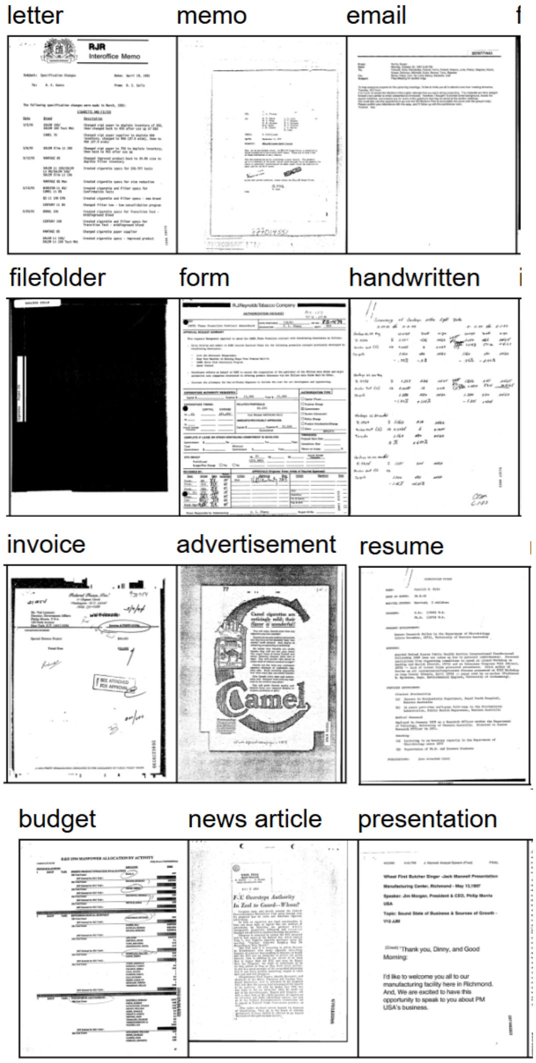


Figure 1: IIT-CDIP train dataset snippet. There 16 different types of scanned documents in the dataset.

4 Proposed Model

4.1 Dataset Configuration

1. FUNSD (Jaume et al., 2019): A dataset for understanding the forms in noisy scanned documents. There are 149 datapoints in train-set and 50 datapoints in test-set. The language in forms is majorly English. More details can be seen in Table 2.

2. SROIE (Huang et al., 2019): The full-form is Scanned Receipts OCR and Information Extraction. This is used for OCR and key information extraction from scanned receipts. It consists of 626 images in train-set and 347 images in test-set. The language in receipts is English. More details can be seen in Table 3.
[1]

[1] Github Repo: https://github.com/ManishPraa24/End-

Table 2: FUNSD dataset configuration.

Attribute	**Details**
Total Images	199 scanned forms
Word-level annotations	30,000
Entities	10,000
Entity Types	Header, Question, Answer & other
Annotation Format	BIO tagging + bounding boxes + key-value links
Image Resolution	100 DPI grayscale

Table 3: SROIE dataset configuration.

Attribute	**Details**
Total Images	1,000 scanned receipts
Key Entities	4
Image resolution	Varied
Annotations	OCR bounding boxes and key-value pairs

4.2 Model Configuration

LayoutLM introduced a powerful pre-training paradigm for scanned documents. However, downstream evaluation was constrained. Earlier work fine-tuned the model on FUNSD and test-set of the SROIE dataset corpora. We find our enhancement far from being trivial for the following reasons:

1. The objectives of pre-training operate at the representation level and do not expose the model to task-specific BIO schemas.

2. Pre-training data consist primarily of scanned books and articles from IIT-CDIP, exhibiting fundamentally different spatial statistics.

3. Character-level fidelity as a Proxy for deployment success.

Thus, our work including the fine-tuning, completes the experiment from the original paper. The model configuration can be seen in the Table 4.

As described in Table 4, we choose 15 number of labels, which represent the BIO-schema heads: HEADER, QUESTION, ANSWER, COMPANY, DATE, ADDRESS, TOTAL, etc. We kept the value of the maximum sequence length same as pre-training. To be memory efficient with the bounding box embeddings, we kept the batch-size as 8. Optimizer is AdamW with weight decay of 0.01. Initially, we kept model warmup comprising of 10% of the steps in order to stabilize the training. For collating the data, we use dynamic padding that gives no batch size errors, and ensures no data loss.

Table 4: Model configuration for fine-tuning on the dataset corpora of FUNSD and SRIOE.

Parameter	**Value**
Number of labels	15 (BIO-schema)
Max. sequence length	512
Batch Size	8
Learning rate	5e-5
Optimizer	AdamW
Epochs	30

4.3 Embeddings and Encoding

We have used the original pipeline of LayoutLM pre-training which contains the following strategies:

1. **Text Encoding**:
   - Input: List of words (from OCR: Tesseract for SROIE, ground-truth for FUNSD).
   - Tokenizer: *AutoTokenizer* (WordPiece, uncased) with *is_split_into_words = True*.
   - Sub-words of the first token for each word inherits the BIO label.

2. **Layout (2D Position) Encoding**
   - Bounding boxes are normalized for every document, making them the size of [0, 1000] x [0, 1000].
   - Formula:

     $$x' = \left\lfloor \frac{x \cdot 1000}{\max(x)} \right\rfloor, \quad y' = \left\lfloor \frac{y \cdot 1000}{\max(y)} \right\rfloor$$

     where x and y are the dimensions of the images.
   - Four coordinates are fed into the learned 2D position embeddings as in (Xu et al., 2019).

to-End-Document-Information-Extraction-and-Verification-System

3. **Input representation**

   - Each token receives:
     [CLS] + Token Embedding + Segment Embedding + 2-D Position Embedding (x1, y1, x2, y2) + *[SEP]*.

4.4 Evaluation Method

4.4.1 Training and Evaluating Loss

During the fine-tuning, we optimize the standard token classification objective by incorporating cross-entropy loss over the predicted logits and ground-truth labels. For a sequence of length T, the loss for a particular instance is calculated as:

$$\mathcal{L} = -\frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \log \left(\frac{\exp\Big(logit_{y_t}[t]\Big)}{\sum_{c=1}^{C} \exp(logit_c[t])} \right), (1)$$

where:

- $\mathcal{A} = t \mid y_t \neq -100$ is set the of active positions.
- y_t ∈ {0, 1, ..., C-1} is the ground-truth label at position t.
- $C = 15$ is the number of classes in our unified BIO schema.
- $logit_c[t]$ is the model's un-normalized score for class c at position t.
- The positions corresponding to the subword continuations, and special tokens ([CLS], [SEP], [PAD]) are assigned label **-100**. This is ignored by the loss function.

Our objective function is similar to that in (Xu et al., 2020).

4.4.2 Token-level Accuracy

For document-image understanding tasks, the token-level accuracy is defined as:

$$Accuracy = \frac{\sum_{t \in \mathcal{A}} 1\{\hat{y}_t = y_t\}}{|\mathcal{A}|} \qquad (2)$$

4.4.3 Macro F1 Score

This is computed for more than 15 BIO classes:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}, Macro - F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c \qquad (3)$$

where P_c and R_c are the precision and recall for class c.

5 Results

The results for baseline LayoutLM on the dataset of FUNSD can be seen in Table 5, and on the dataset corpora of SROIE can be seen in Table 6.

Table 5: Evaluation of FUNSD dataset on baseline LayoutLM model

Metric	Value
Recall	0.041
Precision	0.254
F1	0.068

Table 6: Evaluation of SROIE dataset on baseline LayoutLM model

Metric	Value
Recall	0.271
Precision	0.249
F1	0.249

The macro-F1 score on FUNSD dataset is 0.068 and, that on SROIE dataset is 0.249. Both the scores are very low, compared to that we obtained upon fine-tuning the model on the FUNSD and SROIE datasets. Our fine-tuned model achieved the best macro-F1 score of 0.638 at epoch 19, and best evaluation accuracy of 95.169% at epoch 7. The trajectory of macro-F1 and evaluation accuracy can be seen in Figure 2. We obtained the least validation loss at epoch 5 having value of 0.1544. The trajectory of evaluation loss and training loss during fine-tuning can be seen in the Figure 3.

6 Inference

We parsed three random images, taken from FUNSD, SROIE and Aadhar card dataset, in order to observe the change in inference of the model after fine-tuning.

FUNSD: The inference of pre-trained LayoutLM model on the one of the instance of FUNSD dataset can be seen in Figure 4. Comparing to the inference of the fine-tuned LayoutLM model which can be seen in Figure 4, the pre-trained model fails to classify the text such as 'TO', 'FAX NUMBER', etc. as keys. However, the fine-tuned model classifies the parsed text such as 'TO', 'FAX NUMBER', etc., and 'George Baroody', '(336) 335-7392', etc., as corresponding key-value pairs.

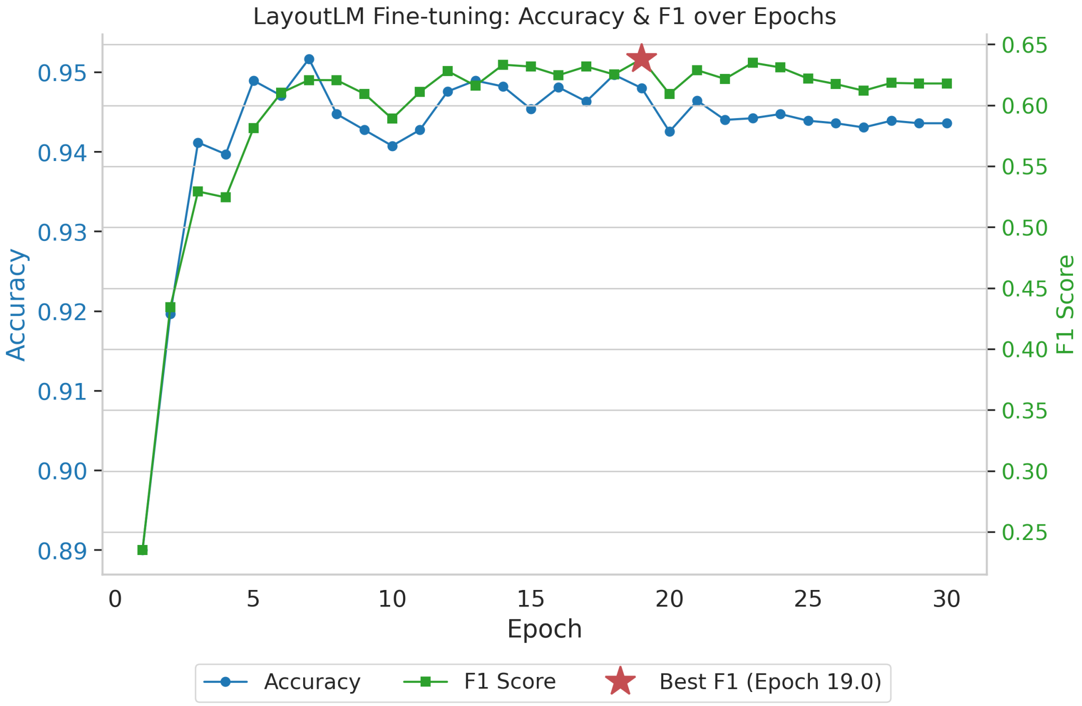


Figure 2: Trajectory of macro-F1 and evaluation accuracy during fine-tuning of the LayoutLM model.

SROIE: Pre-trained model's inference on the instance of SROIE dataset can be seen in Figure 6, and inference obtained using fine-tuned model can be seen in Figure 7. Pre-trained model classified text such as 'Pre auth code . . . ', 'V-Power 97', etc. as B-Company, which is wrong. Comparing to the inference of the fine-tuned model, there is no text classified to the label 'B-Company'.

Aadhar Card: The inference of the dummy Aadhar card obtained using pre-trained model can be seen in Figure 8, and that of fine-tuned model can be seen in Figure 9. Pre-train model classifies entities like 'Name', as B-HEADER. Compared to inference of the fine-tuned model, the 'NAME' is classified as B-QUESTION, enabling a question-answer pair, which is better than the pre-trained inference.

Overall, the fine-tuned model is able to recognize possible question-answer form of text from forms, receipts and aadhar-cards, better than pre-trained model.

7 Future Work

We observe that fine-tuned model still struggles to classify text from Aadhar card, such as 'Male', 'DOB: ¡date¿' to be different and the Aadhar number as a question-answer, which determines that there's still uncertainty existing in the model. We further plan to address these issues by fine-tuning the model on publicly available datasets with annotations. Our last stage will be to make pipeline, which scanned the document and extracts the important key entities based on the type of document, and verify it with the available ground-truth text.

References

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. *ICDAR*. Available at: https://rrc.cvc.uab.es/?ch=13.

Guillaume Jaume, Hazim K Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. *arXiv preprint arXiv:1905.13538*. Available at: https://guillaumejaume.github.io/FUNSD/.

Geewook Kim and 1 others. 2022. clovaai/donut: Official implementation of ocr-free document understanding transformer. https://github.com/clovaai/donut.

David D. Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Iit-cdip test collection 1.0. https://ir.nist.gov/cdip/. Illinois Institute of Technology Complex Document Information Processing (CDIP) test collection.

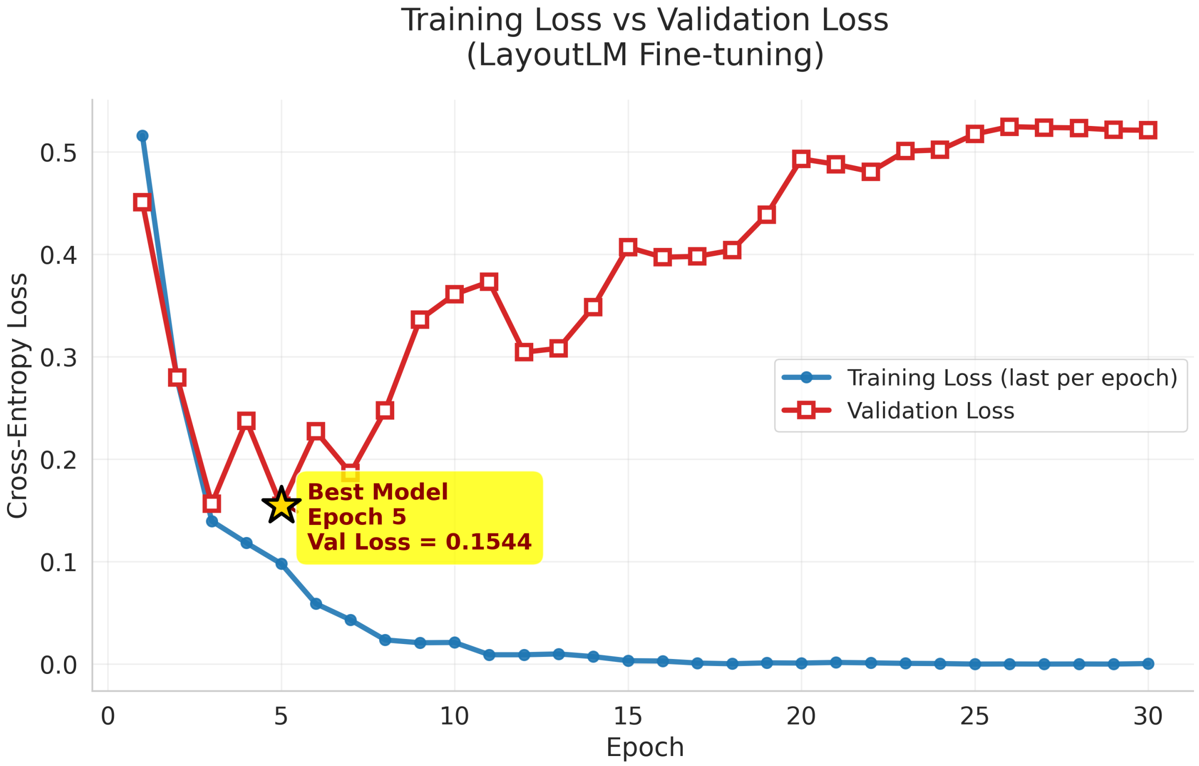
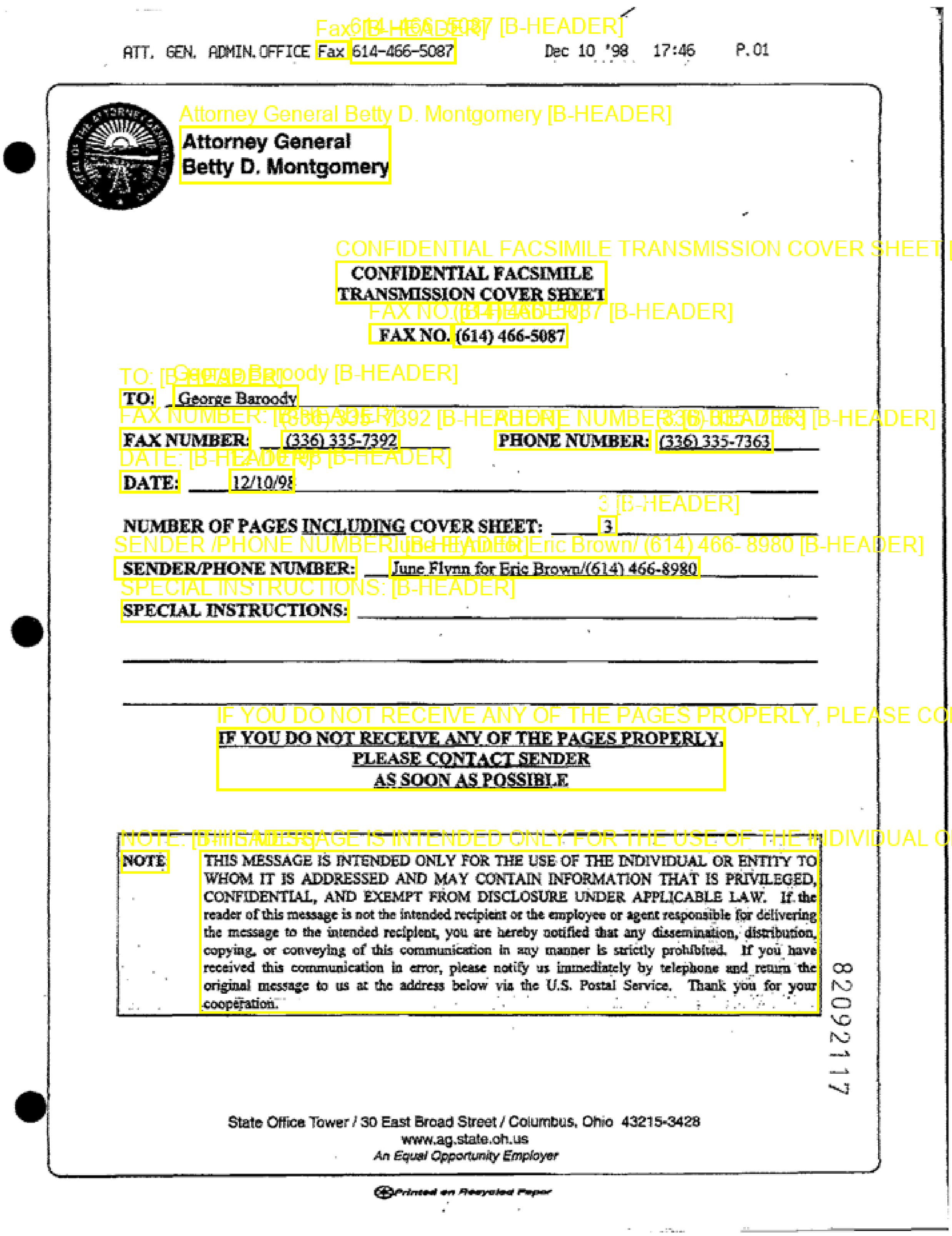


Figure 3: Trajectory of training loss and evaluation loss during the fine-tuning of the LayoutLM model.

Logasanjeev. 2024. indian-id-validator – yolov11 model for aadhaar, pan, and driving license field detection. Hugging Face Model Hub.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. Layoutlm: Pre-training of text and layout for document image understanding. *CoRR*, abs/1912.13318.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.


Fax [B-HEADER]
614-466-5087 [B-HEADER]
Attorney General Betty D. Montgomery [B-HEADER]
CONFIDENTIAL FACSIMILE TRANSMISSION COVER SHEET [B-
FAX NO. [B-HEADER]
(614) 466-5087 [B-HEADER]
TO: [B-HEADER]
George Baroody [B-HEADER]
FAX NUMBER: [B-HEADER]
(336) 335-7392 [B-HEADER]
PHONE NUMBER: [B-HEADER]
(336) 335-7363 [B-HEADER]
DATE: [B-HEADER]
12/10/98 [B-HEADER]
3 [B-HEADER]
SENDER /PHONE NUMBER: [B-HEADER]
June Flynn for Eric Brown/ (614) 466- 8980 [B-HEADER]
SPECIAL INSTRUCTIONS: [B-HEADER]
IF YOU DO NOT RECEIVE ANY OF THE PAGES PROPERLY, PLEASE CONT
NOTE: [B-HEADER]
THIS MESSAGE IS INTENDED ONLY FOR THE USE OF THE INDIVIDUAL OR E

ATT. GEN. ADMIN.OFFICE Fax 614-466-5087 Dec 10 '98 17:46 P.01

Attorney General
Betty D. Montgomery

CONFIDENTIAL FACSIMILE
TRANSMISSION COVER SHEET

FAX NO. (614) 466-5087

TO: George Baroody

FAX NUMBER: (336) 335-7392 PHONE NUMBER: (336) 335-7363

DATE: 12/10/98

NUMBER OF PAGES INCLUDING COVER SHEET: 3

SENDER/PHONE NUMBER: June Flynn for Eric Brown/(614) 466-8980

SPECIAL INSTRUCTIONS:

IF YOU DO NOT RECEIVE ANY OF THE PAGES PROPERLY,
PLEASE CONTACT SENDER
AS SOON AS POSSIBLE

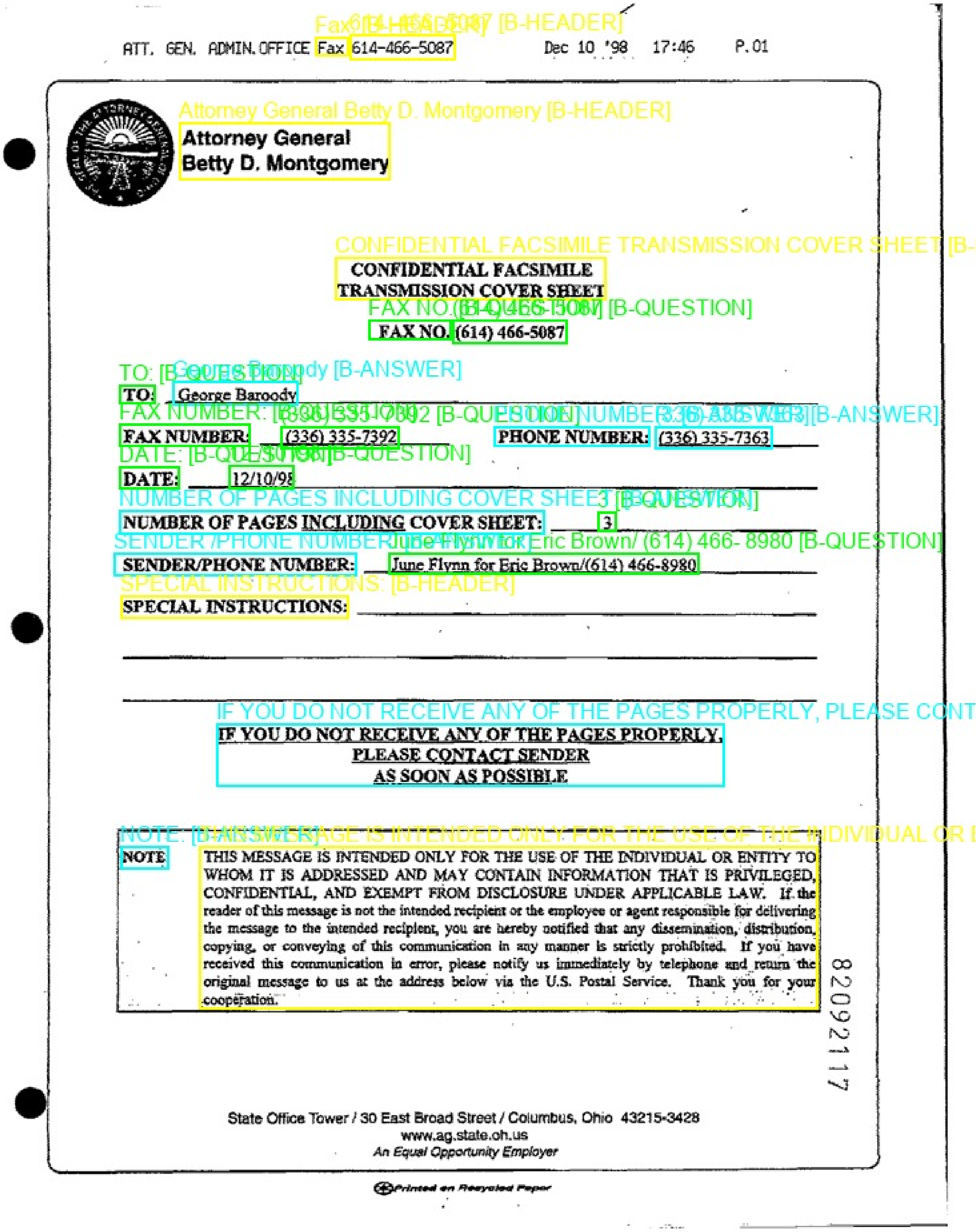
NOTE: THIS MESSAGE IS INTENDED ONLY FOR THE USE OF THE INDIVIDUAL OR ENTITY TO WHOM IT IS ADDRESSED AND MAY CONTAIN INFORMATION THAT IS PRIVILEGED, CONFIDENTIAL, AND EXEMPT FROM DISCLOSURE UNDER APPLICABLE LAW. If the reader of this message is not the intended recipient or the employee or agent responsible for delivering the message to the intended recipient, you are hereby notified that any dissemination, distribution, copying, or conveying of this communication in any manner is strictly prohibited. If you have received this communication in error, please notify us immediately by telephone and return the original message to us at the address below via the U.S. Postal Service. Thank you for your cooperation.

82092117

State Office Tower / 30 East Broad Street / Columbus, Ohio 43215-3428
www.ag.state.oh.us
An Equal Opportunity Employer

Printed on Recycled Paper

Figure 4: Pre-trained LayoutLM inference on one of the instance of the FUNSD dataset.



Fax [B-HEADER]
614-466-5087 [B-HEADER]
Attorney General Betty D. Montgomery [B-HEADER]
CONFIDENTIAL FACSIMILE TRANSMISSION COVER SHEET [B-
FAX NO. [B-QUESTION]
(614) 466-5087 [B-QUESTION]
TO: [B-QUESTION]
George Baroody [B-ANSWER]
FAX NUMBER: [B-QUESTION]
(336) 335-7392 [B-QUESTION]
PHONE NUMBER: [B-ANSWER]
(336) 335-7363 [B-ANSWER]
DATE: [B-QUESTION]
12/10/98 [B-QUESTION]
NUMBER OF PAGES INCLUDING COVER SHEET:
3 [B-QUESTION]
SENDER /PHONE NUMBER: [B-ANSWER]
June Flynn for Eric Brown/ (614) 466- 8980 [B-QUESTION]
SPECIAL INSTRUCTIONS: [B-HEADER]
IF YOU DO NOT RECEIVE ANY OF THE PAGES PROPERLY, PLEASE CONT
NOTE: [B-ANSWER]
THIS MESSAGE IS INTENDED ONLY FOR THE USE OF THE INDIVIDUAL OR I

ATT. GEN. ADMIN.OFFICE Fax 614-466-5087 Dec 10 '98 17:46 P.01

Attorney General
Betty D. Montgomery

CONFIDENTIAL FACSIMILE
TRANSMISSION COVER SHEET

FAX NO. (614) 466-5087

TO: George Baroody

FAX NUMBER: (336) 335-7392 PHONE NUMBER: (336) 335-7363

DATE: 12/10/98

NUMBER OF PAGES INCLUDING COVER SHEET: 3

SENDER/PHONE NUMBER: June Flynn for Eric Brown/(614) 466-8980

SPECIAL INSTRUCTIONS:

IF YOU DO NOT RECEIVE ANY OF THE PAGES PROPERLY,
PLEASE CONTACT SENDER
AS SOON AS POSSIBLE

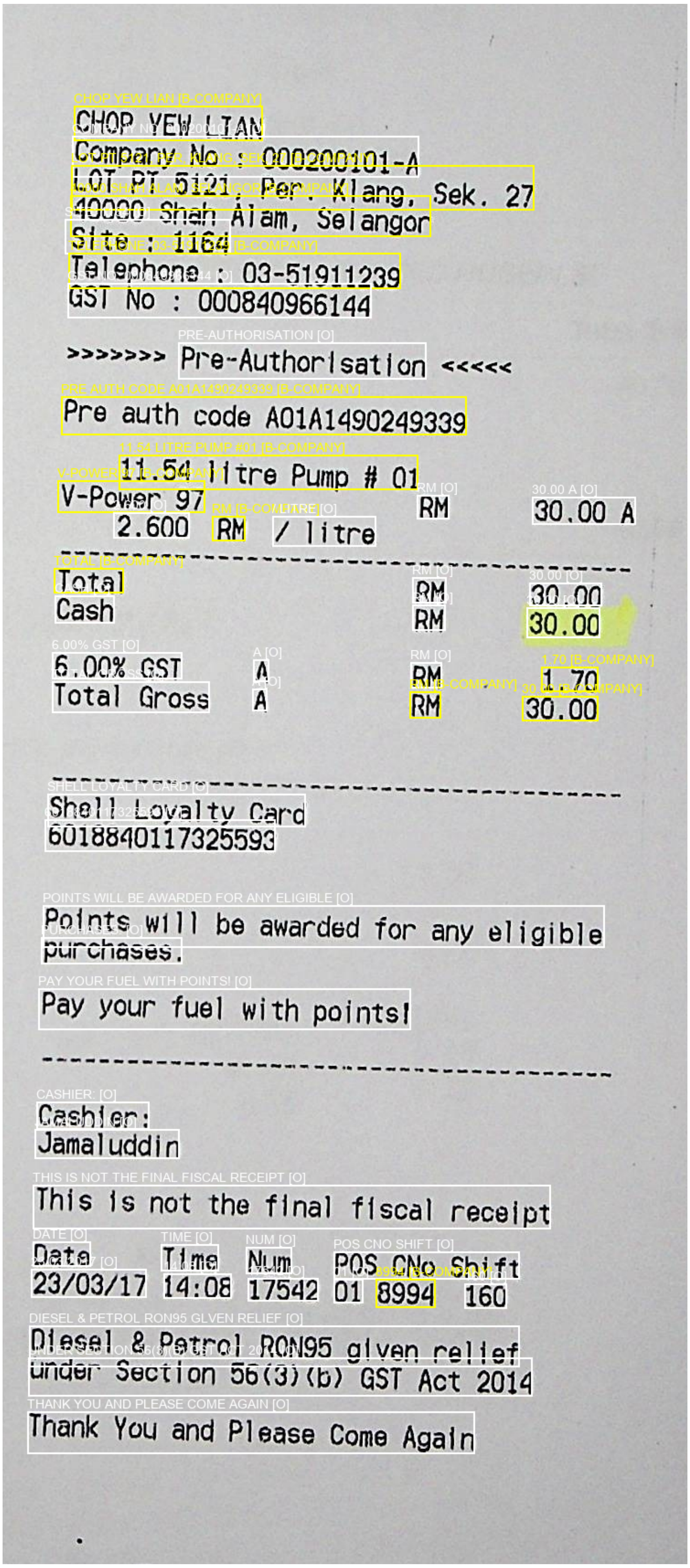
NOTE	THIS MESSAGE IS INTENDED ONLY FOR THE USE OF THE INDIVIDUAL OR ENTITY TO WHOM IT IS ADDRESSED AND MAY CONTAIN INFORMATION THAT IS PRIVILEGED, CONFIDENTIAL, AND EXEMPT FROM DISCLOSURE UNDER APPLICABLE LAW. If the reader of this message is not the intended recipient or the employee or agent responsible for delivering the message to the intended recipient, you are hereby notified that any dissemination, distribution, copying, or conveying of this communication in any manner is strictly prohibited. If you have received this communication in error, please notify us immediately by telephone and return the original message to us at the address below via the U.S. Postal Service. Thank you for your cooperation.

82092117

State Office Tower / 30 East Broad Street / Columbus, Ohio 43215-3428
www.ag.state.oh.us
An Equal Opportunity Employer

Printed on Recycled Paper

Figure 5: Inference of the fine-tuned LayoutLM model on one of the instance of the FUNSD dataset.



CHOP YEW LIAN
Company No : 000200101-A
LOT PT 5121, Per. Klang, Sek. 27
40000 Shah Alam, Selangor
Site : 1164
Telephone : 03-51911239
GST No : 000840966144

>>>>>>>> Pre-Authorisation <<<<<

Pre auth code A01A1490249339

11.54 litre Pump # 01
V-Power 97 RM 30.00 A
2.600 RM / litre

Total		RM	30.00
Cash		RM	30.00
6.00% GST	A	RM	1.70
Total Gross	A	RM	30.00

Shell Loyalty Card
6018840117325593

Points will be awarded for any eligible purchases.

Pay your fuel with points!

Cashier:
Jamaluddin

This is not the final fiscal receipt

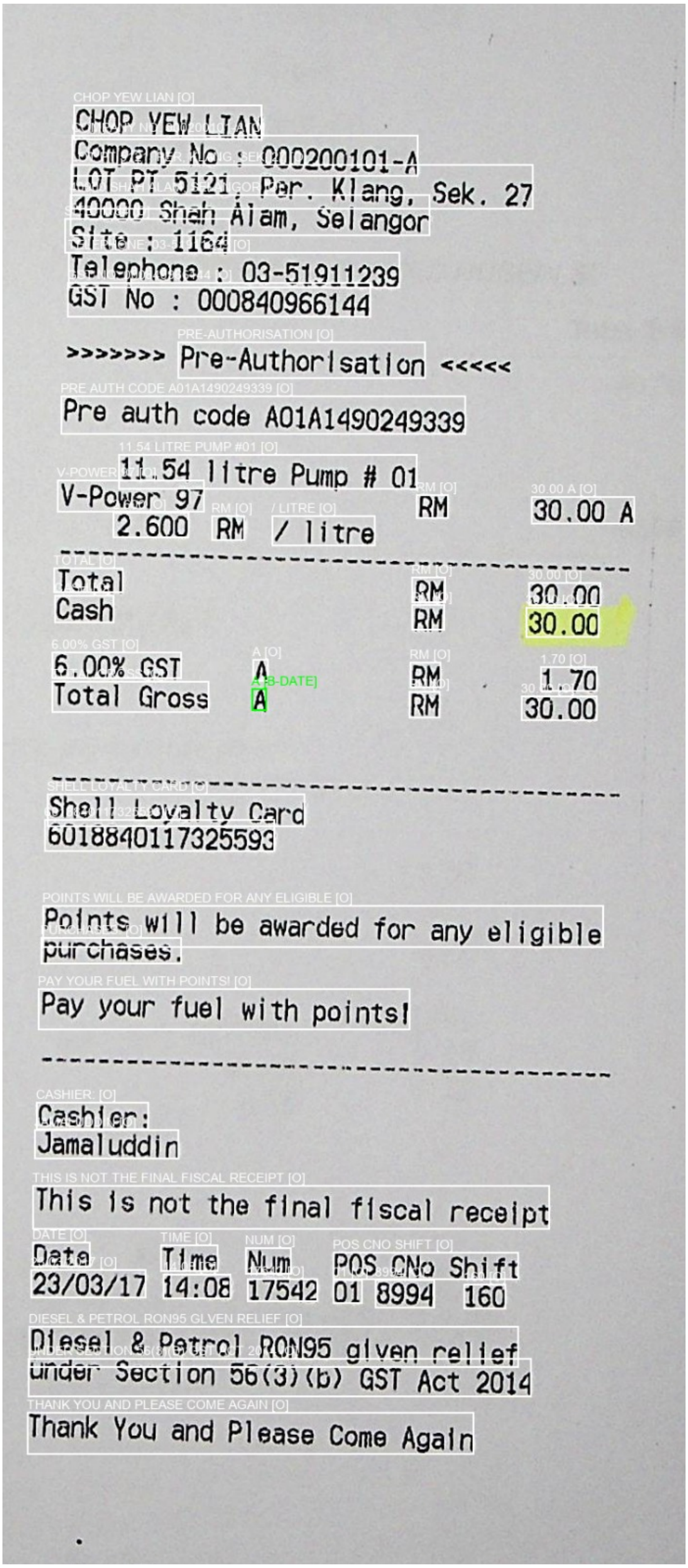
Date	Time	Num	POS	CNo	Shift
23/03/17	14:08	17542	01	8994	160

Diesel & Petrol RON95 given relief
under Section 56(3)(b) GST Act 2014

Thank You and Please Come Again

CHOP YEW LIAN [B-COMPANY]
TELEPHONE: 03-51911239 [B-COMPANY]
PRE-AUTHORISATION [O]
PRE AUTH CODE A01A1490249339 [B-COMPANY]
11.54 LITRE PUMP #01 [B-COMPANY]
V-POWER 97 [B-COMPANY]
RM [O]
30.00 A [O]
RM [B-COMPANY]
LITRE [O]
TOTAL [B-COMPANY]
RM [O]
30.00 [O]
6.00% GST [O]
A [O]
RM [O]
1.70 [B-COMPANY]
RM [B-COMPANY]
SHELL LOYALTY CARD [O]
POINTS WILL BE AWARDED FOR ANY ELIGIBLE [O]
PURCHASES. [O]
PAY YOUR FUEL WITH POINTS! [O]
CASHIER: [O]
THIS IS NOT THE FINAL FISCAL RECEIPT [O]
DATE [O]
TIME [O]
NUM [O]
POS CNO SHIFT [O]
8994 [B-COMPANY]
DIESEL & PETROL RON95 GLVEN RELIEF [O]
THANK YOU AND PLEASE COME AGAIN [O]

Figure 6: Pre-trained LayoutLM inference on one of the instance of the dataset corpora of SROIE.



CHOP YEW LIAN
Company No : 000200101-A
LOT PT 5121, Per. Klang, Sek. 27
40000 Shah Alam, Selangor
Site : 1164
Telephone : 03-51911239
GST No : 000840966144

>>>>>>> Pre-Authorisation <<<<<

Pre auth code A01A1490249339

11.54 litre Pump # 01
V-Power 97 RM 30.00 A
2.600 RM / litre

Total		RM	30.00
Cash		RM	30.00
6.00% GST	A	RM	1.70
Total Gross	A	RM	30.00

Shell Loyalty Card
6018840117325593

Points will be awarded for any eligible purchases.

Pay your fuel with points!

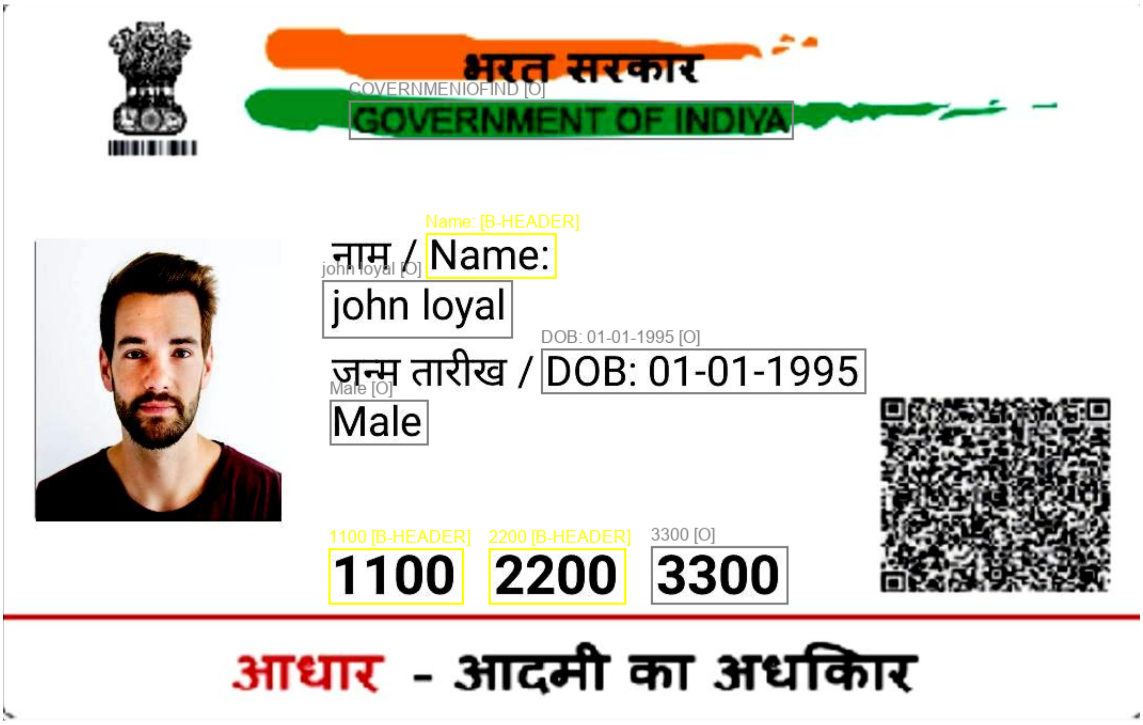
Cashier:
Jamaluddin

This is not the final fiscal receipt

Date	Time	Num	POS CNo Shift
23/03/17	14:08	17542	01 8994 160

Diesel & Petrol RON95 given relief
under Section 56(3)(b) GST Act 2014
Thank You and Please Come Again

Figure 7: Inference of the fine-tuned LayoutLM model on one of the instance of the SROIE dataset.



भरत सरकार

COVERNMENIOFIND [O]

GOVERNMENT OF INDIYA

Name: [B-HEADER]

नाम / Name:

john loyal [O]

john loyal

DOB: 01-01-1995 [O]

जन्म तारीख / DOB: 01-01-1995

Male [O]

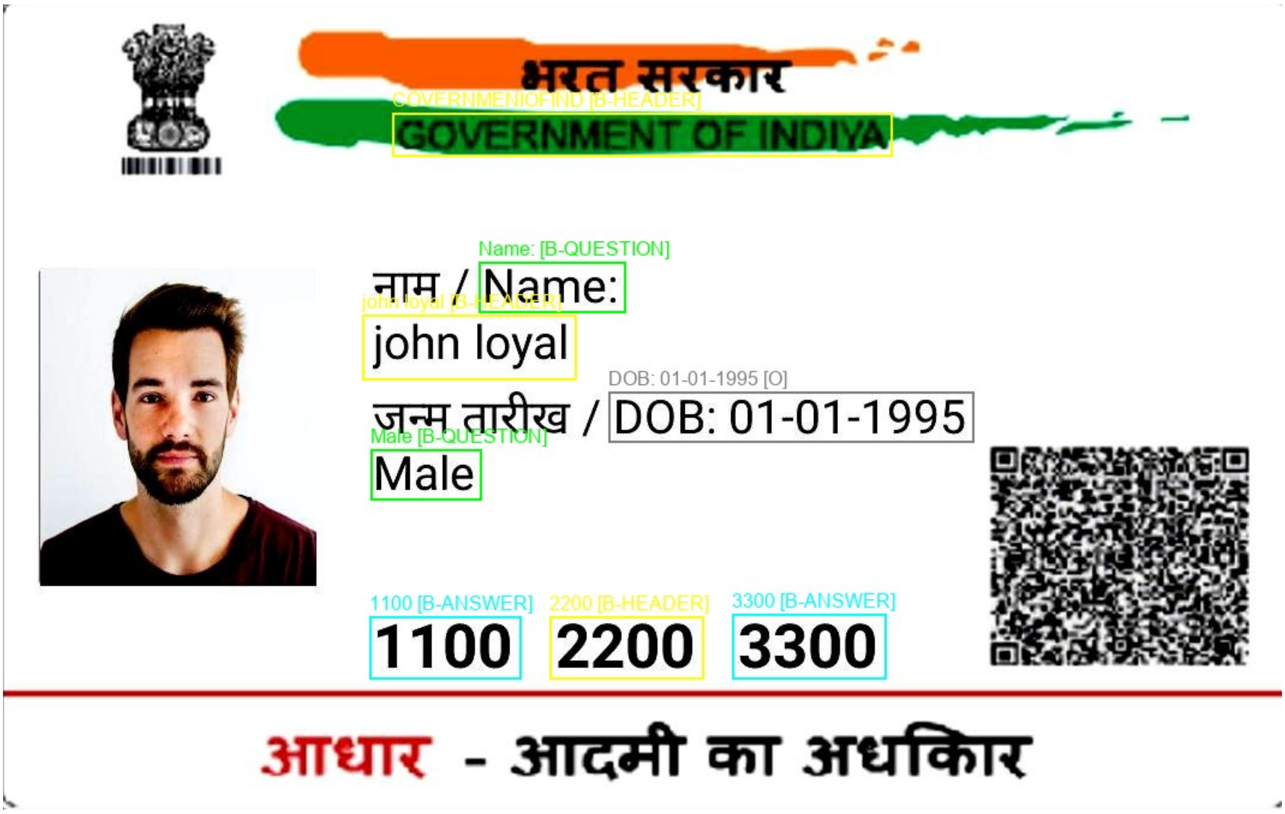
Male

1100 [B-HEADER] 2200 [B-HEADER] 3300 [O]

1100 2200 3300

आधार - आदमी का अधिकार

Figure 8: Pre-trained LayoutLM inference on the dummy Aadhar Card.



भरत सरकार

COVERNMENIOFIND [B-HEADER]

GOVERNMENT OF INDIYA

Name: [B-QUESTION]

नाम / Name:

john loyal [B-HEADER]

john loyal

DOB: 01-01-1995 [O]

जन्म तारीख / DOB: 01-01-1995

Male [B-QUESTION]

Male

1100 [B-ANSWER] 2200 [B-HEADER] 3300 [B-ANSWER]

1100 2200 3300

आधार - आदमी का अधिकार

Figure 9: Inference of the fine-tuned LayoutLM model on the dummy Aadhar Card.