

DocPolarBERT: A Pre-trained Model for Document Understanding with Relative Polar Coordinate Encoding of Layout Structures

Benno Uthayasooriyar^{1,2} Antoine Ly¹ Franck Vermet² Caio Corro³

¹Data Analytics Solutions, SCOR ²Univ Brest, CNRS, UMR 6205, LMBA

³INSA Rennes, IRISA, Inria, CNRS, Université de Rennes

Abstract

We propose a novel self-attention mechanism for document understanding that takes into account text block positions in relative polar coordinate system rather than the Cartesian one. Based on this mechanism, we build DOCPOLARBERT, a layout-aware BERT model for document understanding that eliminates the need for absolute 2D positional embeddings. Despite being pre-trained on a dataset more than six times smaller than the widely used IIT-CDIP corpus, DOCPOLARBERT achieves state-of-the-art results. These results demonstrate that a carefully designed attention mechanism can compensate for reduced pre-training data, offering an efficient and effective alternative for document understanding.

1 Introduction

The standard setting in natural language processing is to consider text that appears as sequential data. In contrast, document understanding aims to analyze data that is presented in two dimensional documents whose layout structure convey crucial information, such as forms and invoices. To this end, one can extend large language models (LLM) to incorporate layout information (Wang et al., 2024a; Tanaka et al., 2024; Luo et al., 2024). However, LLMs pose challenges for real-world deployment due to their high computational costs, often causing bottlenecks when processing large volumes of data in real-time applications. Furthermore, API-based models are often a *no go* when data privacy is a critical concern.

As such, developing light and efficient models is a critical line of work. Similar to BERT (Devlin et al., 2019), prior work proposed to pre-train self-attentive networks (i.e. transformer encoders Vaswani et al., 2017) on large amount of unlabeled documents (Xu et al., 2019; Huang et al., 2022; Tu et al., 2023; Jiang et al., 2025). These models

can then be fine-tuned for downstream tasks like named-entity recognition.

A key characteristic of any document BERT model is how they encode layout structures. Previous works rely on absolute position encoding, often with the addition of a relative position information encoded as bias in the attention mechanism (Xu et al., 2021), in order to improve spatial interaction modeling. We hypothesize that such encoding of layout information can lead to generalization issue. First, absolute positional embeddings are not invariant to translation, which is a key characteristic of many documents of interest: the representation of elements in a table should not be impacted by the position of the table in the document. Second, relative positions between elements always rely on the Cartesian coordinate system. This obfuscate key information (“is above of”, “is at a given distance of”) and can hinder generalization due to data sparsity issues.

As such, we argue that the polar coordinate system can better leverage layout information. Indeed, when dealing with relative positions in forms and tables, the polar coordinate system that distinguishes between angle and distance is a more semantically appropriate choice, capturing:

- Angular information: column headers are often in first row, and this information must prevail for all other rows;
- Distance information: related fields in a form are close (e.g. first name, last name, birth date), but their order may differ.

Therefore, in this work we propose a novel BERT-like encoder for documents that completely removes any absolute positional information, and relies only on relative encoding of layout structures in the polar coordinate system.

Our contributions can be summarized as follows: (1) we propose a novel attention mechanism that in-

corporates relative positional encoding in the polar coordinate system; (2) we pre-train such a model on publicly available data; (3) we evaluate our model by fine-tuning on several downstream tasks, and show it achieves competitive or better results than comparable baselines; and (4) we also conduct a comprehensive set of experiments and ablation studies to validate our design choices and assess the robustness of the proposed approach, covering generalization across complex layouts, discretization strategies of distances, fair comparison with a vision-free LAYOUTLMV3 variant, attention pattern analysis, and computational efficiency.

Code, data and models are publicly available.¹

2 Related Work

Document understanding. Xu et al. (2019) introduced absolute positional embeddings that encode bounding boxes of each text unit (word or group of words) in the input of a BERT model. Following in this direction, researchers have developed multi-modal encoders that include extra vision features, often leading to specific training strategies (Appalaraju et al., 2021; Gu et al., 2021; Xu et al., 2021; Huang et al., 2022; Bai et al., 2023). Nevertheless, as vision models are computationally intensive, there is an interest in alternatives without vision features, to achieve fast inference while maintaining competitive performance (Hong et al., 2022; Wang et al., 2022; Tu et al., 2023; Jiang et al., 2025).

Relative attention. Shaw et al. (2018) introduce *relation-aware self-attention*, which extends the standard attention mechanism with biases based on relative positions of tokens. This design embeds positional relationships directly into the attention process, rather than depending only on absolute position embeddings. Recent research on document understanding (Powalski et al., 2021; Xu et al., 2021; Hong et al., 2022; Huang et al., 2022) follows similar strategies by adapting self-attention to handle two-dimensional spatial relationships. In these models, the attention bias between two text bounding boxes is computed according to their relative spatial arrangement: (1) pairwise relative distances are computed following a specified geometric definition, i.e. distance between top-left corners of the boxes, (2) these continuous distances are discretized into fixed buckets and (3) each discrete

value is mapped to a learned embedding used as a bias in the attention computation. Closer to our work, Hwang et al. (2021) also used relative coordinates in the polar coordinate system.

3 Neural Architecture

In this section, we describe our novel encoding method of layout structures using relative positional information in the polar coordinate system.

3.1 Background: Relative Attention

Without loss of generality, in the following we assume inputs of n tokens and that query, key and value vectors have the same dimension $d > 0$ to simplify notation. We denote $\mathbf{Q}, \mathbf{V} \in \mathbb{R}^{n \times d}$ the query and value matrices, and $\mathbf{k} \in \mathbb{R}^d$ the (single) key vector. For example, for a sentence of n tokens, we consider the attention mechanism for a single token, represented by key \mathbf{k} , and \mathbf{Q}_j and \mathbf{V}_j are the query and value vectors associated with input at position j , respectively.²

The attention mechanism computes an output vector $\mathbf{w} \in \mathbb{R}^d$ as follows (Vaswani et al., 2017):

$$\mathbf{w} \triangleq \mathbf{V}^\top \text{softmax} \left(\sqrt{d}^{-1} \mathbf{Q} \mathbf{k} \right).$$

To introduce relative positional information, Shaw et al. (2018) augment the attention mechanism with input dependent modifications of the key. Let $\mathbf{R} \in \mathbb{R}^{n \times d}$ be a matrix where \mathbf{R}_j is a vector encoding the relative distance with the j -th input, e.g if \mathbf{k} is associated with token i , \mathbf{R}_j could be set to an embedding representing the distance $i - j$. The attention mechanism is modified as follows:

$$\mathbf{w} \triangleq \mathbf{V}^\top \text{softmax} \left(\sqrt{d}^{-1} \begin{bmatrix} \mathbf{Q}_1^\top (\mathbf{k} + \mathbf{R}_1) \\ \vdots \\ \mathbf{Q}_n^\top (\mathbf{k} + \mathbf{R}_n) \end{bmatrix} \right),$$

which can be conveniently rewritten as:

$$= \mathbf{V}^\top \text{softmax} \left(\sqrt{d}^{-1} \left(\mathbf{Q} \mathbf{k} + \text{diag}(\mathbf{Q} \mathbf{R}^\top) \right) \right),$$

where $\text{diag}(\cdot)$ is the matrix diagonal as a vector.

3.2 Input Representation

In practice, inputs are documents that have been processed using an optical character recognition (OCR) model. The output of the OCR model is a sequence of tokens $\mathbf{x} = (x_1 \dots x_n)$ and a sequence of bounding boxes $\mathbf{b} = (b_1 \dots b_n)$, where

¹<https://github.com/buthaya/docpolarbert>

²We write \mathbf{Q}_i the i -th row of matrix \mathbf{Q} as a column vector.

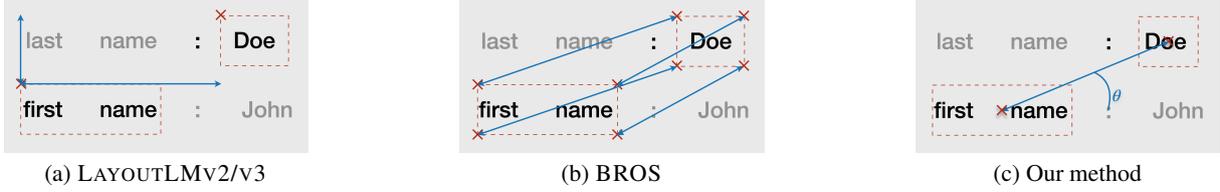


Figure 1: Illustration of relative positional attention in different models. Multiple words can share the same bounding box. Attention is skewed with respect to: (a) the horizontal and vertical distances between the top left corners; (b) distance between the four corners, respectively; (c) distance and angle between centers of the two bounding boxes.

b_i contains information about the position and size of the box containing the i -th token. Several tokens may share the same bounding box information, i.e. the i -th and $(i + 1)$ -th tokens can be in the same box. Note that even though tokens appear in the two-dimensional structure of the document, the linearization provided by the OCR can give crucial information for downstream tasks, and is therefore preserved in practice.

Tokens are embedded as follows:

$$E(x_i) \triangleq E_{\text{voc}}(x_i) + E_{\text{1D-pos}}(i) + \underbrace{E_{\text{2D-pos}}(b_i)}_{\text{for ablation only}},$$

where $E_{\text{voc}}(\cdot)$ is a standard token embedding table, $E_{\text{1D-pos}}(\cdot)$ and $E_{\text{2D-pos}}(\cdot)$ are learned (absolute) positional embedding tables. Following Xu et al. (2019), 1D positional embeddings are similar to the ones of BERT (Devlin et al., 2019), whereas 2D positional embeddings are the sum of four learned embeddings representing discretized absolute position of the bounding box of the token.

3.3 Relative Polar Coordinate Encoding

Let $\mathbf{c}_i = [c_{i,1}, c_{i,2}]^\top$ be the center of the bounding box b_i containing the i -th token expressed in the Cartesian coordinate system. Defining polar coordinates requires specifying an origin and a polar axis. As we aim to represent relative position between one input token (query position) and all other context tokens (key positions), we choose the center of the bounding box including the considered input (query) token as the origin. The polar axis is set to the first Cartesian dimension.

In the following, we assume the considered input is the i -th token. The relative position of the j -th with respect to the i -th one in the polar coordinate system is the tuple $(\rho(\mathbf{c}_i, \mathbf{c}_j), \theta(\mathbf{c}_i, \mathbf{c}_j))$ with:

$$\begin{aligned} \rho(\mathbf{c}_i, \mathbf{c}_j) &\triangleq \|\mathbf{c}_i - \mathbf{c}_j\|_2, \\ \theta(\mathbf{c}_i, \mathbf{c}_j) &\triangleq \text{atan2}(\mathbf{c}_j, \mathbf{c}_i), \end{aligned}$$

where we extend the atan2 to be null when its two inputs are equal, i.e. $\text{atan2}(\mathbf{c}_i, \mathbf{c}_i) = 0$. These relative polar coordinates are embedded using two learned embedding tables $E_{\text{dist}}(\cdot)$ and $E_{\text{angle}}(\cdot)$. To this end, we discretize the real values. For distances, we cap the maximum to a constant ρ_{max} and equally divide $[0, \rho_{\text{max}}]$ into 4 values. For angles, we discretize $(-\pi/2, \pi/2)$ into 8 values. More information is given in Section 4.3.

Let \mathbf{k} be the key associated with the i -th input, and \mathbf{Q} and \mathbf{V} defined as previously. Our relative polar coordinate attention mechanism is defined as follows. We build matrices $\mathbf{S}, \mathbf{T} \in \mathbb{R}^{n \times d}$ as:

$$\begin{aligned} \mathbf{S}_j &\triangleq E_{\text{dist}}(\rho(\mathbf{c}_i, \mathbf{c}_j)), \\ \mathbf{T}_j &\triangleq E_{\text{angle}}(\theta(\mathbf{c}_i, \mathbf{c}_j)). \end{aligned}$$

Then, the output vector \mathbf{w} is defined as:

$$\mathbf{w} \triangleq \mathbf{V}^\top \text{softmax} \left(\sqrt{d}^{-1} \begin{pmatrix} \mathbf{Q}\mathbf{k} \\ + \text{diag}(\mathbf{Q}\mathbf{S}^\top) \\ + \text{diag}(\mathbf{Q}\mathbf{T}^\top) \end{pmatrix} \right).$$

This modified attention mechanism accounts for relative distances between tokens, as well as relative angles around them in the polar basis.

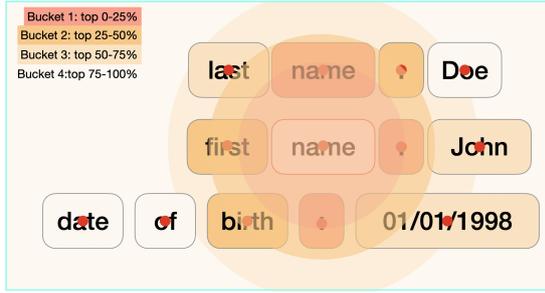
Figure 1 illustrates the key differences of our approach. The full model is illustrated in App. ??.

4 Positional Embeddings & Discretization

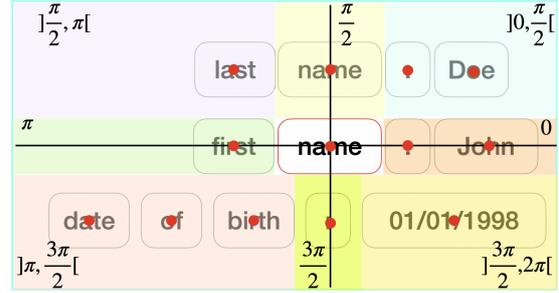
Positional embeddings are little discussed in previous work. In this section, we first review the main discretization strategies for absolute and relative positional embeddings, and then we motivate and propose a new approach for relative distances.

4.1 Absolute Positions in (Xu et al., 2019)

In LAYOUTLM (Xu et al., 2019), absolute token position embeddings are added to token embedding in the network input. For each token x_i , the model adds (1) a *textual position embedding* which corresponds to its sequence index i (regardless of



(a) Distances ρ , discretized using four different buckets.



(b) Angles θ .

Figure 2: Illustration of relative positions for the token “name” and its neighbors in polar coordinates.

its spatial position) and (2) a *layout embedding* corresponding to its bounding box b_i position.

In order to embed the bounding box position, its continuous position information must first be discretized. This transformation can be written as follows. Let $c \in \mathbb{R}$ be one of the four “raw” coordinates of a bounding box b_i (top-left, top-right, bottom-left or bottom-right). Its discretized embedding index c' is computed as follows:

$$c' = \left\lfloor \frac{c}{c_{\max}} \times m_{\text{abs.}} \right\rfloor,$$

where c_{\max} is the maximum vertical (resp. horizontal) page extent, $m_{\text{abs.}}$ is a training hyperparameter, and $\lfloor \cdot \rfloor$ is the floor function. The value $m_{\text{abs.}}$ act as a length normalization constant that defines the number of position embeddings to be learned.

Each discretized coordinate serves as an index for a learned embedding table. In practice, there is a separate embedding table for each of the four bounding box coordinate. The embeddings are then added to input tokens’ embeddings.

4.2 Relative Positions in Xu et al. (2021) and Huang et al. (2022)

LAYOUTLMV2 (Xu et al., 2021) and v3 (Huang et al., 2022) extend absolute positional information with relative information in the attention bias. Instead of using the same discretization strategy for both, they rely on the bucketing strategy of T5 (Raffel et al., 2020) for relative distances: rather than mapping distances linearly, they are assigned to buckets, with higher resolution for short distances and coarser resolution for long ones. This reflects the intuition that local order (e.g. adjacent words) requires precision, while very large distances can be captured approximately.

4.3 Our Approach

DOCPOLARBERT departs from Cartesian coordinates and instead represents spatial relations in polar form. For any pair of bounding box centers (c_i, c_j) , the distance $\rho(c_i, c_j)$ and angle $\theta(c_i, c_j)$ are discretized independently.

Angle. For angles, we discretize $[-\pi, \pi]$ into 8 buckets. Each corresponds to a main direction (above, below, left, right, and four diagonals areas). This reflects common document layouts, such as row/column alignments.

Distance. Distances are partitioned into buckets based on empirical quantiles. Unlike the scaling in the LAYOUTLMV2/V3 bucketing strategy, the quantiles are computed from the distribution of distances *within* each structured document. Thus, the same value of relative distance can be assigned to a different bucket in two different documents. This balances the representation by ensuring each bucket captures a comparable proportion of observed distances per document. Instead of relying on a fixed maximum distance hyperparameter, our strategy adapts to the each document’s distance distribution.

By combining distance and orientation, the representation captures relationships crucial for forms and tables. Figure 2 illustrates the resulting scheme.

We compare this approach with alternative bucketing strategies in Section 6.

5 Pre-training

Neural architecture. We build a BERT-like encoder of 12 layers, with 12 heads per layer. Queries, keys and embeddings are of dimension 768. Beside our custom relative attention, other parts of the neural architecture follows the LAYOUTLM model

Model	# Parameters (\downarrow)
Document Understanding Models	
LAYOUTLMV2 _{BASE}	200M
LAYOUTMASK _{BASE}	182M
LAYOUTLM _{BASE}	160M
LAYOUTLMV3 _{BASE}	133M
RELAYOUT _{BASE}	125M
DOCPOLARBERT _{BASE} (ours)	125M
Text-only Models	
ROBERTA _{BASE}	125M
BERT _{BASE}	110M

Table 1: Comparison of the number of parameters per model.

(Xu et al., 2019), except that we *do not introduce* absolute position information.

The input consists of a global 1D position embedding and a token embedding to capture semantic information. Note that bounding box coordinates are used exclusively to compute distances and angles, which introduce spatial biases in the self-attention mechanism through a polar coordinate representation. No vision features are required for this model.

Our model has \simeq 125M parameters, making it ideal for efficient use in real-world applications as it is smaller than other layout models, see Table 1.

Training objectives. Following standard encoder-based models, we apply a masked language modeling loss with 30% token masking. Additionally, we experiment with the 1D local order prediction (1-LOP) loss (Jiang et al., 2025), which predicts token positions in the linearized OCR output when 1D positional embeddings are masked (30% of tokens).

Data. We initially sought to pre-train on the widely used IIT-CDIP dataset (Lewis et al., 2006), but its publicly available version lacks 2D layout annotations.

To ensure reproducibility of our findings, and eliminate dependency on proprietary OCR systems, we opted for datasets in which text and layout annotations are publicly available: OCR-IDL (Biten et al., 2023) and DOCILE (Šimsa et al., 2023). The DOCILE dataset consists of over 900k unlabeled public invoices, whereas OCR-IDL is a diverse collection of 26M documents (letters, reports, memos, news articles, etc.) sourced from the industry document library³ (IDL), the same repository used to construct IIT-CDIP. For compu-

tational efficiency and dataset balance, we sample 1.8M documents, evenly split between DOCILE and OCR-IDL. While our final corpus is over six times smaller than IIT-CDIP, our results show that it is sufficiently large for model pre-training.

Training settings. We initialize from the weights of the pre-trained ROBERTA model (Liu et al., 2019) and use a minibatch size of 2048 documents. We ran training for 20 epochs with a target learning rate of 1×10^{-4} in a cosine scheduler (Loshchilov and Hutter, 2017) with warmup on 5% of the updates. Pre-training is done on NVIDIA Tesla V100 32GB GPUs.

6 Experiments

6.1 Evaluation on Downstream Tasks

Data. We evaluate our model on five datasets for named-entity recognition on documents. FUNSD (Guillaume Jaume, 2019) is a dataset of 199 forms from the Truth Tobacco Industry Document (TTID) collection. This collection is also the source of the IIT-CDIP (Lewis et al., 2006) and the RVL-CDIP (Harley et al., 2015) datasets. SROIE and CORD v2 (Huang et al., 2019; Park et al., 2019) are datasets of respectively 973 and 1,000 annotated receipts. DOCILE (Šimsa et al., 2023) labeled subset with 6,680 annotated invoices.⁴ PAYSLIPS (Uthayasooriyar et al., 2025) is a dataset of 611 pages of financial documents (pay statements) from the insurance sector. Dataset statistics are given in Table 2.

Training settings. We fine-tune all models for 30 epochs with a minibatch size of 16 and a fixed learning rate of 5×10^{-5} , except for DOCILE* for which we use a learning rate of 1×10^{-5} . To ensure robust evaluation, we report the average F1-score across 10 fine-tuning runs with different seeds, except for DOCILE*, where we conduct only 4 runs to reduce computational costs.

Results. We report results on these datasets for several baselines and for variants of our model in Table 3. For reference, we also report results for models that are not directly comparable with DOCPOLARBERT, either because they are based on different modalities (vision features) or because they are not publicly available, hence we couldn’t reproduce their results.

⁴Since the test set of DOCILE is not publicly available, we use the validation set for testing purposes.

³<https://www.industrydocuments.ucsf.edu>

	PAYSLIPS	DOCILE	FUNSD	SROIE	CORD
Train / Val / Test sizes	485/-/126	6,759/635/1,000	149/-/50	626/-/347	800/100/100
% of O	94.95	89.46	0	83.82	0
Document types	Pay Statements	Invoices	Forms	Receipts	Receipts
Entity types	Dates, Amounts	Text, Dates, Amounts	Text	Text, Dates, Amounts	Text, Dates, Amounts

Table 2: Fine-tuning datasets description.

Model / Pre-training data	Mod.	FUNSD	SROIE	CORD	PAYSLIPS	DOCILE*	AVG
For reference only							
LAYOUTLM _{BASE} (Xu et al., 2019)	T+L+I	79.27	94.67	-	-	-	-
LAYOUTLMV3 _{BASE} (Huang et al., 2022)	T+L+I	90.29	-	96.56	-	-	-
RELAYOUT _{BASE} (Jiang et al., 2025)	T+L	84.64	-	96.82	-	-	-
LAYOUTMASK _{BASE} (Tu et al., 2023)	T+L	92.91	96.87	96.99	-	-	-
Baselines							
LAYOUTLM _{BASE} (IIT-CDIP)	T+L	<u>78.66</u>	94.38	95.66	62.31	58.35	77.87
LAYOUTLM _{BASE} (DOCILE)	T+L	67.24	91.39	91.57	64.74	58.30	74.65
BROS _{BASE} (Hong et al., 2022)	T+L	83.05	96.28	96.50	58.15	57.09	78.21
LILT _{BASE} (Wang et al., 2022)	T+L	77.35	95.31	<u>96.07</u>	65.34	60.05	78.82
DOCPOLARBERT (Ours)							
DOCILE	T+L	57.28	94.02	91.78	76.85	58.14	75.61
+ OCR-IDL	T+L	78.57	96.96	95.97	73.54	<u>59.73</u>	80.95
+ I-LOP	T+L	78.26	<u>96.53</u>	95.15	<u>75.40</u>	61.36	81.34

Table 3: NER F1-score. Modalities are text (T), layout (L), and image (I). Results shown in *italic* are from the original model papers. Bold and underline fonts emphasize best and second results for each dataset.

On all datasets except FUNSD and CORD, our model outperforms baselines. We hypothesize FUNSD’s results are influenced by its origin from IIT-CDIP, used to pre-train all the other models. For CORD, we observed that our model often confuses semantically similar or hierarchically related labels (e.g. subtypes within the “menu” category). This is likely due to limitations of polar-relative encoding in capturing fine-grained structures in dense regions with multiple mentions.

Interestingly, our approach largely outperforms baselines on PAYSLIPS, that contains mainly large tables of numbers. This suggests our relative encoding strategy of layout structures better captures long-range dependencies, particularly in cases where entity values must be correctly associated with corresponding column headers.

6.2 Comparison with Vision-stripped LAYOUTLMV3 Model

To assess the benefit of our polar-based representation in a vision-agnostic setting, we compare DOCPOLARBERT with a modified LAYOUTLMV3 that excludes visual inputs, denoted LAYOUTLMV3*. To this end, we pre-train the LAYOUTLMV3* model on the same corpus as ours, combining DOCILE and OCR-IDL, in order to

isolate the effect of textual and layout information without interference from visual features.

As reported in Table 5, DOCPOLARBERT consistently surpasses LAYOUTLMV3* across all downstream tasks, by a large margin. This indicates that encoding layout using relative polar coordinates better captures structural and semantic patterns from text and layout alone. In practical scenarios where visual data is unavailable, e.g. documents containing sensitive imagery in insurance or healthcare, models designed explicitly for textual and layout features, like DOCPOLARBERT, offer a clear advantage. In contrast, removing visual components from multimodal models like LAYOUTLMV3 leads to degraded performance.

6.3 Attention Analysis

To better understand how the proposed spatial bias affects model behavior, we manually examined attention distributions in fine-tuned models.

First, we focus on samples from the PAYSLIPS dataset where target amounts were correctly classified. For each document, we compare attention patterns of LAYOUTLM, LAYOUTLMV3*, and DOCPOLARBERT with respect to the token corresponding to the predicted amount (Figure 3). In LAYOUTLM and LAYOUTLMV3*, attention

Model	Per-document (1-GPU)	Per-document (4-GPU)	Total dataset time (1-GPU)	Total dataset time (4-GPU)	4-GPU Speedup
LAYOUTLM	9.5	4.8	1873.4	2265.6	0.83×
LAYOUTLMV3*	<u>12.7</u>	<u>5.5</u>	1750.2	1768.1	0.99×
LILT	13.0	7.3	<u>1790.7</u>	2100.9	0.85×
BROS	21.6	7.8	2963.8	2318.7	<u>1.28</u> ×
DOCPOLARBERT	25.9	8.8	3428.1	<u>1879.1</u>	1.82 ×

Table 4: Inference efficiency comparison on the PAYSILIPS dataset. Average per-document latency and total dataset processing time (in ms) across single- and multi-GPU setups. GPUs are NVIDIA V100. Speedup indicates the ratio of 1-GPU to 4-GPU total inference times. Although our model is slower per document, it scales well over 4-GPUs.

	LAYOUTLMV3*	DOCPOLARBERT
FUNSD	66.79	78.57
SROIE	89.83	96.96
CORD	90.90	95.97
PAYSILIPS	44.44	73.54
DOCILE*	59.66	59.73
AVG	70.32	80.95

Table 5: F1-score comparison of LAYOUTLMV3* and DOCPOLARBERT. Both models are pre-trained on OCR-IDL + DOCILE with MLM loss only.

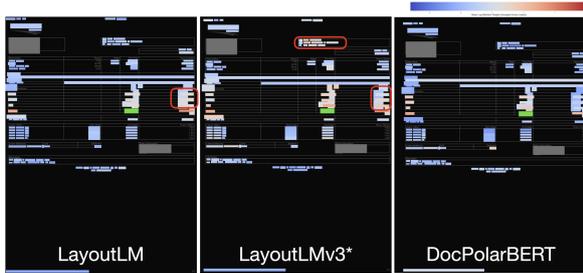


Figure 3: Heatmap of average attention (across all heads and layers) for a NET_PAY_PER_PERIOD token (green box) on a PAYSILIPS document. Blue-to-red values indicate low-to-high attention. We compare LAYOUTLM, LAYOUTLMV3*, and DOCPOLARBERT, highlighting with red bounding boxes the unwanted activations.



Figure 4: Heatmap of average attention for a POST_TAX_DEDUCTIONS_PER_PERIOD token (green box) on a PAYSILIPS document. We compare LAYOUTLM, LAYOUTLMV3*, and DOCPOLARBERT, highlighting with red bounding boxes the unwanted activations. In this example, LAYOUTLM and LAYOUTLMV3 misclassified the target amount, while DOCPOLARBERT correctly identifies it.

weights were frequently assigned to tokens without clear semantic or spatial relevance, such as unrelated text in distant parts of the document. By contrast, DOCPOLARBERT consistently concentrated on contextually and spatially coherent elements, including nearby numeric values and labels such as “Net pay”, “Gross”, or “Tax”, often aligned horizontally or vertically with the target token.

We conducted a similar analysis on instances where only DOCPOLARBERT predicted the correct output while the other models failed (Figure 4). In these cases, LAYOUTLM and LAYOUTLMV3* also directed attention to irrelevant surrounding tokens or to distant header information, whereas DOCPOLARBERT maintained focus on semantically related and spatially aligned elements.

Overall, encoding spatial relations in polar form seems to favor more coherent attention behavior, particularly in documents where layout strongly conveys meaning.

6.4 Length Generalization

Length generalization is a common setting in Transformer evaluation (Bhattachishra et al., 2020; Anil et al., 2022; Wang et al., 2024b, *inter alia*) and structured prediction (Herzig and Berant, 2021; Yao and Koller, 2022; Wu et al., 2023). We propose to adapt this setting to document understanding. Each dataset is sorted by token length, and we split the datasets between train and test sets so that they contain the same number of documents in each split as the original dataset. The training set includes the shortest documents, while the test set contains the longest ones.

We evaluate LAYOUTLM, LAYOUTLMV3*, and DOCPOLARBERT, pre-trained with masked language modeling on DOCILE. The results are shown in Table 6. DOCPOLARBERT achieves the highest average F1-score, slightly surpassing both LAYOUTLM and LAYOUTLMV3*.

However, performance differences across

	LAYOUTLM	LAYOUTLMv3*	DOCPOLARBERT
FUNSD	54.02	70.47	62.10
SROIE	78.38	86.49	82.50
CORD	84.71	87.25	83.56
PAYSLIPS	77.80	74.22	85.10
DOCILE*	60.17	54.45	61.04
AVG	71.02	74.58	74.86

Table 6: F1-scores of LAYOUTLM, LAYOUTLMv3*, and DOCPOLARBERT on the sequence-length generalization task. Models are trained on the shortest documents and tested on the longest ones. All are pre-trained on DOCILE using masked language modeling only.

datasets reveal more nuanced patterns. On FUNSD, DOCPOLARBERT underperforms compared to LAYOUTLMv3*, with similar behavior on SROIE and CORD. This difference can be attributed to the relatively simple spatial layouts of these datasets, where most information aligns along clear horizontal and vertical axes. In such cases, Cartesian coordinates offer an adequate way to capture spatial relationships, while polar coordinate system may add over-complex geometric abstraction.

In contrast, DOCPOLARBERT performs best on PAYSLIPS and DOCILE*, which contain highly structured tables with clear header-value pairings. In these documents, the polar encoding captures relational patterns soundly, even when document size or layout changes. This suggests that polar-relative encoding introduces a strong structural bias beneficial for layout-driven understanding of documents with long tables.

6.5 Efficiency analysis

DOCPOLARBERT shows a noticeable slowdown during inference compared to baseline architectures. As shown in Table 4, per-document inference times are higher, primarily due to the size and structure of the relative bias matrices in our attention mechanism.

Despite the higher per-document cost, DOCPOLARBERT demonstrates superior scaling efficiency on multi-GPU setups. Table 4 shows substantial reductions in inference times over the whole dataset when moving from 1-GPU to 4-GPUs.⁵

⁵4-GPU epoch times can exceed 1-GPU times due to fixed multi-GPU overheads, such as inter-GPU communication and synchronization, which may outweigh parallelization gains for already optimized single-GPU tasks.

	w 2D-Pos	w/o 2D-Pos
FUNSD	76.22	78.26
SROIE	96.52	96.53
CORD	94.63	95.15
PAYSLIPS	72.52	75.40
DOCILE*	61.27	61.36
AVG	80.23	81.34

Table 7: DOCPOLARBERT F1-score comparison with and without absolute 2D positional embeddings.

7 Ablation Study

7.1 Absolute Positional Embeddings

Contrary to previous work, our model does not include absolute 2D positional embeddings. We therefore conduct a study to analyse the impact of absolute positional embeddings in DOCPOLARBERT.

Table 7 compares NER results for DOCPOLARBERT with and without such embeddings. We observe that absolute 2D positional embeddings, that are always provided in baselines, deteriorates downstream task results.

7.2 Discretization Strategies

As discussed in Section 4, previous models use a different discretization method for encoding relative position information. We therefore conduct experiments to evaluate their impact.

In the following, we compare variant of DOCPOLARBERT where the distance coordinate is discretized:

- Using a similar strategy as in LAYOUTLMv2 and v3, with varying number of bins (4 to 32), see Section 4.2;
- Using our quantile-based strategy, but with different number of bins (4 to 32), see Section 4.3.

Bucketing method	LAYOUTLMv2/v3				Quantile-based			
Nb. of buckets	32	16	8	4	32	16	8	4
FUNSD	79.06	78.81	78.51	80.01	79.44	76.71	79.90	78.26
SROIE	97.42	97.69	97.31	97.51	97.74	97.14	97.68	96.53
CORD	96.94	97.38	96.46	97.23	96.77	96.92	96.78	95.15
PAYSLIPS	73.76	73.40	73.37	73.53	73.66	72.81	74.27	75.40
DOCILE*	57.87	58.82	57.82	55.00	58.31	57.83	56.06	61.36
AVG	81.01	81.22	80.70	80.66	81.18	80.28	80.94	81.34

Table 8: NER F1-score of DOCPOLARBERT with the bucketing method from LAYOUTLMv2/v3 vs our quantile-based strategy applied to relative distances. Each variant is pre-trained on OCR-IDL+DOCILE with MLM and 1-LOP.

Results are given in Table 8.

Overall, results remain close across strategies, with less than two points variation in average F1. This shows that DOCPOLARBERT is generally robust to the exact choice of bucketing.

Nonetheless, some dataset-specific patterns are visible when considering the number of buckets. On FUNSD, using fewer buckets appears advantageous, with the T5 4-bucket variant reaching the highest score (80.01). In contrast, SROIE benefits from finer granularity, where the quantile variant with 32 buckets achieve the best results (97.74). CORD shows a consistent preference for 16 buckets across both strategies. For PAYSLIPS and DOCILE*, quantile-based schemes perform slightly better, with the 4-quantile variant reaching the highest scores, which suggests that adapting buckets to the empirical distribution can be advantageous for tabular or invoice-like layouts.

On average, the 4-quantile approach achieves the best performance (81.34), albeit by a narrow margin over the best T5 variant (81.22). Therefore, while the overall effect of bucket choice is limited, adapting bucket counts to dataset characteristics can offer small but consistent benefits.

8 Conclusion

We propose a novel relative positional encoding of layout structure based on the polar coordinate system. This approach allows the model to directly encode semantically crucial information (e.g. table headers are at the top row, no matter how far it is). Experimentally, our approach outperforms comparable models.

9 Limitations

Our proposed document encoder model requires text and layout annotations from an OCR model,

which is different from end-to-end systems that directly process raw documents. Note that although running an OCR can be a costly step, it is nevertheless often done in practical settings, where the data is preprocessed before being archived.

We evaluate our approach only on named-entity recognition, which is the most common benchmark for layout models, but it may not cover all use cases.

Finally, our pre-training data consists of a mix of OCR-IDL and DOCILE, totaling 1.8M documents, whereas many document understanding models leverage IIT-CDIP, which contains 11M documents. We opted for datasets with openly available text and layout annotations to ensure reproducibility. This choice facilitates fair comparison and avoids the reproducibility issues common in document understanding research, where OCR outputs are often proprietary and can heavily affect downstream performance.

Acknowledgement

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015001).

References

- Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. [Exploring length generalization in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 38546–38556. Curran Associates, Inc.
- Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.
- Haoli Bai, Zhiguang Liu, Xiaojun Meng, Li Wentao,

- Shuang Liu, Yifeng Luo, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang, and Qun Liu. 2023. [Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13386–13401, Toronto, Canada. Association for Computational Linguistics.
- Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020. [On the Ability and Limitations of Transformers to Recognize Formal Languages](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, Online. Association for Computational Linguistics.
- Ali Furkan Biten, Rubèn Tito, Lluís Gomez, Ernest Valveny, and Dimosthenis Karatzas. 2023. [OCR-IDL: OCR Annotations for Industry Document Library Dataset](#), page 241–252. Springer Nature Switzerland.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. [Unidoc: Unified pretraining framework for document understanding](#). *Advances in Neural Information Processing Systems*, 34:39–50.
- Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. 2019. [Funsd: A dataset for form understanding in noisy scanned documents](#). In *Accepted to ICDAR-OST*.
- Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. [Evaluation of deep convolutional nets for document image classification and retrieval](#). In *International Conference on Document Analysis and Recognition (ICDAR)*.
- Jonathan Herzig and Jonathan Berant. 2021. [Span-based semantic parsing for compositional generalization](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921, Online. Association for Computational Linguistics.
- Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. [Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.
- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. [Layoutlmv3: Pre-training for document ai with unified text and image masking](#). In *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, page 4083–4091, New York, NY, USA. Association for Computing Machinery.
- 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](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520.
- Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. [Spatial dependency parsing for semi-structured document information extraction](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. Association for Computational Linguistics.
- Zhouqiang Jiang, Bowen Wang, Junhao Chen, and Yuta Nakashima. 2025. [ReLayout: Towards real-world document understanding via layout-enhanced pre-training](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3778–3793, Abu Dhabi, UAE. Association for Computational Linguistics.
- D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. [Building a test collection for complex document information processing](#). In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR06*, page 665–666. ACM.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Ilya Loshchilov and Frank Hutter. 2017. [SGDR: Stochastic gradient descent with warm restarts](#). In *International Conference on Learning Representations*.
- Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. [Layoutllm: Layout instruction tuning with large language models for document understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.
- Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. [Cord: a consolidated receipt dataset for post-ocr parsing](#). In *Workshop on Document Intelligence at NeurIPS 2019*.
- Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. [Going full-tilt boogie on document understanding with text-image-layout transformer](#). In *Document Analysis and Recognition – ICDAR 2021*,

- pages 732–747, Cham. Springer International Publishing.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. [Self-attention with relative position representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.
- Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19071–19079.
- Yi Tu, Ya Guo, Huan Chen, and Jinyang Tang. 2023. Layoutmask: Enhance text-layout interaction in multi-modal pre-training for document understanding. *arXiv preprint arXiv:2305.18721*.
- Benno Uthayasooriyar, Antoine Ly, Franck Vermet, and Caio Corro. 2025. [Training LayoutLM from scratch for efficient named-entity recognition in the insurance domain](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 101–110, Abu Dhabi, UAE. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. [DocLLM: A layout-aware generative language model for multimodal document understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.
- Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. [LiLT: A simple yet effective language-independent layout transformer for structured document understanding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7747–7757, Dublin, Ireland. Association for Computational Linguistics.
- Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. 2024b. [Length generalization of causal transformers without position encoding](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14024–14040, Bangkok, Thailand. Association for Computational Linguistics.
- Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. [ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation](#). *Transactions of the Association for Computational Linguistics*, 11:1719–1733.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. [LayoutLMv2: Multi-modal pre-training for visually-rich document understanding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.
- 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](#).
- Yuekun Yao and Alexander Koller. 2022. [Structural generalization is hard for sequence-to-sequence models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Štěpán Šimsa, Milan Šulc, Michal Uříčář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. [DocILE Benchmark for Document Information Localization and Extraction](#). In *Document Analysis and Recognition - IC-DAR 2023*, pages 147–166, Cham. Springer Nature Switzerland.