Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain

Benno Uthayasooriyar[1, 2] **Antoine Ly**[1] **Franck Vermet**[2] **Caio Corro** [2]
[1]Data Analytics Solutions, SCOR [2]Univ Brest, CNRS, UMR 6205, LMBA
[3]INSA Rennes, IRISA, Inria, CNRS, Université de Rennes

Abstract

Generic pre-trained neural networks may struggle to produce good results in specialized domains like finance and insurance. This is due to a domain mismatch between training data and downstream tasks, as in-domain data are often scarce due to privacy constraints. In this work, we compare different pre-training strategies for LayoutLM. We show that using domain-relevant documents improves results on a named-entity recognition (NER) problem using a novel dataset of anonymized insurance-related financial documents called Payslips. Moreover, we show that we can achieve competitive results using a smaller and faster model.

1 Introduction

Modern natural language processing pipelines heavily rely on pre-trained neural networks, primarily language models (Schwenk and Gauvain, 2005; Jozefowicz et al., 2016; Radford et al., 2019, *inter alia*) and context-sensitive embeddings (Schütze, 1998; Peters et al., 2018; Devlin et al., 2019, *inter alia*). The development of neural architectures based on the attention mechanism (Bahdanau et al., 2015) allows to efficiently pre-train them on GPU using large datasets (Vaswani et al., 2017): most recent networks can contain several hundreds of billions parameters (*e.g.,* Chowdhery et al., 2023).

Despite their experimental success, commercial use of pre-trained neural networks can be limited for the following reasons. Firstly, downstream tasks in information retrieval may require to continuously analyze large amounts of data, which prevents the use of the largest models due to inference time bottleneck. Secondly, applications in specific fields such as financial, medical or insurance, can forbid the use of API-based models due to privacy concerns. Thirdly and lastly, authors may at some point decide to not publicly share latest versions of their models, or to change the license to forbid commercial use.[1] As such, it is increasingly important to ensure replicability and robustness to changes in training data (including for domain transfer) not only for scientific reasons, but also to ensure widespread commercial deployment.

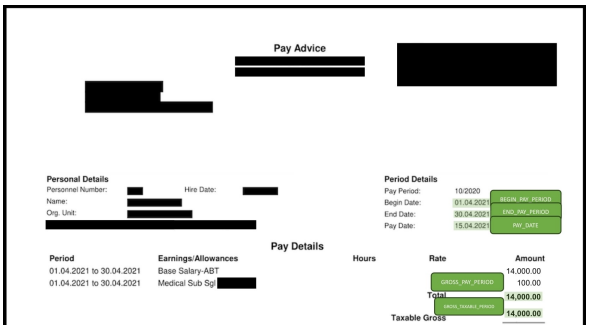

Figure 1: Sample of the newly introduced Payslips dataset for named-entity recognition in the insurance domain.

In this work, we study LayoutLM (Xu et al., 2020) for named-entity recognition (NER) on financial documents from the insurance domain. Our aim is to understand how such a model can be used in a constrained setting: Can performance in downstream tasks be improved by pre-training on domain-specific documents, even when the amount of available data is limited? Can inference time be improved while maintaining downstream performance? To address this, we pre-train several models from scratch using a smaller, but more relevant, set of publicly available documents.

To evaluate these models, we build a novel dataset, Payslips, that contains anonymized insurance pay statements with annotated financial information for NER, detailed in Table 1. Although these documents are private, we have manually anonymized them. Our experiments show that pre-training on documents that are semantically and structurally similar to those in the downstream

[1]See for example LayoutLMv2 and LayoutLMv3.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 101–110
January 19–20, 2025.

task leads to improved performance, even with less training data. Moreover, if speed of inference is crucial, we show that comparable results can be obtained by using only half the number of layers compared to the original LAYOUTLM model.

Our contributions can be summarized as follows:

- We build and release PAYSLIPS, a novel NER dataset of 611 labeled pages of anonymized payslips from the insurance domain;
- We pre-train a LAYOUTLM network using a smaller set of documents (DOCILE, Šimsa et al., 2023);
- We evaluate our model on PAYSLIPS and show that not only does it achieve better F1 scores, but it also has a lower variance;
- We show that a smaller model with half the number of layers maintains performances while improving computational efficiency.

Our code and data are publicly available.[2]

2 Related Work

Contextual embeddings. Peters et al. (2018) first proposed to pre-train a bidirectionnal LSTM on large corpora to learn context-sensitive word embeddings that can be used to improve results on downstream tasks. The BERT model (Devlin et al., 2019) instead uses a self-attentive network (i.e. the encoder part of a transformer) to take full advantage of GPU architectures. However, BERT cannot be trained using the standard language modeling objective as it is not an autoregressive model. Instead, the authors proposed a *masked language modeling* objective where the loss aims to increase a reconstruction term on a hidden part of the input.

Document analysis. For document processing, one must take into account spatial information together with textual content. LAYOUTLM (Xu et al., 2020) extends BERT's positional embeddings with spatial positions. In other words, BERT uses as input embeddings representing the position in the sequence,[3] whereas LAYOUTLM also includes 6-tuples of embeddings describing (discretized) positions and sizes of the boxes containing one or several words. This allows the self-attentive network to capture spatial information, which is especially useful for documents containing tables and/or processed with optical character recognition.[4]

LAYOUTLMV2 (Xu et al., 2021) and V3 (Huang et al., 2022) incorporate more visual information, both as input and in auxiliary training losses. Moreover, the architecture is modified to integrate relative positional information. Li et al. (2021) introduced richer positional information, whereas Wang et al. (2022) focused on language adaptation during the fine-tuning phase. Contrary to these works, we focus on the original LAYOUTLM model as we aim for computational efficiency.

Efficient encoders. The self-attention mechanism has a quadratic-time complexity with respect to the input, which can be slow for long documents. Several works in document analysis (Nguyen et al., 2021; Douzon et al., 2023, *inter alia*) have addressed these drawbacks by integrating more computationally efficient types of attention that are better motivated for document processing. In this work, we instead explore the impact of the number of layers on downstream results.

Label	Train	Test
BEGIN_PAY_PERIOD	236	85
END_PAY_PERIOD	388	100
PAY_DATE	461	101
GROSS_PAY_PERIOD	481	117
GROSS_TAXABLE_PERIOD	245	90
NET_PAY_PERIOD	444	109
PAYG_TAX_PERIOD	499	119
PRE_TAX_DEDUCTION_PERIOD	278	68
POST_TAX_DEDUCTION_PERIOD	243	67
O	60,596	23,228
Total	63,871	24,084

Table 1: Label distribution in PAYSLIPS dataset (word level).

3 Payslips Dataset

We build a novel dataset containing financial pay statements from the insurance domain which we call PAYSLIPS. This dataset consists of a training set of 485 pages and a test set of 126 pages.

The documents originate from data of disability insurance. In the event of a work-related accident, this insurance product compensates the insured person during their recovery period. To determine the indemnity amount, the insurer verifies salary information from each insured person's payslip. To speed up information processing, it is essential to build tools capable of automatically extract-

[2] https://github.com/buthaya/payslips

[3] Note that some models uses positional encoding without relying on an embedding table, see for examples (Vaswani et al., 2017, Section 3.5)

[4] OCR's outputs are composed of boxes containing part of the document text.

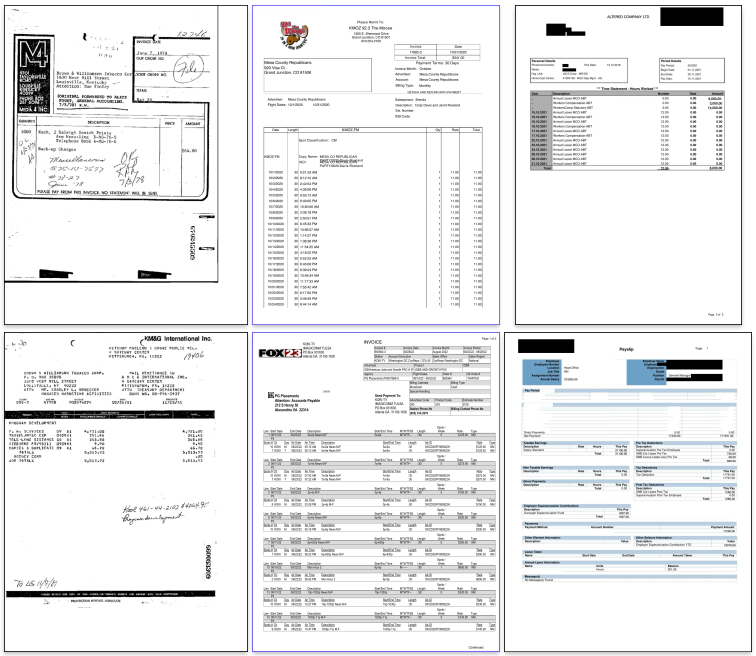

Figure 2: Samples from IIT-CDIP (first column), DOCILE (second column) and PAYSLIPS (third column) datasets. Invoices from DOCILE and pay statements from PAYSLIPS are closer visually and semantically.

ing the useful financial information. To this end, we worked with insurance professionals and identified nine specific fields, as detailed in Table 1. The task is therefore reduced to a standard NER problem, similar to what is done in the FUNSD dataset (Jaume et al., 2019). Unlike datasets such as FUNSD or CORD (Park et al., 2019), PAYSLIPS is notably sparse, with a predominant O class, which poses a challenge for the information extraction task. We explain in more details the particularities and challenges of the PAYSLIPS dataset and other usual datasets for NER in documents in Appendix C.

PAYSLIPS was annotated in-house by people familiar with the documents. Then, samples have been validated by insurance specialists to ensure annotation quality. More details about the annotation process are given in Appendix D.

For privacy reasons, unnecessary or potentially identifying information was altered or deleted. Moreover, images are not shared as they are not used by our LAYOUTLM model, but could give visual cues about the entity emitting the files.

4 LayoutLM

4.1 Neural Architecture

We use the LAYOUTLM model (Xu et al., 2020), which is based on the same neural architecture as BERT (Devlin et al., 2019), but where inputs are tailored to represent texts in 2D documents.

Given a document where the content has been divided into text blocks, each individual block is encoded as follows: (1) words are tokenized; (2) each token is represented by an embedding; (3) 2D positional embeddings are added to word embeddings. The 2D positional embeddings are 6-tuples representing the coordinates of the block in the page's image, and its height and width, discretized and normalized between 0 and 1000.

The original LAYOUTLM could also incorporate an image embedding derived from a vision model. We do not include this input as it slows down the model without significant impact on downstream task results — sometimes the impact is even negative, see (Xu et al., 2020, Table 4).

Then, for the self-attentive part, we use the BASE model, which consists of 12 self-attentive layers. Each layer contains 12 heads of dimension 768, as originally defined by Devlin et al. (2019). Finally, during pre-training, the output contextual embed-

dings are projected into the vocabulary space using a linear layer to compute output logits.

4.2 Pre-training

Loss. We pre-train the model using a Masked Language Modeling (MLM) loss, where part of the input is replaced by dummy embeddings and a negative log-likelihood loss aims to reconstruct the masked part. Xu et al. (2020) also experimented with a Multi-label Document Classification (MDC) loss, a supervised task aiming to classify each page into predefined categories. Their results show that MDC degrades performances, therefore we do not include this loss during pre-training.

Data. LAYOUTLM was pre-trained on the IIT-CDIP dataset (Schmidt et al., 2002; Lewis et al., 2006), which gathers 11 millions documents from the U.S. state lawsuits against the tobacco industry in the 1990s. The authors show that pre-training on this data improves results on several downstream tasks, including NER on SROIE (Huang et al., 2019) and FUNSD (Jaume et al., 2019). Unfortunately, during preliminary experiments we observed that LAYOUTLM tends to under-perform on our internal data. We suspect IIT-CDIP documents are too different in form and content from insurance documents (see Figure 2). We give more insights about these differences in Appendix C. Moreover, adapting information retrieval systems to the insurance domain poses significant challenges due to the sensitivity of the data involved, i.e. we cannot train and distribute models based on internal data due to private data protection laws.

We found no existing datasets of pay statements. However, some relevant invoice datasets are available. Limam et al. (2023) provides a dataset of generated invoices, and RVL-CDIP (Harley et al., 2015) includes a subset of invoices from the IIT-CDIP collection. A more recent and larger dataset, DOCILE (Šimsa et al., 2023), offers a better match in terms of layout and semantics with our downstream task dataset, PAYSLIPS, as shown in Figure 2. It contains approximately 900k unlabeled invoices sourced from two public repositories.[5][6] Although it is more than 10 times smaller than IIT-CDIP, our experimental results shows that it is big enough for pre-training LAYOUTLM.

Technical details. We pre-train LayoutLM from scratch with the MLM loss on the DOCILE dataset, with similar settings to Xu et al. (2020). We use a minibatch size of 80, and ran the training for 5 epochs with a learning rate of 5×10^{-5}. We use a cosine scheduler with warmup on 5% of the updates. Pre-training is done on 8 NVIDIA Tesla V100 16GB GPUs.

Model	F1 DOCILE labeled	F1 PAYSLIPS
Pre-training on IIT-CDIP		
LAYOUTLM$_{\text{BASE}}$	$\mathbf{58.35 \pm 1.63}$	$\underline{62.31 \pm 5.13}$
Pre-training on DOCILE		
LAYOUTLM$_{\text{BASE}}$	$\underline{58.30 \pm 1.52}$	$\mathbf{64.74 \pm 2.92}$
LAYOUTLM$_{\text{6 layers}}$	57.38 ± 1.38	61.80 ± 3.12
LAYOUTLM$_{\text{2 layers}}$	53.89 ± 1.03	54.61 ± 3.71
LAYOUTLM$_{\text{1 layer}}$	51.12 ± 1.53	45.08 ± 3.31

Table 2: F1 scores for named-entity recognition using different pre-training and fine-tuning datasets. Results are averaged on 100 runs with different seeds.

Model	Inference Time (ms)
LAYOUTLM$_{\text{BASE}}$	12.10
LAYOUTLM$_{\text{6 layers}}$	6.15
LAYOUTLM$_{\text{2 layers}}$	2.42
LAYOUTLM$_{\text{1 layers}}$	1.73

Table 3: Inference times per page on the PAYSLIPS dataset. Tests were conducted on a machine equipped with a single NVIDIA Tesla V100 32GB GPU.

5 Experiments

We tackle the NER problem using the standard BIO-tagging approach (Ramshaw and Marcus, 1995), i.e. each token is tagged with either O (not in a mention), B-LABEL (beginning of a mention) or I-LABEL (inside of a mention), where LABEL is any mention label allowed in the dataset. We can then trivially rebuild the full predicted mentions from the predicted BIO tags.

We fine-tune all models with a batch size of 16 for 10 epochs, with a fixed learning rate of 5×10^{-5}.

5.1 Results

We compare the original LAYOUTLM pre-trained on IIT-CDIP with our LAYOUTLM pre-trained (from scratch) on DOCILE on two NER datasets: (1) The subset of the DOCILE dataset which is labeled[7] — it contains 6759 and 635 document pages for training and testing, respectively; (2) Our

[5] https://www.industrydocuments.ucsf.edu/

[6] https://publicfiles.fcc.gov

[7] As the annotation of the test set are not available online, we performed evaluation on the validation set.

novel PAYSLIPS dataset — statistics are reported in Section 3. We fine-tune similarly for both datasets.

We report labeled F1-score averaged on 100 fine-tuning runs in Table 2. Precision and recall are reported in Appendix B. The BASE model (using the full 12 layers) produces similar results on DOCILE no matter if pre-training on IIT-CDIP or DOCILE. However, on our internal PAYSLIPS datasets, our model pre-trained on DOCILE outperforms the original one. Moreover, we observe that our pre-trained model exhibits a way lower variance between fine-tuning runs.

To cope with the high and continuous flow of documents, an insurer might require a faster model. Therefore, we also experimented using a smaller number of self-attentive layers, see Table 2. Inference times per model are reported in Table 3. On PAYSLIPS, when pre-training on DOCILE using only 6 layers, we achieve comparable scores to the off-the-shelf LAYOUTLM model, while dividing the inference time by almost 2.

5.2 Statistical Significance

Domain-specific datasets are often of small sizes, so comparing F1-scores may lead to wrong conclusion if they are not statistically significant. We follow the original Message Understanding Conference (MUC, Chinchor, 1992; Chinchor et al., 1993) and rely on the approximate randomization method (Noreen, 1989), which does not require assumptions on the data distribution. For this test, the null hypothesis is "*The proposed system and the baseline system do not differ in F1*". The difference is computed in term of absolute F1 difference over many random data splits. Pseudo-code is given in Appendix A.

In our case, we compare the LAYOUTLM pre-trained on IIT-CDIP to the one pre-trained on DOCILE, both being fine-tuned on PAYSLIPS. As we did 100 fine-tunings, we took two models with a F1-score difference below 1.00 for the test. The obtained significance value, 0.0019, is lower than 0.01 and thus considered highly significant, according to Chinchor (1992, Figure 3).

6 Conclusion

In this work, we pre-train from scratch a LAYOUTLM model using the DOCILE dataset. Importantly, we show that our model obtain better results on a novel domain-specific NER dataset. This shows that it is still possible to develop fast and state-of-the-art in-house models that allow commercial usage.

We also release our novel PAYSLIPS dataset that can be used to challenge document processing models in financial domains.

7 Limitations

The novel PAYSLIPS dataset is of small size compared to many standard benchmarks. Unfortunately, specialized domains like insurance not only induce expensive annotation costs, but it is also difficult to obtain authorization to publicly release the data. This issue is also common in other domains like biomedical NLP. Another issue is that PAYSLIPS is highly specialized, so interest may be limited.

Experimental results highlight that NLP models may not be useful for production yet, as the F1 scores are below 65.

Acknowledgement

We thank the anonymous reviewers for their comments and suggestions. This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015001).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

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.

Nancy Chinchor. 1992. The statistical significance of the MUC-4 results. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Nancy Chinchor, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: An analysis of the third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–450.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek

Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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.

Thibault Douzon, Stefan Duffner, Christophe Garcia, and Jérémy Espinas. 2023. Long-range transformer architectures for document understanding. In *Document Analysis and Recognition – ICDAR 2023 Workshops*, pages 47–64, Cham. Springer Nature Switzerland.

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)*.

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.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *Preprint*, arXiv:1602.02410.

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*, SIGIR '06, page 665–666, New York, NY, USA. Association for Computing Machinery.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural pre-training for form 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 6309–6318, Online. Association for Computational Linguistics.

Mahmoud Limam, Marwa Dhiaf, and Yousri Kessentini. 2023. Fatura dataset.

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. 2021. Skim-attention: Learning to focus via document layout. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2413–2427, Punta Cana, Dominican Republic. Association for Computational Linguistics.

E.W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Heidi Schmidt, Karen Butter, and Cynthia Rider. 2002. Building digital tobacco industry document libraries at the university of california, san francisco library/center for knowledge management. *D-Lib Magazine*, 8(9):1082–9873.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Holger Schwenk and Jean-Luc Gauvain. 2005. Training neural network language models on very large corpora. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 201–208, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

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. 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.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, 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 - ICDAR 2023*, pages 147–166, Cham. Springer Nature Switzerland.

A Statistical Significance

In the context of working with small test sets, it is important to validate that differences in experimental results are not attributable to randomness. To achieve this, (1) we run 100 times each fine-tuning experiment, using different random seeds for both data shuffling and initialization of the linear layer, and (2) we conduct statistical significance testing.

We follow the same procedure as the Message Understanding Conference (Chinchor, 1992; Chinchor et al., 1993) and rely on approximate randomization testing. This test is performed on a test set using two systems, A and B. For 9999 iterations, the test compares: (1) the difference in average F1-score between A and B on the test set with (2) the difference in average F1-score between two shuffled sets, each containing a mix of the F1-scores of A and B on the test set. The significance level is then computed as the percentage of iterations in which the difference in F1-score of the shuffled sets exceeds the actual difference in F1-score between A and B. The entire pseudo-code for this test is given in Algorithm 1.

Algorithm 1 Approximate Randomization testing

function AR(f_{baseline}, f_{proposed}, $\mathbf{x}_{\text{test}}$)
  Input: f_{baseline}, f_{proposed} : the models to compare and $\mathbf{x}_{\text{test}}$ the test set of size $\mathbf{N}$
  Output: α the significance value
  $\mathbf{y}_{\text{baseline}} \leftarrow \mathbf{N}$ predictions of the baseline model.
  $\mathbf{y}_{\text{proposed}} \leftarrow \mathbf{N}$ predictions of the proposed model.
  Compute the mean F1-score for each set of predictions: $\overline{\mathbf{F}_{\text{baseline}}}$ and $\overline{\mathbf{F}_{\text{proposed}}}$
  $\Delta F1_{\text{original}} = \left|\overline{\mathbf{F}_{\text{proposed}}} - \overline{\mathbf{F}_{\text{baseline}}}\right|$
  $n_{ge} \leftarrow 0$ ▷ Counter
  for $i \leftarrow 1$ to 9999 **do**
    $\mathbf{y} \leftarrow \mathbf{y}_{\text{baseline}} \cup \mathbf{y}_{\text{proposed}}$
    Shuffle $\mathbf{y}$
    Split $\mathbf{y}$ into two subsets $\mathbf{y}_A$ and $\mathbf{y}_B$, each of the same size.
    Compute the mean F1-score for each shuffled subset: $\overline{\mathbf{F}_A}$ and $\overline{\mathbf{F}_B}$
    $\Delta F1_{\text{shuffled}} = \left|\overline{\mathbf{F}_A} - \overline{\mathbf{F}_B}\right|$
    if $\Delta F1_{\text{shuffled}} \geq \Delta F1_{\text{original}}$ **then**
      $n_{ge} \leftarrow n_{ge} + 1$ ▷ Increment counter
  return $\alpha = \frac{n_{ge}+1}{9999+1}$ ▷ Significance level

B Precision and Recall

In addition to the F1-scores presented in Table 2, we provide a detailed precision and recall metrics in Table 4. We observe that on PAYSLIPS, the gain is mainly due to an increase in precision when pretraining on DOCILE. It is also interesting to note that when going from 12 to only 6 layers, the drop in performance is, again, due to a drop in precision.

C Datasets in Document Understanding tasks

In the field of Document Understanding, state-of-the-art models can experience a decline in performance when applied to domain-specific tasks compared to their results on standard benchmark datasets. Models like LAYOUTLM are typically evaluated on NER using datasets such as FUNSD, SROIE, and CORD. Table 5 highlights differences in size, types of categories to extract, and sparsity, which contribute to the complexity of domain specific NER. Firstly, document types vary significantly across datasets, impacting downstream task performance. Document analysis and receipt analysis are two very different tasks, and typically, F1 scores for SROIE and CORD tend to be higher (Xu et al., 2020, 2021; Huang et al., 2022; Wang et al., 2022, *inter alia*) than for FUNSD, DOCILE (labeled), or our newly introduced PAYSLIPS dataset. Secondly, sparse datasets, with fewer annotated entities, pose different challenges compared to non-sparse datasets with a higher density of annotations. In Table 5, we see datasets such as FUNSD and CORD, where each word belongs to a category, contrasted with datasets that focus on specific parts of the documents, and other words are categorized as OUTSIDE these entities of interest. Additionally, the primary entities vary notably between datasets with text heavy categories (e.g., FUNSD), and datasets of invoices and receipts that are filled with numbers. Specifically, in invoice-like documents such as DOCILE and PAYSLIPS, the complex and diverse layouts present challenges in understanding which amounts belong to which categories. In receipts, the amounts are often very close to an item name or a word directly describing the amount (e.g, 'total:'). This variation highlights two key points : the importance of efficiently leveraging layout information, and the different emphasis required on text understanding versus numerical understanding across datasets.

Numerical information emphasis can be ad-

Model	Pre-training dataset	Precision	Recall	F1
Fine-tuned on DOCILE **labeled**				
LAYOUTLM$_{\text{BASE}}$	IIT-CDIP	**57.79**	55.25	**58.35 ± 1.63**
LAYOUTLM$_{\text{BASE}}$	DOCILE	57.22	**59.45**	58.30 ± 1.52
LAYOUTLM$_{6\text{ LAYERS}}$	DOCILE	56.59	58.20	57.38 ± 1.38
LAYOUTLM$_{2\text{ LAYERS}}$	DOCILE	52.65	55.25	53.89 ± 1.03
LAYOUTLM$_{1\text{ LAYER}}$	DOCILE	49.71	52.68	51.12 ± 1.53
Fine-tuned on PAYSLIPS				
LAYOUTLM$_{\text{BASE}}$	IIT-CDIP	65.70	**59.80**	62.31 ± 5.13
LAYOUTLM$_{\text{BASE}}$	DOCILE	**71.47**	59.53	**64.74 ± 2.92**
LAYOUTLM$_{6\text{ LAYERS}}$	DOCILE	64.59	59.62	61.80 ± 3.12
LAYOUTLM$_{2\text{ LAYERS}}$	DOCILE	61.80	49.66	54.61 ± 3.71
LAYOUTLM$_{1\text{ LAYER}}$	DOCILE	51.66	40.29	45.08 ± 3.31

Table 4: Precision, Recall, and F1 scores for named-entity recognition using different pre-training and fine-tuning datasets. Results are averaged on 100 runs with different seeds.

dressed in the data used during pre-training. Most layout-aware encoder networks are pre-trained on IIT-CDIP, a collection of 40 million pages of documents from the Tobacco industry, published by UCSF. These documents, dating back to the 1990s, are primarily images with noise introduced during the scanning process, complicating the extraction of high-quality OCR outputs and potentially impacting model performance. In contrast, the DOCILE dataset consists mainly of electronic documents with minimal noise and highly legible text. Furthermore, DOCILE is composed exclusively of invoices, which are text-light and number-heavy, making it more suitable for financial domain-specific applications, whereas IIT-CDIP has more potential for training generalizable networks.

The size of the dataset also explains the continued use of IIT-CDIP for pre-training in the litterature (Xu et al., 2020; Huang et al., 2022; Bai et al., 2023, *inter alia*). With over ten times the volume of DOCILE, it remains a valuable resource for handling all kinds of documents.

D PAYSLIPS construction details

The PAYSLIPS dataset was obtained to automate the financial assessment at the claims and underwriting stages of a disability product. Accelerating this process allows underwriters and claims specialists to focus on less menial tasks while reducing the response time for a new policy or the payment of a claim. The underwriting specialists provided the Data Science team with an anonymous version of 611 pay statements. These documents were free of non-relevant Personal Identifiable Information (PII) such as names, addresses, ID numbers, and banking information. The raw data was then processed through an in-house OCR solution to obtain the text and layout of each page at the word level. An extensive annotation procedure was then initiated, during which several Data Scientists followed rules defined with the underwriters regarding the entities to extract. As such, only the amounts for the concerned period were annotated, as opposed to the year-to-date (YTD) amounts. Once the annotation procedure was completed, fine-tuning could be done on this data. The results presented in this paper are based on this version of the dataset. However, after discussions with SCOR's legal department, we could not share this version of the dataset as it still contained identifiable information about the company issuing the payments and the insured persons. To create a shareable version, we had to manually alter several amounts and the remaining sensitive information. The amounts were altered while ensuring the consistency and logical relationships between them, to preserve the coherence of the task.

Dataset	**Train / Val / Test size** (# pages)	**% of O** (word level)	**Document Type**	**Main entity types**
FUNSD (Jaume et al., 2019)	149 / - / 50	0	Forms	Text
SROIE (Huang et al., 2019)	626 / - / 347	83.82	Receipts	Text, Dates, Amounts
CORD (Park et al., 2019)	800 / 100 / 100	0	Receipts	Text, Dates, Amounts
DOCILE labeled (Šimsa et al., 2023)	6,759 / 635 / 1,000	89.46	Invoices	Text, Dates, Amounts
PAYSLIPS (ours)	485 / - / 126	94.95	Pay Statements	Dates, Amounts

Table 5: Overview of annotated datasets for named-entity recognition in documents. The percentage of O labels is calculated based on the combined train, validation and test sets, except for DOCILE labeled, where test annotations are unavailable, and the percentage is based on the train and validation sets.