DocLayLLM: An Efficient Multi-modal Extension of Large Language Models for Text-rich Document Understanding

Supplementary Material

A. Pre-traing tasks

We adopt pre-training tasks as shown in Table A. These tasks facilitate the alignment of layout and visual features with the LLM's feature space while enhancing the LLM's understanding of document content.

Task	Description
Document Description	Provide a brief overview of the document.
Text and Box Reconstruction	Recover the coordinates of bounding boxes of all the OCR text.
Layout Analysis	Determine the layout type (e.g., Title, Author, Paragraph, etc) of a giving area or locate specific layout elements.
Table Analysis	Decode the structure of tables and identify the positions of elements within.
Mask Language Model	Restore masked words in the provided OCR text.
Mask Position Model	Reconstruct the box for text elements missing the coordinates of the bounding box.
Geometric Analysis	Calculate distances or directions between two specified text elements.

Table A. The training tasks during the pre-training stage.

B. More Implementation Details

We implement our DocLayLLM using Llama2-7B-Chat [9] and Llama3-8B-Instruct [3]. The hyper-parameters for both pre-training and supervised fine-tuning are detailed in Table B. As shown in the table, our DocLayLLM demonstrates efficiency, requiring fewer training resources while maintaining high performance. This underscores the method's capability to deliver robust results without the need for extensive computational power, making it a resource-efficient solution for text-rich document understanding tasks.

Parameters	Pre-Training	Supervised Fine-Tuning
LoRA Rank	64	64
Batch Size	512	64
Max Length	2560	2560
Precision	bf16	bf16
Trainable Parameters	170M/Llama2; 178M/Llama3	170M/Llama2; 178M/Llama3
Fixed Parameters	6.7B/Llama2; 8.0B/Llama3	6.7B/Llama2; 8.0B/Llama3
Learning Rate	1e-4	2e-5
Weight Decay	0.01	0.01
Scheduler	cosine	cosine
Adam Betas	[0.9, 0.999]	[0.9, 0.999]
Adam Epsilon	1e-8	1e-8
Epoch	1	3

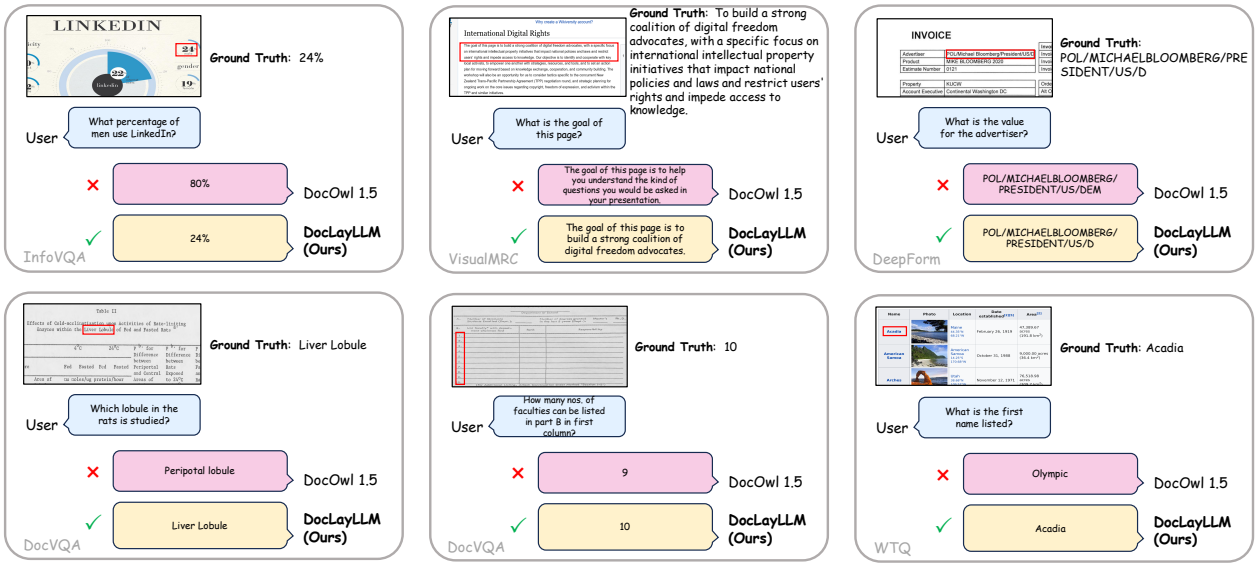
Table B. Hyper-parameters of DocLayLLM.

C. More Qualitative Examples

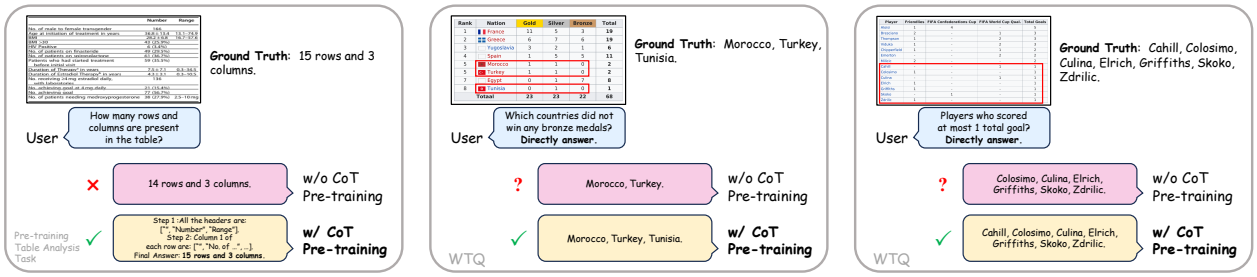
We also provide additional qualitative examples of our DocLayLLM. As shown in the comparison between DocLayLLM and the SOTA OCR-free method DocOwl 1.5 [2] in Figure Aa, our DocLayLLM demonstrates superior document understanding capabilities, delivering accurate answers in examples from InfoVQA and VisualMRC. Furthermore, in the DeepForm example, we observe that our design to integrate OCR information helps reduce the occurrence of hallucinated outputs compared to OCR-free methods. Moreover, in the examples from DocVQA and WTQ, DocLayLLM reliably exhibits robust table comprehension abilities. These results collectively highlight the effectiveness of our design in incorporating OCR information.

Furthermore, we also present the results of whether Table-Structure-Aware CoT is used during the pre-training stage. As shown in Figure Ab, models incorporating CoT demonstrate a more comprehensive understanding of table structures, leading to more accurate and thorough answers in table-related downstream document understanding tasks. This validates the effectiveness of our proposed CoT pre-training approach.

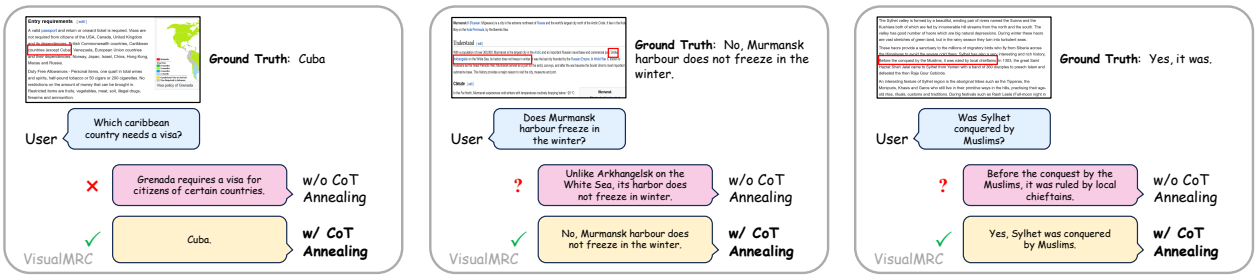
Additionally, we visualized the outputs with and without the use of CoT Annealing. As shown in the visualization of VisualMRC in Figure Ac, DocLayLLM employing CoT Annealing tends to provide more straightforward and accurate answers. This is particularly evident in yes-or-no questions, where the model without CoT Annealing often fails to directly respond with a clear "yes" or "no" but repeats the sentence in the document where the answer is located. In contrast, the model using CoT Annealing typically provides a direct answer first, followed by an explanation. These observations indicate that CoT Annealing enables the model to answer questions more directly, thereby enhancing its performance.



(a) Qualitative comparisons with DocOwl 1.5 [2] across various benchmarks. The document-oriented VQA tasks include InfoVQA [5], VisualMRC [8], and DocVQA [4]; the KIE task includes DeepForm [7]; and the Table Understanding task includes WTQ [6].



(b) Qualitative comparisons between the use and absence of CoT Pre-training. **w/o Pre-training** indicates the absence of CoT at the pre-training stage, while **w/ CoT Pre-training** denotes its application. **"?"** represents that the answer is ambiguous.



(c) Qualitative comparisons between the use and absence of CoT Annealing. **w/o CoT Annealing** indicates the absence of CoT Annealing, while **w/ CoT Annealing** denotes its application. **"?"** represents that the answer is ambiguous.

Figure A. Further qualitative comparisons of DocLayLLM against the SOTA OCR-free method and under various settings.

D. Input Length Analysis

In our ablation study, we evaluated the performance of different methods for incorporating OCR information. This section further examines the input length of OCR information under various approaches. The analysis was conducted using Llama3 [3] as the base model, with its tokenizer applied for tokenization. Table C presents a comparison of the average input length of OCR information across several benchmarks under two configurations: **(I)** encoding OCR bounding box coordinates as plain text, following the approach of ICL-D3IE [1], and **(II)** encoding OCR bounding box coordinates using a layout embedder LE.

The results clearly show that encoding with LE significantly reduces the input length, thereby enhancing effi-

ciency during both training and inference. These findings underscore the efficiency of our proposed DocLayLLM.

Input Method	Document-oriented VQA		KIE	
	DocVQA	VisualMRC	DeepForm	KLC
(I)	1571.80	6269.17	4952.87	457.58
(II)	**455.17**	**2095.35**	**1198.97**	**125.40**

Table C. The average input length of OCR information across various benchmarks, comparing different ways to input OCR bounding box coordinates.

E. OCR Result Impacts

Since DocLayLLM requires OCR result input, we explored the impact of OCR quality on the performance of DocLayLLM. In the results presented in the main text, we used the official OCR results when evaluating on the DocVQA benchmark. To assess the model's applicability in real-world scenarios, we employed a commercial OCR engine [1] to process DocVQA and used the recognized text for further testing DocLayLLM's performance. The results in Table D suggest that the reported results in the main text have not fully reflected the potential of DocLayLLM. The model could achieve even better performance with real-world OCR results.

Furthermore, as illustrated in Figure B, we observed that when OCR errors occur, DocLayLLM has the capability to correct these errors and produce the final correct answer. This further substantiates the robustness of DocLayLLM in real-world scenarios.

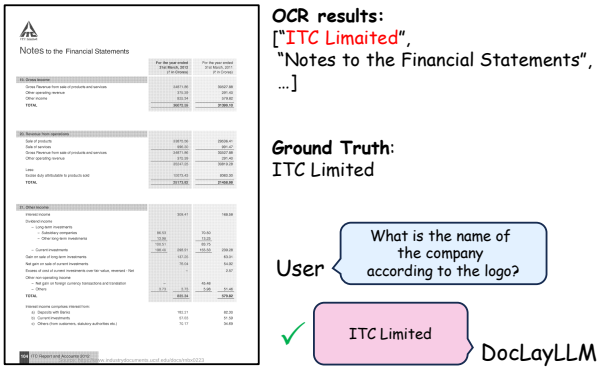


Figure B. Illustration of DocLayLLM's OCR error correction capability.

[1] https://www.textin.com/

	ANLS↑
Official OCR	86.52
Commercial Engine	**87.52**

Table D. Performances of DocLayLLM on DocVQA benchmark with different OCR results.

References

[1] Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. ICL-D3IE: In-Context Learning with Diverse Demonstrations Updating for Document Information Extraction. In *Proc. ICCV*, pages 19428–19437, 2023. 2

[2] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. In *Findings EMNLP*, pages 3096–3120, 2024. 1, 2

[3] Meta AI Llama Team. The Llama 3 Herd of Models. *arXiv: Comp. Res. Repository*, abs/2407.21783, 2024. 1, 2

[4] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In *Proc. WACV*, pages 2199–2208, 2021. 2

[5] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. In *Proc. WACV*, pages 1697–1706, 2022. 2

[6] Panupong Pasupat and Percy Liang. Compositional Semantic Parsing on Semi-Structured Tables. In *Proc. ACL*, pages 1470–1480, 2015. 2

[7] Stacey Svetlichnaya. DeepForm: Understand Structured Documents at Scale. https://wandb.ai/stacey/deepform_v1, 2020. Accessed: 2024-08-14. 2

[8] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. VisualMRC: Machine Reading Comprehension on Document Images. In *Proc. AAAI*, pages 13878–13888, 2021. 2

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv: Comp. Res. Repository*, abs/2307.09288, 2023. 1