Arctic-TILT
Business Document Understanding at Sub-Billion Scale

Łukasz Borchmann* Michał Pietruszka† Wojciech Jaśkowski Dawid Jurkiewicz
Piotr Halama Paweł Józiak‡ Łukasz Garncarek Paweł Liskowski
Karolina Szyndler Andrzej Gretkowski Julita Ołtusek Gabriela Nowakowska§
Artur Zawłocki Łukasz Duhr Paweł Dyda Michał Turski§

Snowflake AI Research

https://huggingface.co/Snowflake/snowflake-arctic-tilt-v1.3
https://github.com/Snowflake-Labs/arctic-tilt

§Adam Mickiewicz University †Jagiellonian University ‡Warsaw University of Technology

Abstract

The vast portion of workloads employing LLMs involves answering questions grounded on PDF or scanned content. We introduce the Arctic-TILT achieving accuracy on par with models 1000× its size on these use cases. It can be finetuned and deployed on a single 24GB GPU, lowering operational costs while processing rich documents with up to 400k tokens. The model establishes state-of-the-art results on seven diverse Document Understanding benchmarks, as well as provides reliable confidence scores and quick inference, essential for processing files in large-scale or time-sensitive enterprise environments. We release Arctic-TILT weights and an efficient vLLM-based implementation on a permissive license.

1 Introduction

General-purpose LLMs and their multi-modal counterparts provide a crucial advantage in process automation: they can be applied immediately, eliminating the expensive and time-consuming efforts of creating dedicated system architecture and model development. Though they are suitable choices for prototyping and building proof-of-concept solutions, once the case is validated, it becomes essential to consider the demands of real-world deployments, such as cost-efficiency (Fu et al., 2024; Ong et al., 2024), finetunability (Liu et al., 2022), and ensuring accurate confidence calibration (Van Landeghem, 2024).

We consider these issues in the context of Document Understanding (DU), where it is commonly required to integrate textual, layout and graphical clues to obtain the required information and introduce the Arctic-TILT, designed to address the needs of broad-use deployments, cost efficiency, and domain adaptations for a fraction of the cost of the leading models. The proposed solution appears competitive with orders of magnitude larger models on business and long document benchmarks.

2 Related Works

Traditionally, extracting tables or information from documents involved distinct steps like form recognition, field detection, and value extraction (Medvet et al., 2011; Rusiñol et al., 2013; Peanho et al., 2012; Tian et al., 2016; Le et al., 2019; Baek et al., 2019; Holt and Chisholm, 2018; Carbonell et al., 2019), each requiring separate models or heuristic pipelines. Later efforts moved towards more end-to-end graph-based methods (Liu et al., 2019; Hwang et al., 2021; Yu et al., 2021; Wang et al., 2024, *inter alia*). Recently, DU research closely paralleled advances in LLMs, converging on unified text-to-text formulations (Mathew et al., 2021b,a; Borchmann et al., 2021).

Despite being elegant, pure text-based methods

* See Appendix I for contributions.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 264–283
July 28-30, 2025

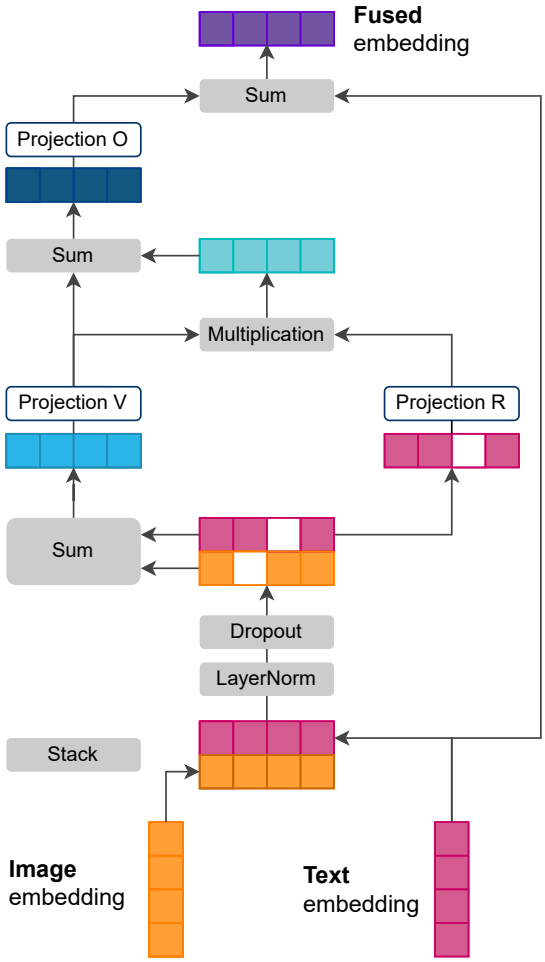


Figure 1: Our modality fusion. It can be seen as attention with role (Schlag et al., 2019) simplified as we calculate it over a pair of aligned text and image tokens.

fall short in layout-intensive tasks. This has led to the emergence of approaches extending LLMs with visual encoders (Li et al., 2023; Wu et al., 2023), layout modalities (Fujitake, 2024), or both (Mao et al., 2024; Li et al., 2024; Tang et al., 2023). Other research leverage multimodal instruction-following datasets (Dai et al., 2023; Zhang et al., 2023; Ye et al., 2023b, *inter alia*) or introduce auxiliary objectives like text-image matching (Peng et al., 2022; Tang et al., 2023; Xu et al., 2020; Bai et al., 2022; Feng et al., 2024). Finally, some works approach DU using vision-only models (Kim et al., 2021, 2022; Lee et al., 2023a; Beyer et al., 2024).

The key dimension involves balancing model performance and deployment constraints. We advocate for lightweight DU models due to their superior memory efficiency and inference speed—crucial for practical or edge deployments—aligned with prior work emphasizing cost-effectiveness (Fu et al., 2024; Zhao et al., 2024; Ong et al., 2024).

TILT	Arctic-TILT
Vision Encoding and its Fusion with Text	
sum of text & image	fusion by tensor product
first layer only	every encoder layer
Pretraining and finetuning	
400k steps of adaptation	900k steps
SFT on 4 datasets	SFT on 17 datasets
Transformer	
dense attention, vanilla	sparse attention, SLED
max 9k tokens	max 400k tokens
basic optimization	heavy optimization
Licensing and availability	
closed, proprietary	open source

Table 1: Comparison of TILT and Arctic-TILT.

3 Arctic-TILT

We build on the TILT encoder-decoder model, which extends T5 (Raffel et al., 2020) by incorporating (1) an attention bias based on horizontal and vertical distances and (2) image embeddings capturing token visual neighborhood (Powalski et al., 2021). To overcome its limitations, we introduce novel modality fusion, attention sparsity, training recipe, and optimized training/inference. The improved model is referred to as Arctic-TILT (see Table 1).

3.1 Fusion of Text and Vision

TILT integrates visual and textual semantics by summing word embeddings with RoI-pooled bounding box representations, using a U-Net-based image encoder. Features are fused once, immediately after embedding. However, ablations from Powalski et al. (2021) indicate that TILT's visual backbone contributes less to performance than layout features, suggesting that single-step summation loses critical visual details. We attribute this to (a) the long backpropagation path weakening visual gradients and (b) summation failing to capture higher-order text-spatial interactions. To address this, we replace TILT's one-time fusion with a layer-wise fusion mechanism using tensor product representations. This approach enables progressive interaction between modalities in each encoder block, with gating elements reducing noise.

Fusion by Tensor Product. Specifically, we opt for the fusion of modalities inspired by approximation of tensor product representations (Smolensky, 1990; Schmidhuber, 1993; Schlag et al., 2019). Given the text and image embeddings $t, i \in \mathbb{R}^d$, we

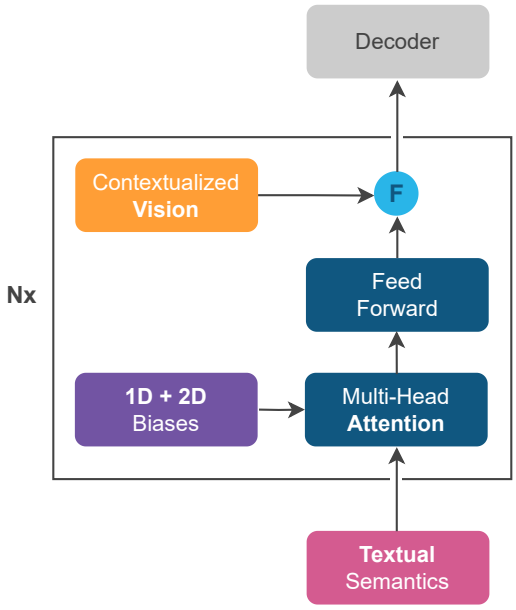


Figure 2: The Arctic-TILT encoder block combines *Vision* from U-Net and *Textual Semantics* from input embeddings through *Fusion* (F) operation. The *Multi-Head Attention* is augmented with *1D and 2D positional biases*. This procedure is repeated in each layer (*Nx*).

calculate the fused embedding with: $\text{Fuse}(t, i) = O(V(t + i) \odot (\mathbf{1} + Rt)) + t$ where V, R, and O are $\mathbb{R}^{d \times d}$ trainable parameters. In practice, we use a variant of this mechanism with layer norm and dropout (Figure 1 and Listing 1).

Placement. We found that placing the fusion module after FFNs (Figure 2) is most beneficial. Additionally, by applying it after every encoder layer, we mitigate the vanishing gradient effect and enable the model to focus on different visual features as its comprehension of the document improves.

3.2 Long Context Support

Concerning the product-oriented nature of our work, it is essential to cover a significant fraction of real-world documents of potentially arbitrary lengths while operating within limited resources. The outlined optimizations are guided by the need to handle as much context as possible on widely available A10 and L4 GPUs equipped with 24GB vRAM. We assume a single-GPU setup and measure the impact of applied techniques and architectural changes on the maximum context length used during the finetuning and inference.

Chunked processing. To address the quadratic complexity of encoder self-attention, we employ a variant of fusion-in-decoder/SLED (de Jong et al., 2023; Pietruszka et al., 2022; Ivgi et al., 2022), using zero chunk padding. This approach restricts encoder attention to a bounded-width neighborhood around its diagonal, forming a block diagonal matrix and thus linearly reducing attention weights relative to sequence length (see Appendix E).

Nested stack checkpointing. Applying gradient checkpointing across the entire 24-layer encoder stack substantially reduces memory requirements, storing activations only for the final layer needed by the decoder. This decreases memory usage dramatically—for example, from 96GB to just 4GB when processing 1M tokens—at the cost of an extra encoder forward pass.

Random chunks. Concatenated chunk embeddings may still exceed memory limits in the decoder cross-attention. Although the model supports 230k tokens during training, we further extended this by randomly discarding chunks, allowing exposure to different document parts over epochs.

Beyond primary techniques, we apply additional optimizations. Mixed-precision training with *bfloat16* and disabled weight caching reduce RAM usage, doubling inference input length. Recomputing decoder projections per layer instead of caching key-value pairs extends inference context to 389k tokens. Offloading decoder activations from GPU to CPU minimizes peak GPU memory at the cost of increased processing time. Lastly, memory-efficient attention reduces attention overhead (Rabe and Staats, 2022).

Ultimately, our optimizations culminate in significant memory usage improvements, allowing us to effectively train and deploy Arctic-TILT for documents up to 500 pages[1] on a single 24GB GPU. The step-by-step summary is studied in Table 2.

3.3 Pretraining and finetuning

The training process began with a self-supervised pretraining from the T5 large model (Raffel et al., 2020). Following the introduction of TILT architecture changes, which included U-Net (Ronneberger et al., 2015) and 2D biases, as well as text-vision post-fusion, the model underwent further self-supervised pretraining for a total of 900k steps based on documents from the CCpdf (Turski

[1]Specifically, 390k input tokens with an output of 128 tokens, corresponding to 780 tokens per page on average.

	Inference	Training
Vanilla TILT	9k	4k
+ attention sparsity	87k	41k
+ mixed precision	179k	51k
+ memory efficient attention	183k	56k
Inference-only optimizations		
+ no cross-attention KV cache	389k	
Training-only optimizations		
+ nested checkpointing		230k
+ CPU offloading		256k
+ random chunks		389k

Table 2: Max input length (tokens) consumed during training and inference given single 24GB GPU. Tested for documents up to 500 pages (389k tokens).

et al., 2022) and OCR-IDL (Biten et al., 2022).

Finally, the model was finetuned on QA and KIE datasets. In this phase, we increase the number of supervised datasets to 17, compared to TILT's original choice of four. The datasets chosen represent critical aspects of DU tasks, including, but not limited to, forms, financial reports, charts, invoices, insurance documents, contracts, and legal documents (detailed in Appendix D).

The model features 822M parameters and the total computational cost of its training is slightly less than 10 days on 8xH100 GPUs.

4 Experiments

We evaluate our approach across multiple DU benchmarks spanning diverse tasks and document types. DocVQA (Mathew et al., 2021b) assesses systems on QA over scanned documents, while SlideVQA (Tanaka et al., 2023) addresses challenges in densely packed presentation slides. MMLongBench-Doc (Ma et al., 2024) targets extensive multi-page documents. Kleister NDA (Stanisławek et al., 2021) emphasizes precise legal-domain information extraction, whereas Kleister Charity and VQA-CD (Mahamoud et al., 2022) focus respectively on financial reports and corporate purchase documents. InfographicsVQA (Mathew et al., 2021a) highlighting multimodal reasoning. Finally, we also include long-context summarization tasks from PubMed-Lay and ArXiv-Lay (Nguyen et al., 2023). Across these datasets, input sizes, domain coverage, and visual complexity vary significantly—from single-page invoices or forms to multi-page legal contracts.

Comparison in Table 3 include only models previously recognized for achieving SOTA performance in their respective settings, alongside Arctic-TILT results presented in Generalist and Specialist variants. All baseline scores are sourced from third-party publications claiming superior performance over previous models. See Appendix G for comparisons with additional open-source models.

4.1 Document Visual QA and KIE

Zero-shot Performance. As shown in Table 3, our model evaluated in the zero-shot setting often achieves near-SOTA performance out of the box (e.g., on VQA-CD, DUDE and Kleister Charity). However, on Kleister NDA—where the questions are more complex—its performance is less competitive. On the recently introduced MMLongBench-Doc (Ma et al., 2024), which evaluates zero-shot performance on documents up to 400 pages, we exceed several much larger LLMs and LVLMs (e.g., Mixtral 8x7B, QWen-Plus, Claude-3 Opus, InternVL) by substantial margins. Models such as Gemini 1.5 Pro and GPT-4o do outperform us, but they reportedly contain hundreds of times more parameters (full results in Table 8). Section 4.3 explores how our model's performance improves with limited annotated data, comparing it to GPT-4o.

Multi-page. Among six multi-page QA/KIE datasets, we achieve new SOTA results on four (MP-DocVQA, Kleister Charity, Kleister NDA, and DUDE), outperforming larger general-purpose LLMs such as GPT-4 Family (Vision Turbo and Omnia) and specialized DU models like ERNIE-Layout, LAMBERT, GRAM, and BigBird-Pegasus+Layout. We attribute these gains to our explicit modeling of long-context interactions.

Three of these datasets include labeled positions of the target answers, allowing us to analyze performance based on where the relevant information appears in each document. Figure 3 shows a *primacy bias*, with higher accuracy when the key text occurs near the beginning of the input (Liu et al., 2024a). Overall, considering the input sequence length in tokens, Arctic-TILT sets new SOTA on four out of the six longest datasets in Table 3, demonstrating particular strength on multipage inputs where many existing DU models struggle.

Single-page. In settings with single-page excerpts or standalone images (shorter inputs), our model still performs strongly. It surpasses TILT by 2 points on the DocVQA dataset (Mathew et al., 2021b) and also outperforms GPT-4 Vision. Notably, Arctic-TILT achieves state-of-the-art re-

Dataset	Industrial	Multipage	State-of-the-Art (Params, Score)			Our (Specialist, Generalist)	
MP-DocVQA	✓	✓	GRAM	859M	80.3	**81.2**	76.9
Kleister Charity	✓	✓	LAMBERT	125M	83.6	**88.1**	86.9
Kleister NDA	✓	✓	ERNIE-Layout	355M	88.1	**94.3**	38.3
DUDE	✓/✗	✓	GPT-4Vt + OCR	200B+	53.9†	**58.1**	55.9
MMLongBench-Doc	✓/✗	✓	GPT-4o	200B+	**42.8**†	—	25.8
SlideVQA	✗	✓	GPT-4Vt + OCR	200B+	**57.3**†	55.1	40.4
ArXiv-Lay	✗	✓	BigBird...+Layout	581M	41.2	**44.4**	—
PubMed-Lay	✗	✓	BigBird...+Layout	581M	42.1	**44.8**	—
DocVQA	✓	✗	InternVL 2.0 Pro	108B+	**95.1**†	90.2	88.6
VQA-CD	✓	✗	QALayout	8M	42.5	**90.7**	88.7
InfographicsVQA	✗	✗	InternVL 2.0 Pro	108B+	**86.8**†	–	57.0

Table 3: Arctic-TILT (822M params) compared to the previous state-of-the-art. Our model remains competitive and excels when input is a long, business document. Original metrics used for each dataset; † denotes generalist score.

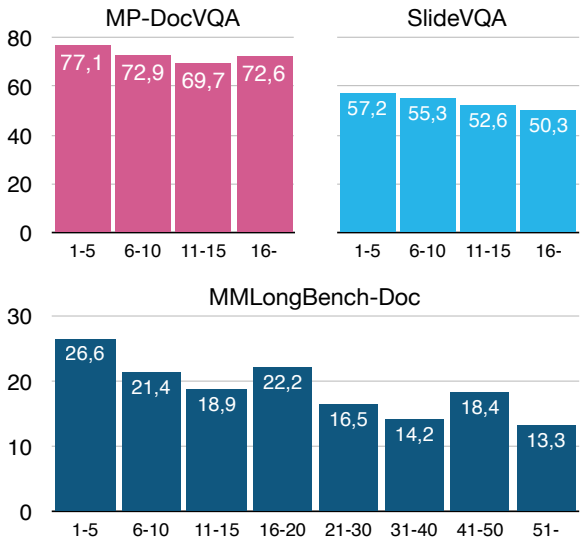


Figure 3: Scores depending on the evidence location.

sults on the newly introduced VQA-CD dataset (Souleiman Mahamoud et al., 2022), which includes invoices and purchase orders. Although we observe a gap between Arctic-TILT and InternVL 2.0 Pro (108B+ parameters) on certain benchmarks like InfographicsVQA (Mathew et al., 2021a), the model's overall performance in single-page tasks remains competitive. We attribute some limitations of Arctic-TILT to the varied aspect ratios and unusual formats in these tasks, making it challenging for our comparatively small, 8M-parameter visual backbone to encode every layout robustly.

Strengths and Weaknesses. Qualitative analysis using the DUDE diagnostic subset (see Appendix G) reveals that Arctic-TILT outperforms other state-of-the-art (SOTA) models on both abstractive and extractive questions, while ranking second-best for list-based and unanswerable queries. This suggests robust handling of complex data but also indicates potential areas for improvement on less typical answer types, which could be addressed by adjusting the supervised fine-tuning (SFT) data mix. On the SlideVQA dataset (Tanaka et al., 2023) our model achieves a score 2 points lower than GPT-4 Vision. We attribute this shortfall to the predominantly horizontal format of slides—a layout not specifically targeted in our current pretraining mix.

4.2 Layout-Aware Summarization

To complement our VQA and KIE results, we also investigate Arctic-TILT's capacity for capturing layout information and long-range dependencies in the LoRaLay collection of summarization tasks. Unlike most other summarization benchmarks, LoRaLay includes scientific documents with rich structure rather than simple text blocks (Nguyen et al., 2023). As shown in Table 3, Arctic-TILT outperforms the previous SOTA on both ArXiv-Lay and PubMed-Lay by several points. Notably, this is achieved without any specialized pretraining objectives tailored to summarization, underscoring our model's general ability to handle complex, layout-intensive inputs.

4.3 Adapting to Novel Use Cases

Arctic-TILT introduces optimizations to enhance training under minimal memory constraints, improving adaptability in production settings for out-of-domain examples and novel use cases. Thus, we evaluate its zero-shot accuracy improvement when finetuned on up to 25 annotated documents from holdout datasets, including Ghega patents (Medvet et al., 2011) and a private payment stub dataset (see Appendix F). As shown in Figure 4, Arctic-TILT rapidly approaches GPT-4o's accuracy with just five examples and surpasses it with slightly more.

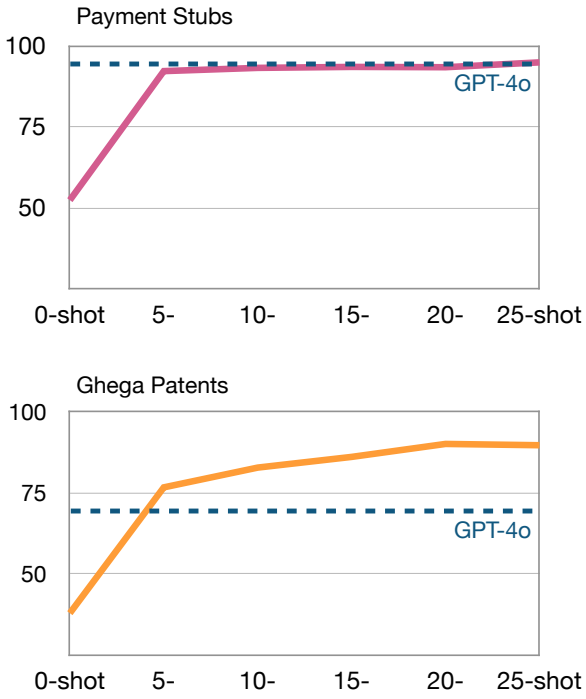


Figure 4: Improvement of Arctic-TILT zero-shot accuracy given finetuning on up to 25 annotated documents.

These results highlight the advantages of specialized, smaller LLMs over general-purpose models, emphasizing cost-effectiveness and adaptability.

4.4 Confidence Calibration

Following van Landeghem et al. (2023), we evaluate *Expected Calibration Error* (ECE) and *Area Under the Risk-Coverage Curve* (AURC) on the DUDE dataset. Confidence scores are derived from per-token lists, where we use the minimum score instead of the geometric mean, as it proved empirically superior. Results show exceptional calibration, with an SOTA ECE of 7.6 (previous best: 19.0), indicating strong alignment between confidence and accuracy. Our AURC of 25.3 (previous best: 44.0) further demonstrates effective uncertainty estimation, allowing for appropriate low-confidence assignments to ambiguous predictions requiring human review. Beyond DUDE, we analyze 18k samples from 14 datasets (Figure 5). The results confirm consistently low ECE and well-calibrated confidence scores, as accuracy follows the diagonal $y = x$ in the calibration plot.

4.5 Computational Efficiency

The imperative for businesses to rapidly and efficiently process substantial document volumes calls for models that maximize throughput and

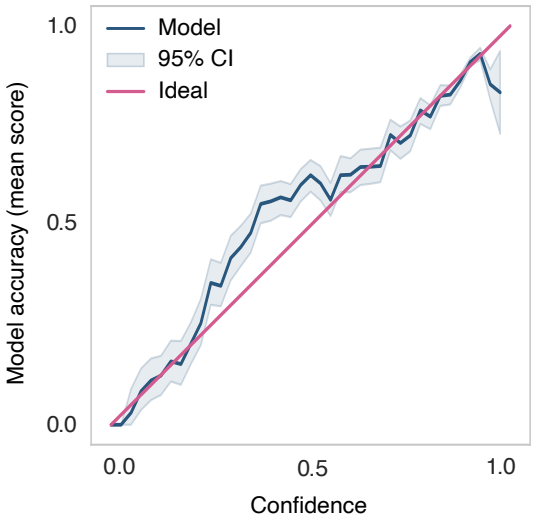


Figure 5: Arctic-TILT calibration.

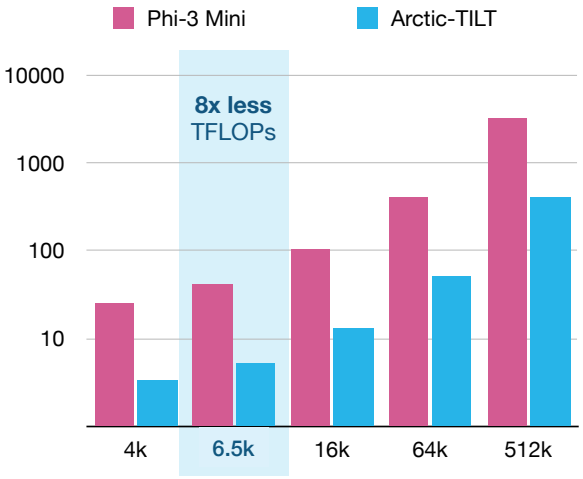


Figure 6: Arctic-TILT's computational efficiency (TFLOPs, lower is better) compared to Phi-3 Mini on VQA/KIE given inputs ranging from 4k to 512k tokens.

operational efficiency. To address this aspect of the model, we analyze the inference floating point operations per second (TFLOP) required for Arctic-TILT compared to Phi-3 Mini (Abdin et al., 2024), an example of a decoder-only model featuring 3.8B parameters and optimized by resorting to the attention sliding window. The latter was selected as a well-known reference model concerning the limited memory and compute regime we aim at, though it is not capable of achieving satisfactory accuracy on Document Understanding tasks.

Results presented in Figure 6 indicate that Arctic-TILT consistently demands lower TFLOP across all context lengths for our primary use case of VQA/KIE,[2] reflecting its smaller parameter size.

[2]We assume the output of 8 tokens—longer than the aver-

Importantly, concerning the input of 6.5k tokens, the mean input length for VQA/KIE tasks considered before, we require 8× less operations.

4.6 Ablation Study

We systematically alter one of four key differences between Arctic-TILT and Vanilla TILT to evaluate their individual contributions (Table 4).

Concerning the fusion positioning with respect to the multi-head attention (MHA) and the fusion mechanism used, results suggest that the approach from Section 3.1 is optimal. Replacing fusion by TP with Vanilla TILT fusion (*original fusion*) leads to a loss of 1.6 points on average. Similarly, placing fusion after the MHA and before the FFN (*our but pre-fusion*) is worse than placing it before the MHA (*Arctic-TILT*) by 1.5 points. We see that employing sparsity with blocks of 1024 tokens with no overlap (Section 3.2) outperforms alternative variants. Specifically, Vanilla TILT (*original, dense*) cannot consume the entire content of some lengthy documents, leading to the loss of 15 points. Similarly, varying block sizes between 1024 and 2048 tokens that either overlap with 128 tokens or have no overlap, we see that they lead to the loss of at least 1.7 points on average. Analysing the impact of additional self-supervised pretraining introduced in Arctic-TILT (Section 3.3), we see they offer an advantage of 4.6 points on average, indicating that the Vanilla TILT (*original pretraining*) was undertrained. Finally, change in the introduced supervised finetuning data (Section 3.3) markedly enhanced the model's performance across all evaluated tasks.

Overall, we found that any deviation from the proposed setup leads to the degradation of scores obtained by the model on downstream tasks.

	Charity	DUDE	MP-	NDA	Δ
Arctic-TILT	82.7	44.6	76.7	72.1	–
original fusion	80.3	43.4	76.6	69.5	-1.6
our but pre-fusion	79.4	44.2	77.3	69.2	-1.5
original, dense	55.0	38.6	66.1	56.6	-15.0
1024/128 sparsity	78.6	43.3	75.7	68.2	-2.6
2048/0 sparsity	79.0	43.6	76.9	69.6	-1.9
2048/128 sparsity	80.3	43.7	76.4	69.0	-1.7
original pretraining	79.1	43.6	75.0	69.0	-4.6
original SFT data	40.2	37.1	73.5	17.0	-27.1

Table 4: Results of the ablation study (0-shot ANLS).

5 Summary

We have introduced the Arctic-TILT model, which addresses TILT's limitations in handling multimodal input, suboptimal training procedure, and maximum context length. By analyzing the results and considering the cost-efficiency of the designed solution, we provided practical insights into designing capable, lightweight models for the industry.

Arctic-TILT demonstrates state-of-the-art or competitive performance across seven diverse benchmarks, often outperforming models significantly larger, especially on long, business-centric documents. Our work illustrates that strategic design and optimization can rival the capabilities of larger, more resource-intensive models.

Importantly, Snowflake is releasing the Arctic-TILT model weights and an efficient vLLM-based implementation to the public, enabling broader access and application of this cost-effective and high-performance solution for Document AI.

Acknowledgements

We extend our heartfelt thanks to Tomasz Dwojak and Daniel Campos, whose feedback on the manuscript and valuable suggestions have greatly enhanced the quality of this work. We are also grateful to Łukasz Słabinski, Michał Gdak, Tomasz Stanisławek, Nikolai Scholz, and Vivek Raghunathan; their support and guidance as managers were instrumental throughout this research. To conclude, we thank Rafał Kobiela for his assistance with the cloud infrastructure, and Staszek Pasko for coordinating the open source process and discussing the product aspects of this project.

References

Marah Abdin et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question an-

age target length of evaluation datasets from Section 4.1.

swering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection.

Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, et al. 2022. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. *arXiv preprint arXiv:2212.09621*.

Lucas Beyer et al. 2024. PaliGemma: A versatile 3B VLM for transfer.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering.

Ali Furkan Biten, Rubèn Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. 2022. OCR-IDL: OCR Annotations for Industry Document Library Dataset.

Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. 2024. GRAM: Global Reasoning for Multi-Page VQA.

Łukasz Borchmann. 2024. Notes on Applicability of GPT-4 to Document Understanding.

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. DUE: End-to-end Document Understanding Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Manuel Carbonell, Alicia Forn'es, Mauricio Villegas, and Josep Llad'os. 2019. TreyNet: A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages. *ArXiv*, abs/1912.10016.

Zhe Chen et al. 2024. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites.

Aakanksha Chowdhery et al. 2022. PaLM: Scaling Language Modeling with Pathways.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Kenny Davila, Fei Xu, Saleem Ahmed, David A. Mendoza, Srirangaraj Setlur, and Venu Govindaraju. 2022. ICPR 2022: Challenge on Harvesting Raw Tables from Infographics (CHART-Infographics). In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4995–5001.

Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end Document Recognition and Understanding with Dessurt.

Michiel de Jong, Yury Zemlyanskiy, Joshua Ainslie, Nicholas FitzGerald, Sumit Sanghai, Fei Sha, and William Cohen. 2023. FiDO: Fusion-in-Decoder optimized for stronger performance and faster inference.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. 2024. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization?

Masato Fujitake. 2024. LayoutLLM: Large Language Model Instruction Tuning for Visually Rich Document Understanding. In *International Conference on Language Resources and Evaluation*.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. *LAMBERT: Layout-Aware Language Modeling for Information Extraction*, page 532–547. Springer International Publishing.

Xavier Holt and Andrew Chisholm. 2018. Extracting Structured Data from Invoices. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 53–59, Dunedin, New Zealand.

Anwen Hu, Haiyang Xu, Jiabo Ye, Mingshi Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *ArXiv*, abs/2403.12895.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Mingshi Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *ArXiv*, abs/2409.03420.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial Dependency Parsing for Semi-Structured Document Information Extraction.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2022. Efficient Long-Text Understanding with Short-Text Models.

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.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *European Conference on Computer Vision (ECCV)*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut: Document Understanding Transformer without OCR. *CoRR*, abs/2111.15664.

Anh Duc Le, Dung Van Pham, and Tuan Anh Nguyen. 2019. Deep Learning Approach for Receipt Recognition.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023a. Pix2struct: Screenshot parsing as pretraining for visual language understanding.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. *ArXiv*, abs/2210.03347.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023b. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2024. Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *ArXiv*, abs/2403.04473.

Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, and Jingdong Wang. 2024. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *ArXiv*, abs/2405.21013.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations.

Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. 2022. QAlayout: Question Answering Layout Based on Multimodal Attention for Visual Question Answering on Corporate Document. In *Document Analysis Systems*, pages 659–673, Cham. Springer International Publishing.

Zhiming Mao, Haoli Bai, Lu Hou, Jiansheng Wei, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Visually Guided Generative Text-Layout Pre-training for Document Intelligence. *ArXiv*, abs/2403.16516.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. InfographicVQA.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. DocVQA: A Dataset for VQA on Document Images.

Eric Medvet, Alberto Bartoli, and Giorgio Davanzo. 2011. A probabilistic approach to printed document understanding. *International Journal on Document Analysis and Recognition (IJDAR)*, 14:335–347.

Laura Nguyen, Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2023. LoRaLay: A Multilingual and Multimodal Dataset for Long Range and Layout-Aware Summarization.

Jordy van Landeghem. 2024. *Intelligent Automation for AI-driven Document Understanding*. Ph.D. thesis, KU Leuven.

Jordy van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. 2023. Document Understanding Dataset and Evaluation (DUDE).

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. RouteLLM: Learning to Route LLMs with Preference Data.

Claudio Antonio Peanho, Henrique Stagni, and Flavio Soares Correa da Silva. 2012. Semantic information extraction from images of complex documents. *Applied Intelligence*, 37:543–557.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-Layout: Layout Knowledge Enhanced Pretraining for Visually-rich Document Understanding.

Michał Pietruszka, Łukasz Borchmann, and Łukasz Garncarek. 2022. Sparsifying Transformer Models with Trainable Representation Pooling.

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.

Markus N. Rabe and Charles Staats. 2022. Self-attention Does Not Need $O(n^2)$ Memory.

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. *JMRL*.

Sachin Raja, Ajoy Mondal, and C. V. Jawahar. 2023. ICDAR 2023 Competition on Visual Question Answering on Business Document Images. In *Document Analysis and Recognition - ICDAR 2023*, pages 454–470, Cham. Springer Nature Switzerland.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Marçal Rusiñol, Tayeb Benkhelfallah, and Vincent Poulain dAndecy. 2013. Field Extraction from Administrative Documents by Incremental Structural Templates. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1100–1104.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. 2019. Enhancing the Transformer with Explicit Relational Encoding for Math Problem Solving.

J. Schmidhuber. 1993. Reducing the Ratio Between Learning Complexity and Number of Time Varying Variables in Fully Recurrent Nets. In *ICANN '93*, pages 460–463, London. Springer London.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216.

Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. 2022. QAlayout: Question Answering Layout based on multimodal Attention for visual question answering on corporate Document. *Acoustics Research Letters Online*.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts. In *Document Analysis and Recognition – ICDAR 2021*, pages 564–579, Cham. Springer International Publishing.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying Language Learning Paradigms.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting Text in Natural Image with Connectionist Text Proposal Network.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for Multi-Page DocVQA.

Michał Turski, Tomasz Stanisławek, Filip Graliński, and Karol Kaczmarek. 2022. Building the High Quality Corpus for Visually Rich Documents from Web Crawl Data.

Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kang Gu, and Sameena Shah. 2024. DocGraphLM: Documental Graph Language Model for Information Extraction.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2024. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A Limitations

Although our approach demonstrates state-of-the-art performance on a range of document-related tasks, it is primarily tailored for unstructured or semi-structured Document Understanding. This focus imposes limitations when applied to non-DU tasks such as Scene Text VQA (Biten et al., 2019), where text may appear in complex outdoor scenes with highly variable lighting, orientation, and font usage. Likewise, because of relying exclusively on MLM pretraining and lightweight visual encoder, Arctic-TILT may struggle with VQA assuming the dominant image component (Antol et al., 2015). Next, because of the SFT datasets' composition and compact model size, it cannot follow complex instructions, and its intended use is limited to QA and summarization tasks. Finally, as discussed earlier and visible in Figure 3, accuracy can suffer if the key answer appears very late in the document. Example failure cases illustrating these limitations are provided in Appendix H.

B Contribution

We have:

- introduced the Arctic-TILT model, which addresses TILT's limitations in handling multimodal input, suboptimal training procedure, and maximum context length;
- established state-of-the-art performance on seven benchmarks demanding text, vision, and layout comprehension;
- demonstrated that within the industrial applications setting and while keeping the parameter count below 1B, one could achieve performance better or comparable to vastly larger models;
- presented a novel modality fusion mechanism inspired by tensor product representations, and have shown how effectively apply it across the transformer encoder;
- demonstrated how, with well-designed attention sparsity patterns and numerous other optimizations, consume extensive input sequences during training and inference, given a single cost-efficient GPU, while maintaining competitive accuracy of the model;
- demonstrated that all of the architectural decisions can be drawn from the systematic ablation study we conducted;
- provided insights that can be applied to design future generations of multimodal models, particularly for visually rich document processing.

Our work illustrates that strategic design and optimization can rival the capabilities of larger, more resource-intensive models.

C Why TILT as a Starting Point?

We argue that the effectiveness of the DU model depends primarily on its ability to understand specific document formats and structures in the most

document-native way possible, which can only be guaranteed by equipping the model with layout-aware architectural biases as early as possible.

Though a number of large vision-only models have been proposed (Kim et al., 2022; Davis et al., 2022; Lee et al., 2022), smaller models with an explicit OCR step still outperform them. Notably, even GPT-4 Vision benefits from the availability of OCR-recognized text (Borchmann, 2024). Although document intelligence requires visual features (e.g., to recognize checkboxes, signatures, text colors, and formatting), the text and its spatial arrangement are most important. This necessitates models with heavy textual and lightweight visual encoders, such as TILT.

Secondly, the imperative for businesses to rapidly and efficiently process substantial document volumes calls for models that maximize throughput while also maximizing operational efficiency. Smaller, specialized models, tailored for such tasks, often surpass their larger LLM counterparts, which struggle to meet these criteria due to their higher computational demands and processing times. The motivation for these is not only practical, as regulations such as GDPR, CCPA, or Chinese digital laws may require specific types of information to be processed locally. This need is fulfilled with smaller, specialized models that can be deployed on broadly available GPUs and thus are not restricted to a handful of regions.

The original TILT offers impressive performance despite keeping the number of parameters below 1B because of a well-balanced parameter budget and relying on encoder-decoder architecture, which, despite lower popularity compared to decoder-only models, offers better quality in compute-matched setups (Raffel et al., 2020; Chowdhery et al., 2022; Wang et al., 2022; Tay et al., 2023). Besides, we prefer them because achieving optimal attention sparsity patterns is more straightforward with separate encoder and decoder modules.

The encoder-decoder model with a sizeable textual backbone and small visual encoder, equipped with layout architectural bias that has previously established state-of-the-art results, appears a viable starting point for building a modern DU system.

D Datasets for Supervised Finetuning

Training of Arctic-TILT included SFT phase on twelve publicly available and five in-house annotated datasets. The first group included Kleister Charity, Kleister NDA (Stanisławek et al., 2021), CHART-Infographics (Davila et al., 2022), DeepForm★ (Borchmann et al., 2021), DocVQA (Mathew et al., 2021b), DUDE (Van Landeghem et al., 2023), FUNSD (Jaume et al., 2019), InfographicVQA (Mathew et al., 2021a), SQuAD 2.0 (Rajpurkar et al., 2018), TAT-DQA (Zhu et al., 2022), VQA-CD (Mahamoud et al., 2022), and VQAonBD (Raja et al., 2023).

Private datasets were based on QA annotations of IRS990 forms, insurance reports, company annual reports, synthetic invoices, and charity annual reports. To give the research community a grasp on the characteristics of this collection, we provide the most important statistics and examples of questions in Figure 7 and Table **??**.

E Used Hyperparameters

Chunking setup. Given hyperparameters—core chunk length c, overlap size o, and prefix length l—the input of length $C = n \cdot c$ is divided as follows: chunk 1 contains prefix tokens followed by input tokens $0, \ldots, c - l$. Subsequent chunks $i + 1$ start with prefix tokens followed by tokens $t - o + 1, \ldots, t - o + c - l$, where chunk i used tokens up to position t. We studied the size of the attention block, as well as the overlap size of consecutive blocks. To our surprise, the best setup for inference was 1024 tokens attention size with no overlap, and these conclusions are independent of the setup overlap/attention size during training. The abstract illustration of this concept is present in Figure 9.

Learning rate scheduling and precision. We observed a non-trivial inference between the two hyper-parameters. Compared to *fp32* pretraining, *bf16* pretraining with more aggresive learning rate scheduling was able to catch up, and with same learning rate scheduling was observably worse. We ended up with using *cosine_luh* scheduler with 1% training steps with constant learning rate of 1e-3 (warm-up), followed by 89% training steps with linear decay down to 2e-4, followed by cosine scheduling for the remaining 10% steps decaying to 5e-5. Same observations were drawn during finetuning.

Training protocol. The finetuning phase's hyperparameters are set as 100k steps at batch size 128 with the *AdamWScale* optimizer. We set loss

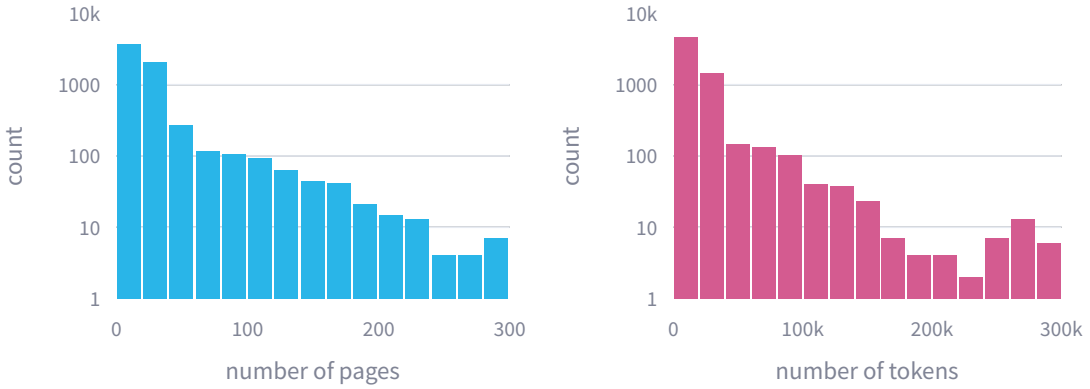


Figure 7: Lengths of documents included in five private datasets (number of tokens and pages).

Dataset	Sample Questions	Documents	Annotations
IRS990	What is the percentage of public support in the year of the report? What is the sum of the total liabilities in US dollars? What is the Employer Identification Number?	3,097	38,025
Insurance Reports	What was the value of total premiums written in the Surplus & Self-Procured category in 2016? Who is the director of the Alaska Division of Insurance? For whose contributions were tax credits claimed in 2015?	50	1,702
Company Annual Reports	What is the name of the chief executive officer? What is the total net income for report year? What is the tier 1 capital ratio?	648	3,522
Synthetic Invoices	What is the number of items on the invoice? What is the total net amount of the item described on the invoice? What is the description of the item of the transaction?	2,707	17,274
Charity Annual Reports	What is the independent auditor's name? What are the charity's total funds in the bank and in hand? What is the name of the organisation's chairman?	161	4,025

Table 5: Outline of private datasets used for SFT.

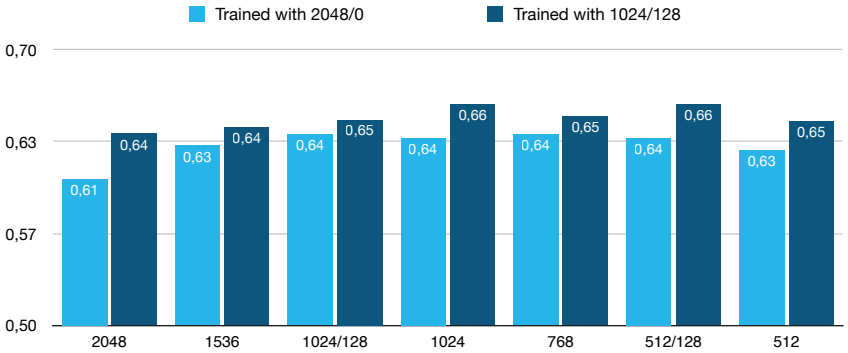


Figure 8: Impact of chunk size and overlap size (chunk/overlap) on a downstream inference for two models, assuming in-house dataset of business use cases. We observe no positive impact of overlap for sufficiently long input sequences, such as 1024 tokens.

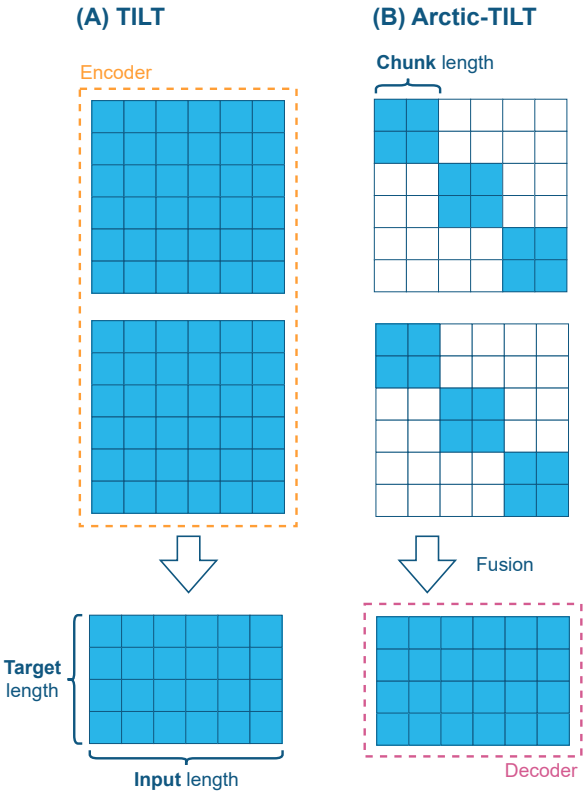


Figure 9: An illustration of sparse attention matrices assuming a two-layer encoder and decoder. The original TILT (A) consumes the complete input at once, in contrast to Arctic-TILT (B) with blockwise attention

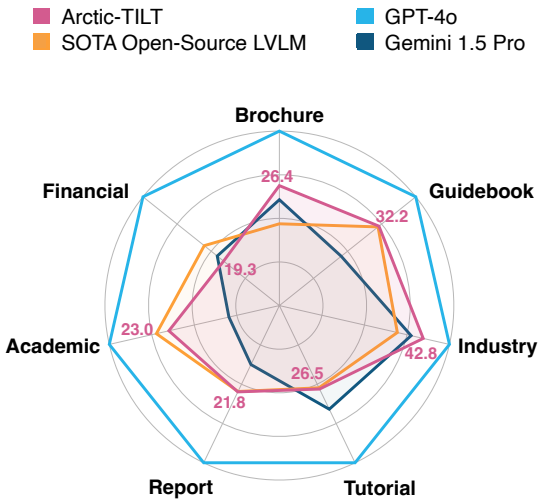


Figure 10: Fine-grained MMLongBench-Doc results. Arctic-TILT appears better than or comparable to the best open-source LVLMs and Gemini 1.5 Pro despite having at least 30x fewer parameters.

reduction to mean and weight decay to $1e{-}5$. Additionally, we used case augmentation of the whole triple consisting of the document, question, and answer. Specifically, if we detect that the document is not already cast to upper or lowercase, we create an augmented version of the three-tuple question-document-answer by casting them all to that case, similarly to Powalski et al. (2021). This means that there are up to three versions of each data point, such as the original one, uppercase, and lowercase.

Downstream tasks evaluation. For downstream task evaluation on benchmarks providing trainset (DocVQA, MP-DocVQA, DUDE, Kleister Charity, Kleister NDA, SlideVQA, InfographicsVQA, VQA-CD) we performed additional training with Optuna (Akiba et al., 2019) hyperparameter tuning. We performed 10-40 studies optimizing the following hyperparameters:

- *case augmentation* (on, off) – augment dataset with lowercased/uppercased version of training samples, in case they are statistically distinguishable;
- *answer variants sampling* (on, off) – for questions with multiple versions of the correct answer (e.g. `100`, `$100`), we either pick the same, or sample variant per epoch;
- *dropout* sampled with uniform distribution from the interval $(0, 0.2)$;
- *weight decay* sampled with log-uniform distribution from the interval $(1e{-}6, 1e{-}2)$;
- *learning rate* sampled with log-uniform distribution from the interval $(1e{-}4, 5e{-}3)$.

F Finetuning Study

GPT-4o baseline. Following the findings of Borchmann (2024), we assume input images of 2048px along longer dimensions (usually height) and similar prompts. The latter were subject to further per-dataset optimization to cover the convention used in considered datasets (final form presented in Table **??**).

Payment Stubs. The private dataset used for evaluation consists of American payment stubs, i.e., documents obtained by an employee regarding the salary received. The test split contains 39 documents with 448 annotations. Since all come from different companies, their layouts differ significantly. Questions aim to extract employee and employer names, dates, addresses and information from payment tables, where each row consists of payment type, hours worked, and payment amount,

Dataset	Prompt
Payment Stubs	Replace [ANSWER] with a value in the template given question and document. ↩ Question: [TEXT] ↩ Template: Based on the context, the answer to the question would be "[ANSWER]". ↩ ↩ Normalize amounts to two decimal places, without thousand separator and without dollar sign. ↩ Normalize states using postal abbreviations, e.g., TX or NJ.
Ghega Patents	Replace [ANSWER] with a value in the template given question and document. ↩ Question: [TEXT] ↩ Template: Based on the context, the answer to the question would be "[ANSWER]". ↩ ↩ Normalize dates to YYYY-MM-DD format except question about priority which should remain similar to "DD.MM.RRRR (country code) (optional number)."

Table 6: Final prompts used for GPT-4o baselines.

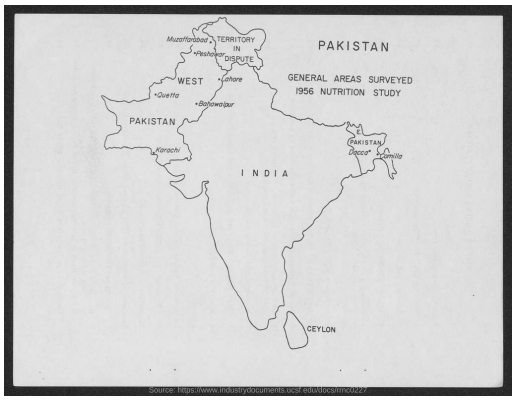


Figure 11: Question: Is Dacca in the West or East Pakistan? Arctic-TILT: West. Ground Truth: East. Due to a lack of advanced visual comprehension, the model cannot determine the city's precise location within the country.

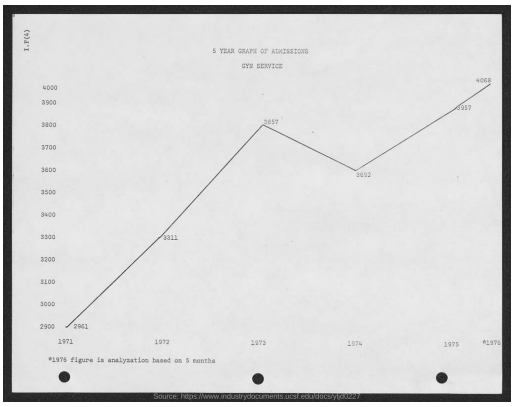


Figure 12: Question: In which year did admissions fall? Arctic-TILT: 1971. Ground Truth: 1974. Due to a lack of advanced visual comprehension and the limited presence of chart data in SFT datasets, the model cannot interpret the line plot correctly.

e.g., 'What is the name of the US state of the employee's address?' or 'When does the pay period finish?'

G Broader Evaluation Tables

This section presents a detailed performance analysis of various models on DUDE (Table 7), MMLongBench-Doc (Table 8), and a broad range of datasets featured in the main part of the paper (Table 9). Additionally, a fine-grained analysis of the top four models' performance is illustrated in Figure 10.

H Qualitative Examples of Model Errors

Qualitative analysis of model answers reveals limitations such as varied signs of limited visual comprehension (Figure 11, Figure 12, Figure 14, Figure 15), problems with counting (Figure 13). Additionally, because of relying on a third-party OCR engine, Arctic-TILT can copy from the provided textual layer that sometimes contains incorrectly recognized words (Figure 16).

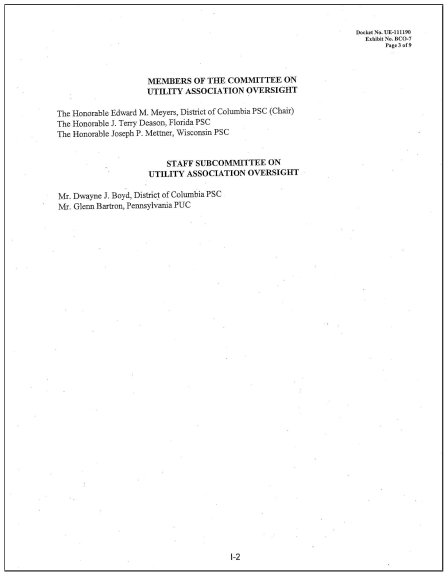
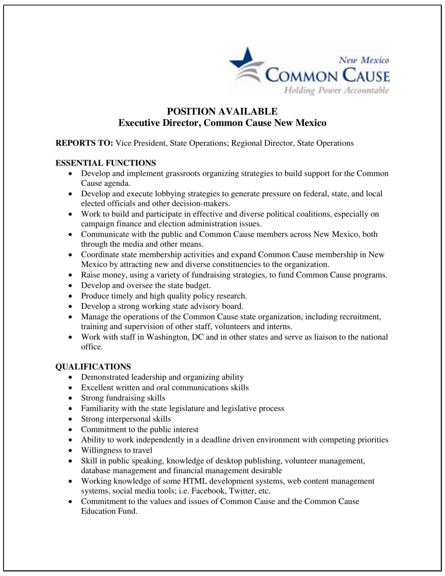


Figure 13: Question: How many Members of the Committee on Utility Association Oversight are there? Arctic-TILT: 4. Ground Truth: 3. Like many heavier LLMs, Arctic-TILT struggles with counting objects.



New Mexico
COMMON CAUSE
Holding Power Accountable

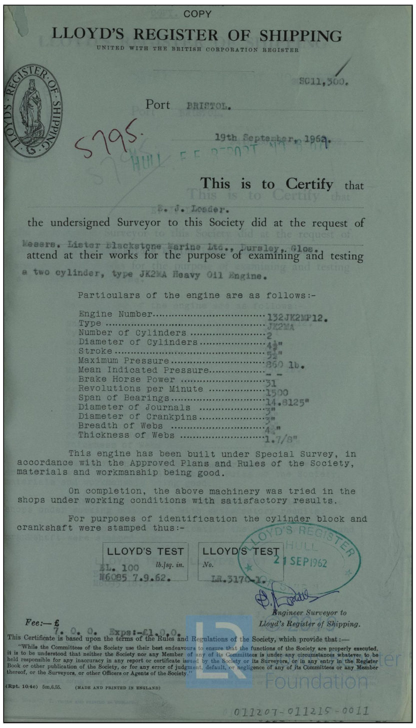
POSITION AVAILABLE
Executive Director, Common Cause New Mexico

REPORTS TO: Vice President, State Operations; Regional Director, State Operations

ESSENTIAL FUNCTIONS
- Develop and implement grassroots organizing strategies to build support for the Common Cause agenda.
- Develop and execute lobbying strategies to generate pressure on federal, state, and local elected officials and other decision-makers.
- Work to build and participate in effective and diverse political coalitions, especially on campaign finance and election administration issues.
- Communicate with the public and Common Cause members across New Mexico, both through the media and other means.
- Coordinate state membership activities and expand Common Cause membership in New Mexico by attracting new and diverse constituencies to the organization.
- Raise money, using a variety of fundraising strategies, to fund Common Cause programs.
- Develop and oversee the state budget.
- Produce timely and high quality policy research.
- Develop a strong working state advisory board.
- Manage the operations of the Common Cause state organization, including recruitment, training and supervision of other staff, volunteers and interns.
- Work with staff in Washington, DC and in other states and serve as liaison to the national office.

QUALIFICATIONS
- Demonstrated leadership and organizing ability
- Excellent written and oral communications skills
- Strong fundraising skills
- Familiarity with the state legislature and legislative process
- Strong interpersonal skills
- Commitment to the public interest
- Ability to work independently in a deadline driven environment with competing priorities
- Willingness to travel
- Skill in public speaking, knowledge of desktop publishing, volunteer management, database management and financial management desirable
- Working knowledge of some HTML development systems, web content management systems, social media tools; i.e. Facebook, Twitter, etc.
- Commitment to the values and issues of Common Cause and the Common Cause Education Fund.

Figure 14: Question: What colors are in the logo of the Common Cause? Arctic-TILT: blue, red. Ground Truth: blue, grey. Our image encoder consumes grayscale images, yielding color recognition based on guessing or approximations.



COPY

LLOYD'S REGISTER OF SHIPPING
UNITED WITH THE BRITISH CORPORATION REGISTER

LLOYD'S · REGISTER · OF · SHIPPING

SC11,300.

Port BRISTOL.

5795.

19th September, 1962.

HULL F.E. REPORT

This is to Certify that

E. J. Loader.

the undersigned Surveyor to this Society did at the request of

Messrs. Lister Blackstone Marine Ltd., Dursley, Glos.

attend at their works for the purpose of examining and testing

a two cylinder, type JK2MA Heavy Oil Engine.

Particulars of the engine are as follows:-

Engine Number........ 132JK2MP12.
Type JK2MA
Number of Cylinders 2
Diameter of Cylinders........ 4½"
Stroke........ 5½"
Maximum Pressure........ 860 lb.
Mean Indicated Pressure........ - -
Brake Horse Power 31
Revolutions per Minute 1500
Span of Bearings........ 14.8125"
Diameter of Journals .... 3"
Diameter of Crankpins........ 3"
Breadth of Webs .... 4½"
Thickness of Webs .... 1.7/8"

This engine has been built under Special Survey, in accordance with the Approved Plans and Rules of the Society, materials and workmanship being good.

On completion, the above machinery was tried in the shops under working conditions with satisfactory results.

For purposes of identification the cylinder block and crankshaft were stamped thus:-

LLOYD'S TEST lb./sq. in.	LLOYD'S TEST No.
BL. 100 M6085 7.9.62.	LR.3170-1.

LLOYD'S REGISTER HULL 21 SEP 1962

Engineer Surveyor to Lloyd's Register of Shipping.

Fee:—£ 7. 0. 0. Exps:-£1.0.0.

This Certificate is based upon the terms of the Rules and Regulations of the Society, which provide that:—

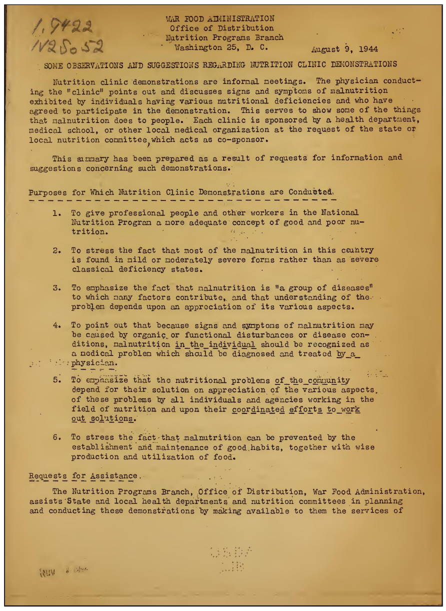
"While the Committees of the Society use their best endeavours to ensure that the functions of the Society are properly executed, it is to be understood that neither the Society nor any Member of any of its Committees is under any circumstances whatever to be held responsible for any inaccuracy in any report or certificate issued by the Society or its Surveyors, or in any entry in the Register Book or other publication of the Society, or for any error of judgment, default, or negligence of any of its Committees or any Member thereof, or the Surveyors, or other Officers or Agents of the Society."

(Rpt. 10/4c) 5m,6,55. (MADE AND PRINTED IN ENGLAND)

Lloyd's Register Foundation

011207-011215-0011

Figure 15: Question: What is the date on the stamp? Arctic-TILT: 19th September, 1962. Ground Truth: 1962-09-21. Arctic-TILT returns the first date found in the document, struggling to discriminate between the dates that appear in different visual contexts.



1,9422
N25052

WAR FOOD ADMINISTRATION
Office of Distribution
Nutrition Programs Branch
Washington 25, D. C.

August 9, 1944

SOME OBSERVATIONS AND SUGGESTIONS REGARDING NUTRITION CLINIC DEMONSTRATIONS

Nutrition clinic demonstrations are informal meetings. The physician conducting the "clinic" points out and discusses signs and symptoms of malnutrition exhibited by individuals having various nutritional deficiencies and who have agreed to participate in the demonstration. This serves to show some of the things that malnutrition does to people. Each clinic is sponsored by a health department, medical school, or other local medical organization at the request of the state or local nutrition committee, which acts as co-sponsor.

This summary has been prepared as a result of requests for information and suggestions concerning such demonstrations.

Purposes for Which Nutrition Clinic Demonstrations are Conducted.

1. To give professional people and other workers in the National Nutrition Program a more adequate concept of good and poor nutrition.

2. To stress the fact that most of the malnutrition in this country is found in mild or moderately severe forms rather than as severe classical deficiency states.

3. To emphasize the fact that malnutrition is "a group of diseases" to which many factors contribute, and that understanding of the problem depends upon an appreciation of its various aspects.

4. To point out that because signs and symptoms of malnutrition may be caused by organic or functional disturbances or disease conditions, malnutrition in the individual should be recognized as a medical problem which should be diagnosed and treated by a physician.

5. To emphasize that the nutritional problems of the community depend for their solution on appreciation of the various aspects of these problems by all individuals and agencies working in the field of nutrition and upon their coordinated efforts to work out solutions.

6. To stress the fact that malnutrition can be prevented by the establishment and maintenance of good habits, together with wise production and utilization of food.

Requests for Assistance.

The Nutrition Programs Branch, Office of Distribution, War Food Administration, assists State and local health departments and nutrition committees in planning and conducting these demonstrations by making available to them the services of

Figure 16: Question: What is the mentioned branch of war food administration? Arctic-TILT: Nutrition Program Branch. Ground Truth: Nutrition Programs Branch. Because of relying on a third-party OCR engine, Arctic-TILT can copy from the provided textual layer that contains incorrectly recognized words (here in singular form instead of plural).

I Contributions

ŁB. Performing early-stage architecture ablations, writing most of the paper and preparing figures, final fusion by TP module design and related ablation studies, implementation of GPT-4o baseline for Arctic-TILT SFT, analysis of results on public benchmarks, overseeing initial model sparsification experiments, study of long context utilization, self-supervised pretraining of the model.

ŁD. Various contributions related to configurations and automatizations of experiments.

PD. Optimization of various data pipelines (image processing, loading, metric computation). Datasets updates and modifications (main focus on increasing loading speed and OCR correctness).

ŁG. Technical leadership and participation in the codebase implementation, performance optimization, and attention sparsity efforts. Writing parts of the paper.

AG. Efficient implementation of attention sparsity. Contributions to codebase implementation and memory optimizations. Preparation and cleaning of some datasets. Experiments with model sizes and architectures.

PH. Major contributions to memory- and compute-efficiency, including evaluation of long context approaches with a theoretical memory model, implementing nested checkpointing and mixed-precision, and empirically evaluating the complete solution's performance and memory usage characteristics.

WJ. Leading SFT efforts, including finetuning the model's final version, experiments with hyperparameters, training protocols, and dataset composition. Model performance improvements. Various contributions in the model and training code. Preparation and cleaning of some datasets.

PJ. Memory fragmentation handling. Implementation and creation of semi-synthetic long document Needle-in-a-Haystack benchmark used in early experiments. Data management, curating final training and performing a few downstream evaluations (DUDE, VQA_CD). Wrote parts of the paper. Performed final ablations.

DJ. Leading efforts to increase the context length from 25 to 500 pages (conceptualization, brainstorming, planning, guidance). Conducting initial memory and throughput experiments, as well as final stress tests and quality experiments for very large context lengths. Implementation of CPU offloading. Performing few-shot finetuning experiments. Delivering results for SlideVQA, Kleister Charity, and Kleister NDA datasets.

PL. Implemented TP fusion and conducted ablation studies focusing on module placement. Devised enhanced training protocol, performed hyperparameter tuning, and contributed to various model improvements. Performed experimental evaluation on DocVQA and InfographicsVQA.

GN. Preparation and management of datasets, the idea behind creating semi-synthetic long documents, automation of data processing pipeline, conducting experiments and analyzing the results with long documents.

JO. Selection and improvements of the training datasets (analysis of data quality, filtering the data, fixing quality issues), optimizations of image encoder and data processing.

MP. Performing early-stage architecture ablations (researching, implementing, and studying effects) that lead to co-authoring TP fusion (i.e., proposing initial attn-based version, module placement study, and fusion in every layer). Leading efforts in writing parts of the paper (technical optimizations, training, related works, analysis, structuring and rewriting).

KS. The idea behind attention sparsification, ablation studies of various approaches, implementation of required prototypes, and analysis of the results.

MT. Training loop optimization (in terms of processing time and data efficiency), performing downstream evaluations (DocVQA, MP-DocVQA, MMLongBench-Doc), dataset preparation and cleaning, error analysis, and organization of work.

AZ. Various contributions to the development of the codebase.

```
    class TiltLayerNorm(nn.Module):
        """
        This is essentially the T5 modification of layer norm, referred to as RMS norm.

        Args:
            dim: the dimension of vectors to be normalized, i.e. the last dimension of the input tensor
            eps: small positive value added to computed second moment for numerical stability
        """

        def __init__(self, dim: int, eps: float = 1e-6) -> None:
            super().__init__()
            self.w = nn.Parameter(torch.ones(dim))
            self.eps = eps
            self.init_weights()

        def forward(self, inp: Tensor) -> Tensor:
            dtype = inp.dtype
            x = inp.to(torch.float32)
            squared_norm = x.pow(2).mean(dim=-1, keepdim=True)
            x = x * torch.rsqrt(squared_norm + self.eps)
            return self.w * x.to(dtype)

        def init_weights(self, factor: float = 1.0) -> None:
            self.w.data.fill_(factor * 1.0)


    class TiltPostFusionModule(nn.Module):
        """
        Introduced in the Arctic-TILT paper.

        Args:
            d_model: dimension of input vectors
            dropout: probability of dropout applied to input embeddings
            layer_norm: the module responsible for input embeddings
        """

        def __init__(self, d_model: int, dropout: float, layer_norm: TiltLayerNorm):
            super().__init__()
            self.layer_norm = layer_norm
            self.to_v = nn.Linear(d_model, d_model, bias=False)
            self.to_out = nn.Linear(d_model, d_model, bias=False)
            self.to_r = nn.Linear(d_model, d_model, bias=False)
            self.dropout = nn.Dropout(dropout)

        def forward(self, text_queries: Tensor, image_queries: Tensor) -> Tensor:
            """
            Compute module's forward pass.

            Args:
                text_queries (Tensor): Tensor representing the primary input in the fusion, which is text-
     based, or mixed.
                image_queries (Tensor):  Tensor representing the secondary input in the fusion, which is
     image-based.
            """
            bs, l, d = text_queries.shape
            inputs = torch.stack([text_queries, image_queries], dim=-2)
            inputs = inputs.view(bs * l, 2, d)
            normed_inputs = self.dropout(self.layer_norm(inputs))
            normed_primary_input = normed_inputs[:, 0]
            out: Tensor = self.to_v(normed_inputs.sum(-2))
            out = out + out * self.to_r(normed_primary_input)
            out = self.to_out(out)
            out = out.view(bs, l, d)
            return text_queries + out
```

Listing 1: Complete Arctic-TILT modality fusion module.

Method	ANLS↑	ECE↓	AURC↓	AUROC	Extract↑	Abstract↑	List↑	Unanswerable↑
Arctic-TILT 0.8B	**58.1**	**7.6**	**25.3**	52.9	**62.7**	**56.5**	46.7	62.6
GPT-4 Vt + Azure OCR	53.9	55.8	43.2	50.0	59.7	52.5	**57.9**	51.3
GRAM	53.4	44.0	44.0	50.0	56.8	52.3	20.0	65.4
GRAM C-Former	51.0	46.1	46.1	50.0	55.1	50.5	17.3	61.0
DocGptVQA	50.0	22.4	42.1	87.4	51.9	48.3	28.2	62.0
DocBlipVQA	47.6	30.6	48.6	78.3	50.7	46.3	30.7	55.2
model_0327	46.6	19.0	44.0	88.5	55.2	46.6	17.9	47.3
T5-concat	38.7	24.9	43.4	51.1	37.3	37.5	16.8	52.9
Multi-Modal T5 (2023-04-20)	37.9	59.3	59.3	50.0	41.5	40.2	20.2	34.7
Multi-Modal T5 (2023-04-19)	37.9	59.3	59.3	50.0	41.5	40.2	20.3	34.7
Hi-VT5	35.7	61.4	61.0	50.0	28.3	33.0	10.6	**62.9**
Hi-VT5 w. token type	35.6	28.0	46.0	48.8	30.9	35.1	11.8	52.5
QAP	11.6	41.7	90.8	50.1	0.1	0.1	0.0	62.0

Table 7: DUDE performance metrics for different methods. Bolded is the best result in a given criteria.

Model	#Param	Context Window	ACC	F1
GPT-4o	-	128k	42.8	44.9
GPT-4V(vision)	-	128k	32.4	31.2
Arctic-TILT	822M	390k	25.8	–
Gemini-1.5-Pro	-	32k	31.2	24.8
GPT-4o	-	128k	30.1	30.5
GPT-4-turbo	-	128k	27.6	25.9
Mixtral-Instruct-v0.1	8x22B	64k	26.9	24.7
Claude-3 Opus	-	32k	26.9	24.5
DeepSeek-V2	-	32k	24.9	19.6
Gemini-1.5-Pro	128k	-	22.8	20.6
QWen-Plus	-	32k	18.9	13.4
Mixtral-Instruct-v0.1	8x7B	32k	17.0	16.9
Mistral-Instruct-v0.2	7B	32k	16.4	13.8
ChatGLM-128K	6B	128k	16.3	14.9
InternLM-Chat-V1.5	26B	8k	13.5	13.0
InternLM-XC2-4KHD	8B	8k	8.8	8.9
MiniCPM-Llama3-V2.5	8B	8k	8.5	8.6
EMU2-Chat	37B	2k	8.3	5.5
Claude-3 Opus	200k	-	7.6	7.4
DeepSeek-VL-Chat	7.3B	4k	7.4	5.4
Idefics2	8B	8k	7.0	6.8
mPLUG-DocOwl 1.5	8.1B	4k	6.9	6.3
Monkey-Chat	9.8B	2k	6.2	5.6
Qwen-VL-Chat	9.6B	6k	6.1	5.4
CogVLM2-LLAMA3-Chat	19B	8k	4.4	4.0

Table 8: Performance metrics for different models. Results follow Ma et al. (2024).

Model	Size	MP-DocVQA	Kleister Charity	Kleister NDA	DUDE Bench-Doc	MMLongVQA	Slide-Lay	ArXiv-Lay	PubMed VQA	Doc CD	VQA VQA	Infographic VQA
Arctic-TILT	823M	81.2	88.1	94.1	58.1	–	55.1	44.4	44.8	90.2	90.7	–
Arctic-TILT†	822M	76.9	86.9	38.3	55.9	25.8	40.4	–	–	88.6	88.7	57.0
ERNIE Layout (1)	355M	–	–	88.1	–	–	–	–	–	88.4	–	–
LAMBERT (2)	125M	–	83.6	81.8	–	–	–	–	–	–	–	–
Hi-VT5 (3)	784M	73.5	–	–	49.2	–	–	–	–	–	–	–
InternVL 1.5 (4)	26B	–	–	–	–	–	–	–	–	90.9	–	72.5
InternVL2-Pro	108B	–	–	–	–	–	–	–	–	95.1	–	83.3
Claude-3 Opus (OCR)	72B	–	–	–	–	26.9	–	–	–	89.3	–	–
Gemini-1.5-Pro (OCR)	–	–	–	–	46.0	31.2	–	–	–	93.1	–	81.0
LayoutLMv2 (5)	426M	–	–	85.2	–	–	26.5	–	–	86.7	–	28.3
GPT-4o (OCR)	–	–	–	–	–	30.1	–	–	–	–	–	–
GPT-4 Vi + OCR (6)	–	–	–	–	53.9	–	57.3	–	–	–	–	–
mPLUG-DocOwl 1.5 (7)	8.1B	–	–	–	–	6.9	–	–	–	81.6	–	50.4
mPLUG-DocOwl 2 (8)	8B	–	–	–	–	6.9	–	–	–	80.7	–	46.4
TextMonkey (9)	9.7B	–	–	–	–	–	–	–	–	64.3	–	28.2
GRAM (10)	859M	79.7	–	–	53.4	–	–	–	–	–	–	–
UReader (11)	86M	–	–	–	–	–	–	–	–	65.4	–	42.2
BigBird... + Layout (12)	581M	–	–	–	–	–	–	41.2	42.1	–	–	–
Pix2Struct Large (13)	1.3B	–	–	–	–	–	–	–	–	76.6	–	40.0
Donut (14)	176M	–	–	–	–	–	–	–	–	67.5	–	11.6
StrucTexTv3 (15)	1.8B	–	–	–	–	–	–	–	–	72.8	–	68.6
DocKylin (16)	7B	–	–	–	–	–	–	–	–	77.3	–	46.6

Table 9: Comparison of DU-models and LLMs on a broad range of datasets. (1) Peng et al. (2022), (2) Garncarek et al. (2021), (3) Tito et al. (2023), (4) Chen et al. (2024), (5) Xu et al. (2020), (6) Borchmann (2024), (7) Hu et al. (2024a), (8) Hu et al. (2024b), (9) Liu et al. (2024b), (10) Blau et al. (2024), (11) Ye et al. (2023a), (12) Nguyen et al. (2023), (13) Lee et al. (2023b), (14) Kim et al. (2022), (15) Lyu et al. (2024), (16) Zhang et al. (2024).