snowflake®

STATE-OF-THE-ART DOCUMENT AI ON A SINGLE 24GB GPU

Łukasz Borchmann

ML IN PL 2025

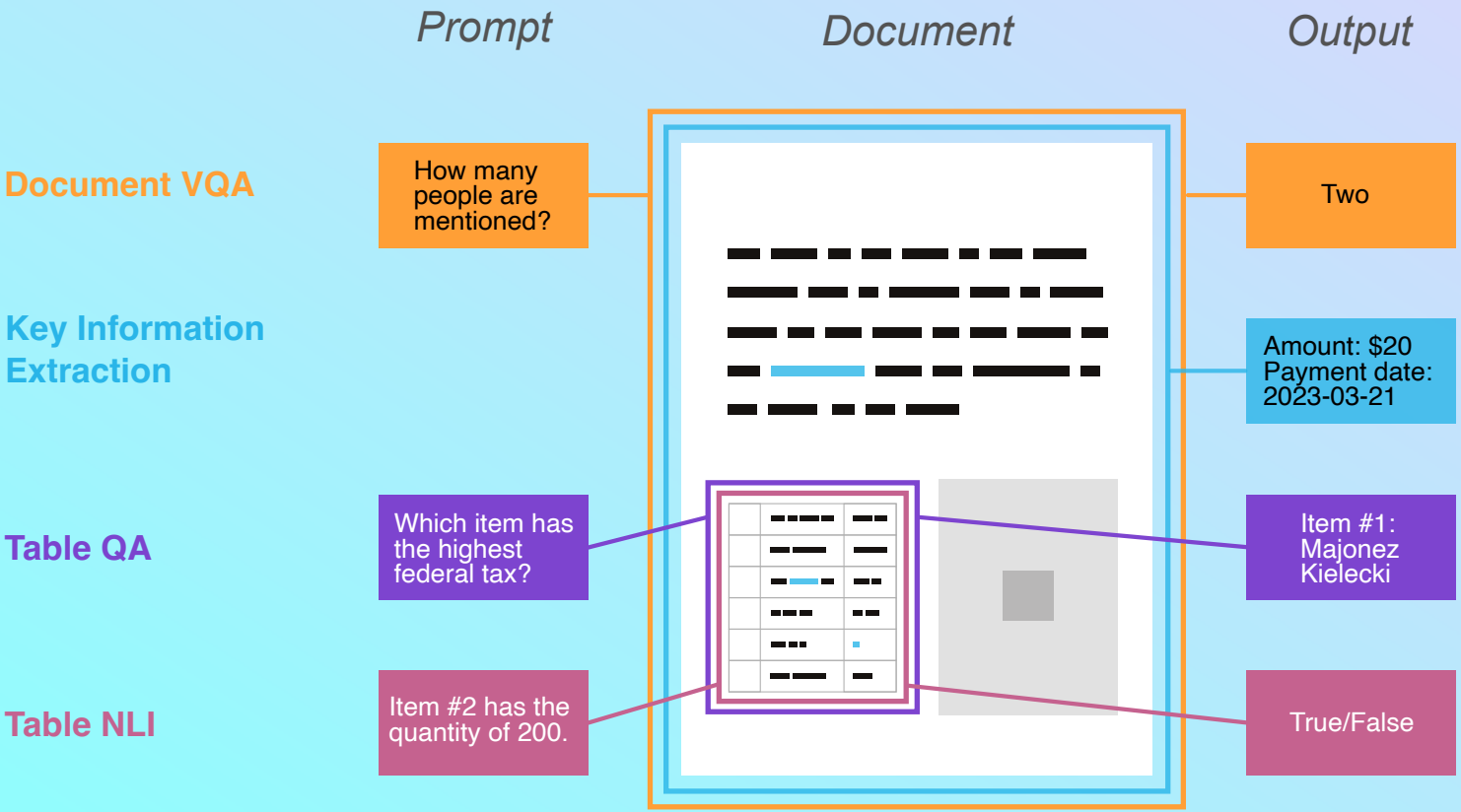
WE ARE
OPEN SOURCE

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

What do we mean by *Understanding?*

The term is specifically defined as the **capacity to convert a document into meaningful information** (see, e.g. Haralick, 1994; Yacoub, 2003; Dehghani, 2018).

The exact nature of this information depends on the task under consideration and **can range from the location of document components to the answer valid for some content-related questions formulated in natural language**.


Prompt
Document
Output
Document VQA
How many people are mentioned?
Two
Key Information Extraction
Amount: $20
Payment date: 2023-03-21
Table QA
Which item has the highest federal tax?
Item #1: Majonez Kielecki
Table NLI
Item #2 has the quantity of 200.
True/False

Total Debt in 3Q10: 4,430.5 mUSD

2D Document Comprehension

(US$ Mn)	2009	LTM	3Q09	3Q10
Revenues	1,998.	2,451.3	477.1	664.6
% Domestic	42.4%	42.8%	47.4%	46.7%
Gross Profit	456.7	717.9	99.6	221.9
Gross margin	30%	29%	34%	41%
EBIT	257.4	451.6	49.2	149.1
EBITDA	311.5	749.7	116.8	232.2
EBITDA margin	26%	31%	24%	35%
(Interest Expense)	(212.2)	(229.3)	(53.4)	(64.5)
(Capes)	(346.1)	(304.3)	(194.4)	(92.1)
(Change working cap)	222.8	(161.9)	84.5	(138.0)
EBITDA-IE-Capex- ΔWC	176.1	54.1	(46.5)	(54.4)
Cash	1,452.2	2,148.7	1,384.6	2,148.7
Short Term Debt	803.4	756.3	874.2	756.3
Long Term Debt	2,922.1	3,674.2	2,784.8	3,674.2
Total Debt	3,725.4	4,430.5	3,659.1	4,430.5
Net Debt	2,273.3	2,281.8	2,274.5	2,281.8
Summarized Metrics	**2009**	**LTM**	**3Q09**	**3Q10**
EBITDA/Interest	2.4x	3.3x	2.2x	3.5x
Total Debt/EBITDA	7.x	5.9x	6.9x	5.9x

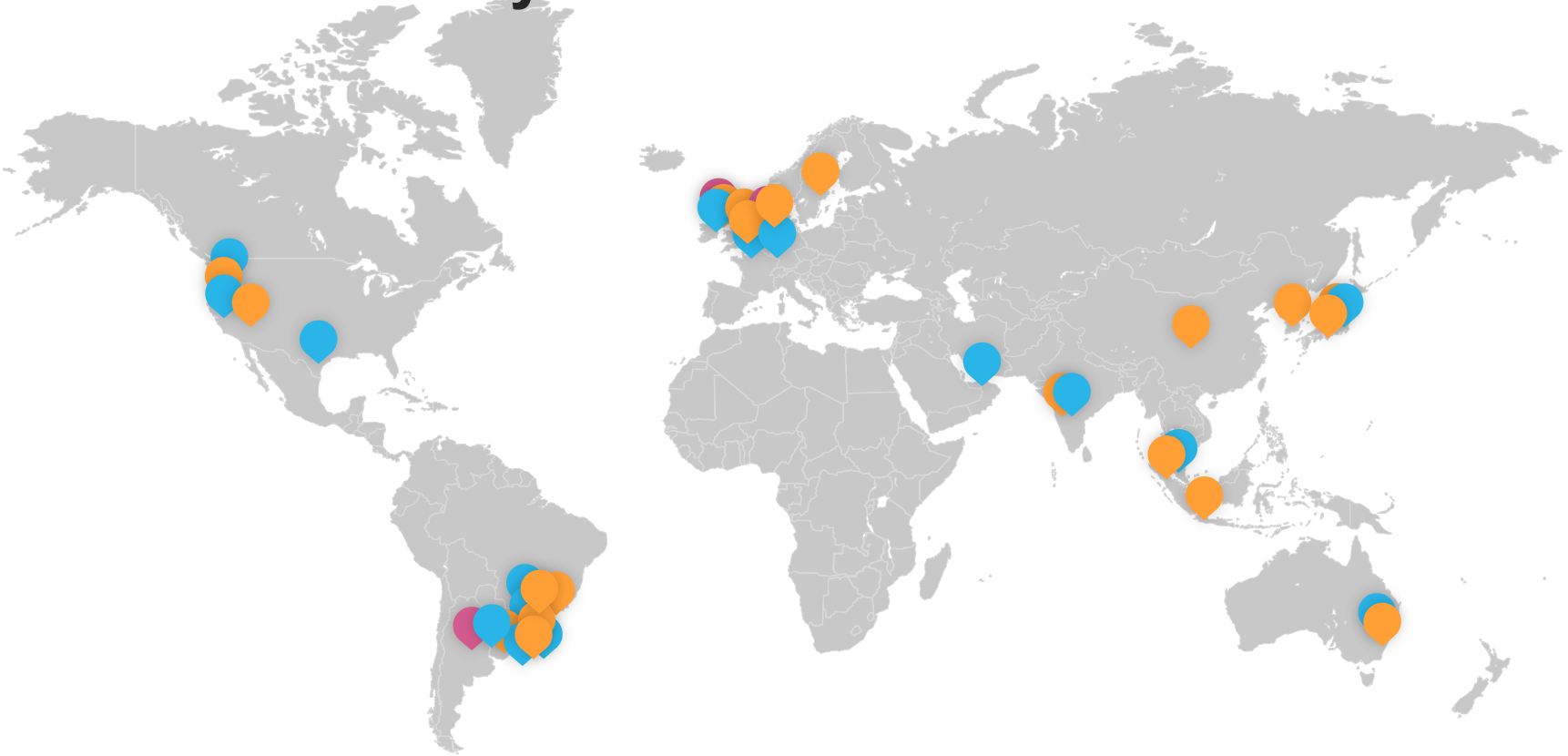
Total Debt in 3Q10: ?

1D Methods

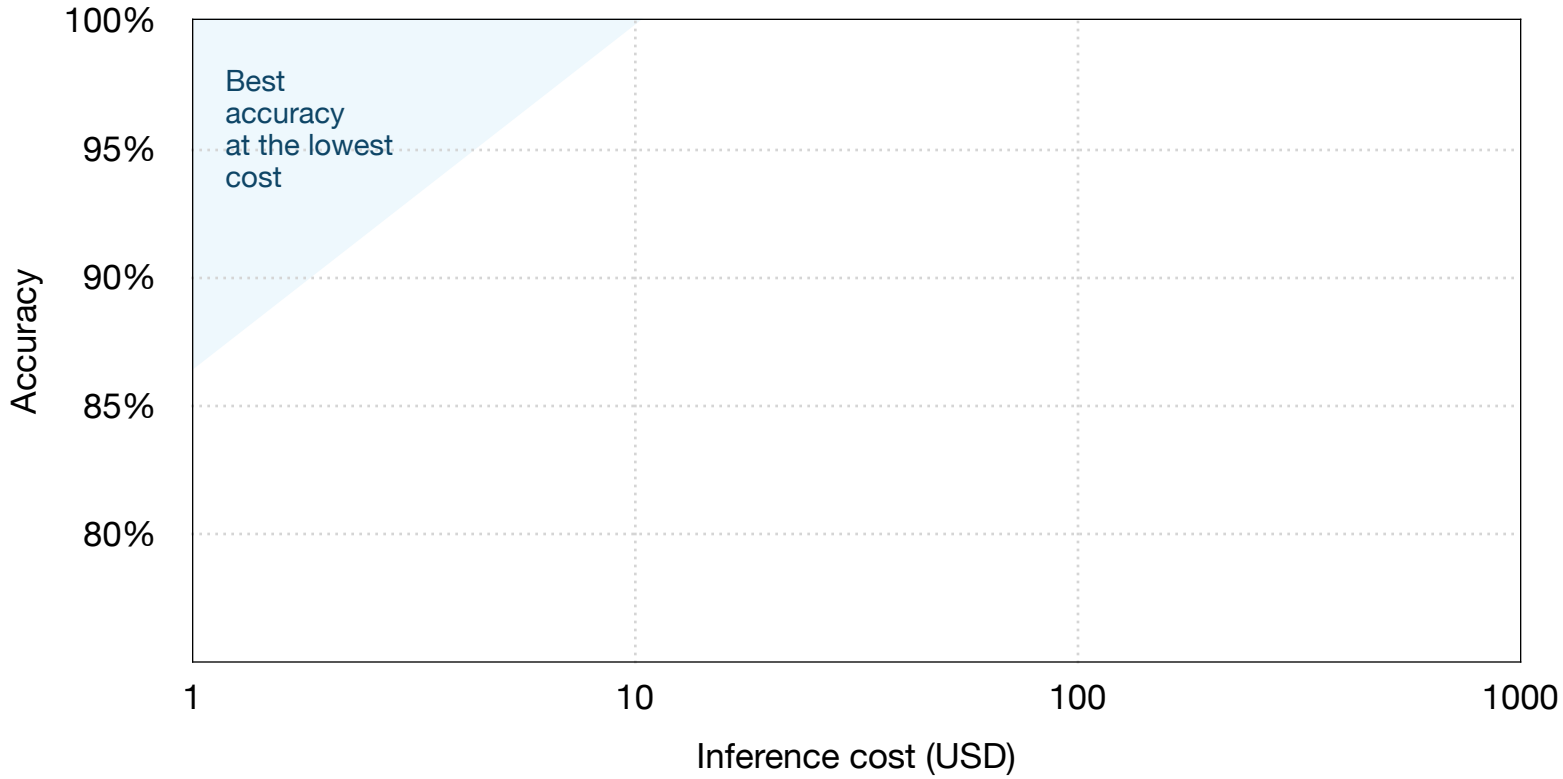
EBITDA-IE-Capex-ΔWC 176.1 54.1 (46.5)(59.4) Cash 1,452.2 2,148.7 1,384.6 2,148.7Short Term Debt 803.4 756.3 874.2 756.3 495 Long Term Debt 2,922.1 3,674.2 2,784.8 3,674.2 Total Debt 3,725.4 4,430.5 3,659.1 4,430.5 Net Debt 2,273.3 2,281.8 2,274.5 2,281.8 Summarized Metrics 2009 LTM 3Q09 3Q10 EBITDA/Interest 2.4x 3.3x 2.2x 3.5xTotal Debt/EBITDA 7.3x 5.9x 6.9x 5.9x

WHAT DOES
THE PRODUCT NEED?

Global Availability



Tier-One Results for Tier-Two Cost



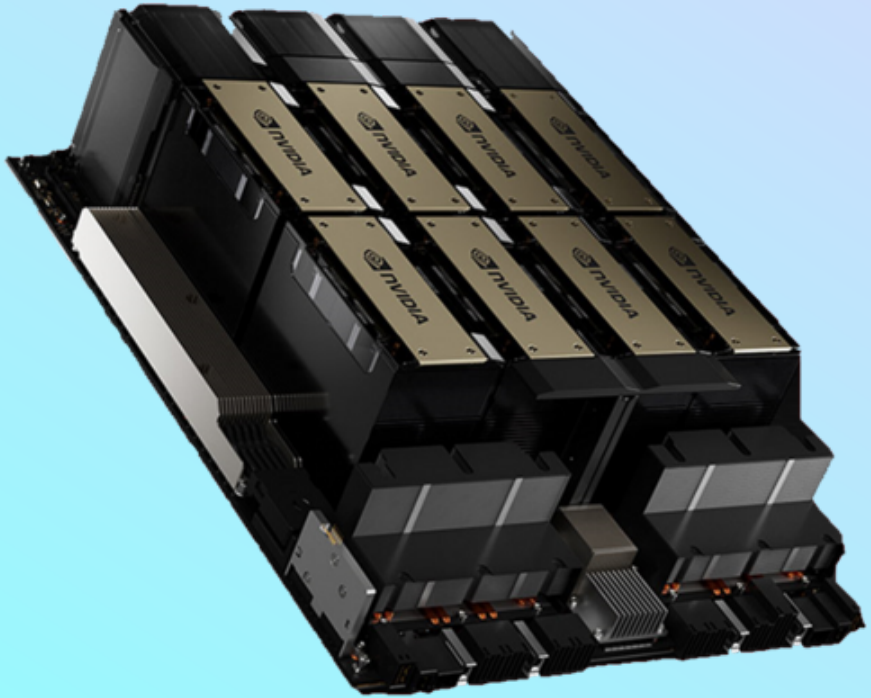
+ Couple More, e.g.

Long Context Support

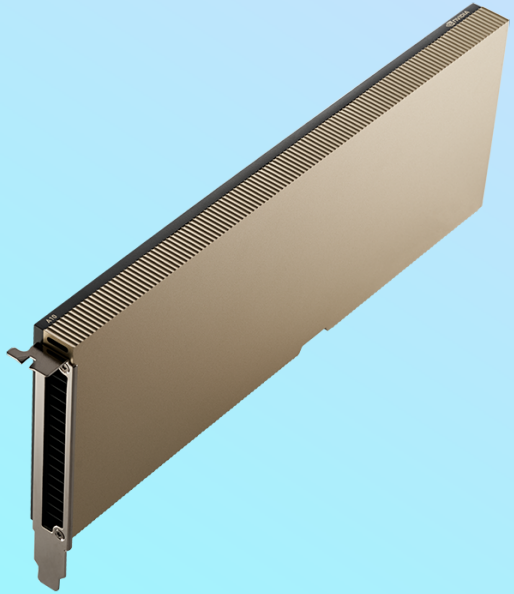
Scalability

Reliable Confidence Scores

H200



A10



Assumption: the best one can do with a single A10 or L4 GPU (24GB vRAM)

Both inference and **fine-tuning**

Why fine-tuning is that important?

LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report

Justin Zhao, Timothy Wang
Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky,
Piero Molino, Travis Addair, Devvret Rishi

Predibase

Abstract

Low Rank Adaptation (LoRA) has emerged as one of the most widely adopted methods for Parameter Efficient Fine-Tuning (PEFT) of Large Language Models (LLMs). LoRA reduces the number of trainable parameters and memory usage while achieving comparable performance to full fine-tuning. We aim to assess the viability of training and serving LLMs fine-tuned with LoRA in real-world applications. First, we measure the quality of LLMs fine-tuned with quantized low rank adapters across 10 base models and 31 tasks for a total of 310 models. We find that 4-bit LoRA fine-tuned models outperform base models by 34 points and GPT-4 by 10 points on average. Second, we investigate the most effective base models for

hosts 25 LoRA fine-tuned Mistral-7B LLMs on a single NVIDIA A100 GPU with 80GB memory. LoRA Land highlights the quality and cost-effectiveness of employing multiple specialized LLMs over a single, general-purpose LLM.

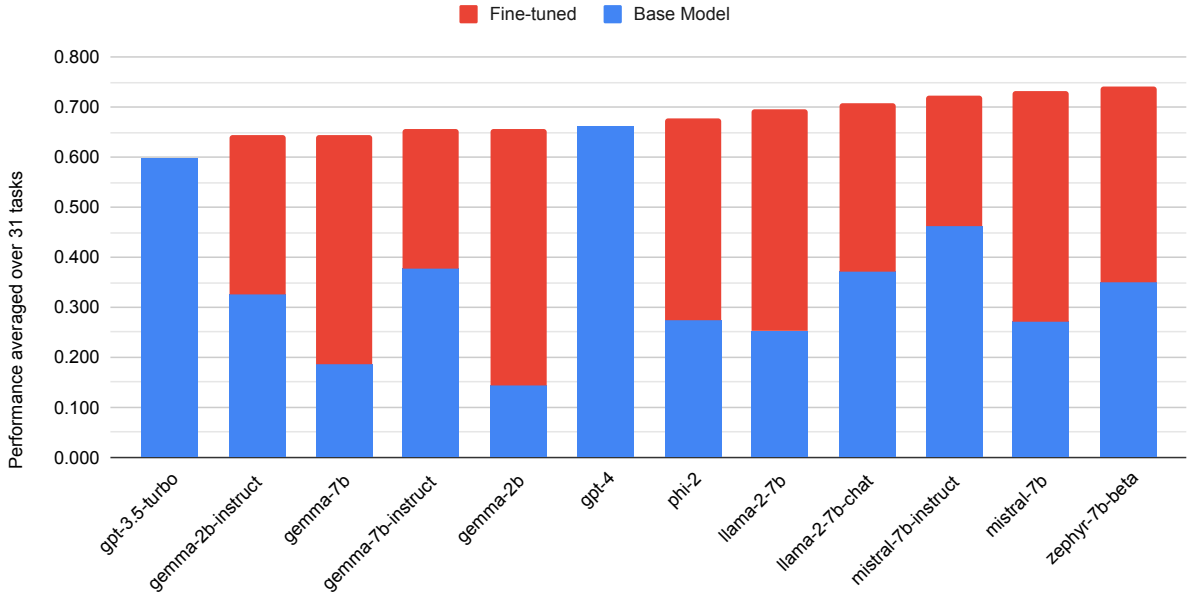
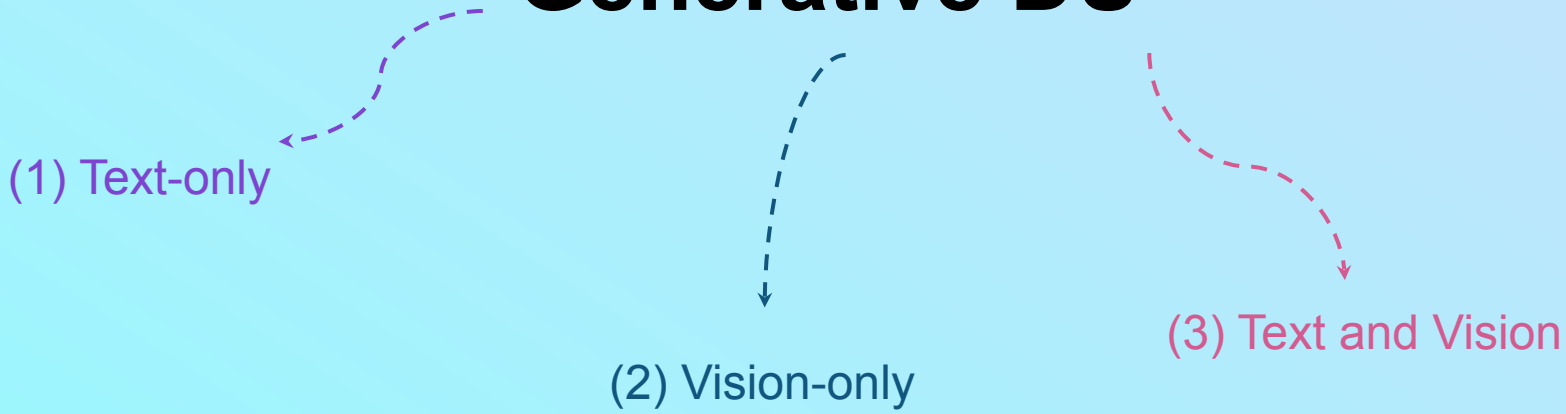
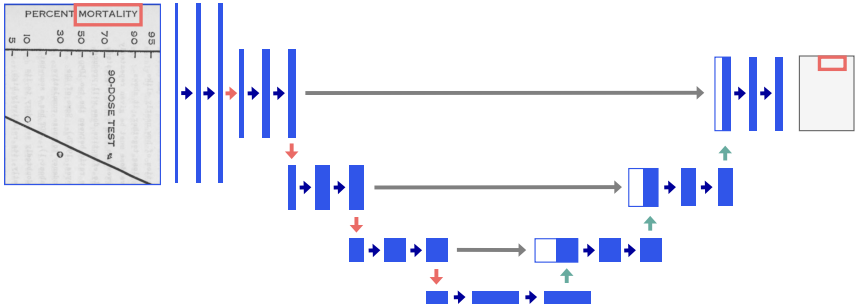


Figure 1: Average model performance for GPT-3.5, GPT-4, and 310 LLMs, before and after fine-tuning with LoRA, across 31 different tasks and 10 different base models. Zephyr-7b and Mistral-7b models exhibit the best performance after LoRA-based fine-tuning.

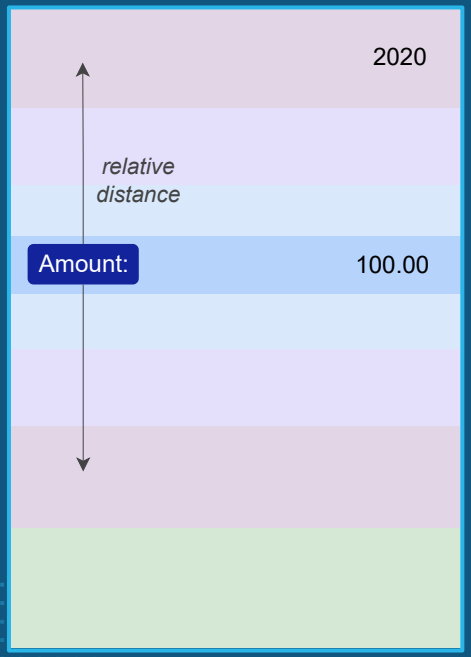
Generative DU

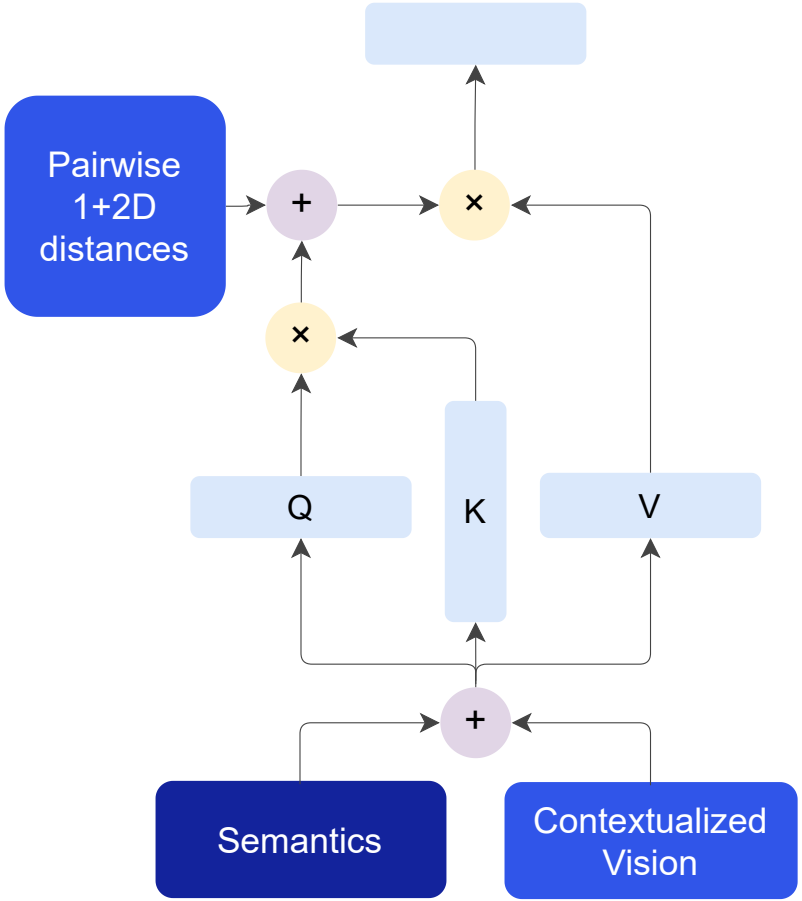


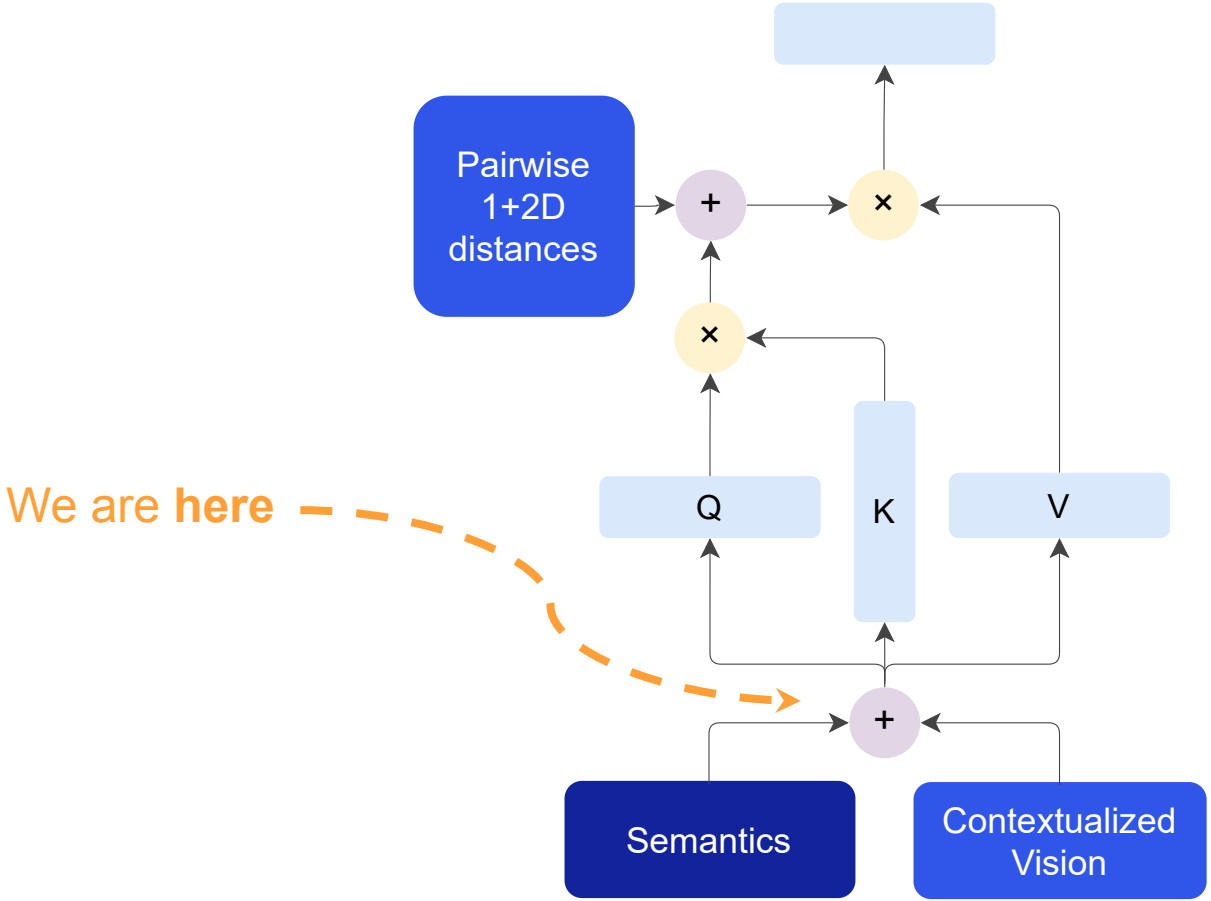
U-Net Visual Encoder



2D positional biases




Pairwise
1+2D
distances
+
×
×
Q
K
V
+
Semantics
Contextualized
Vision


Pairwise
1+2D
distances
+
×
×
Q
K
V
We are **here**
+
Semantics
Contextualized
Vision

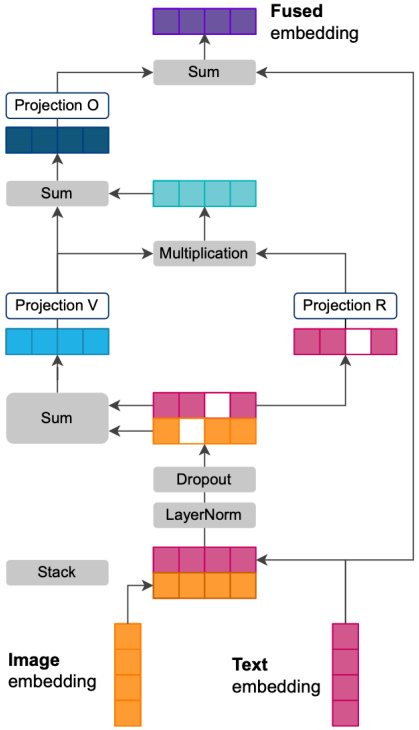


Figure 2: Arctic-TILT modality fusion. It can be seen as attention with role vector (Schlag et al., 2019) simplified concerning we calculate it over a pair of aligned text and image tokens.

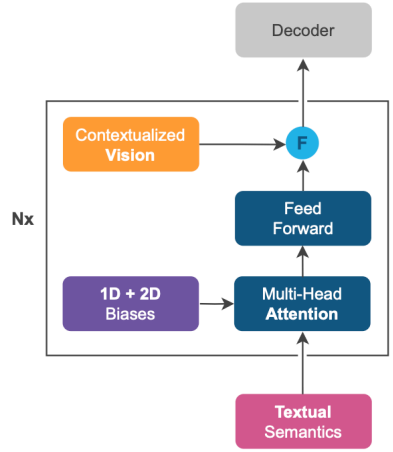


Figure 3: The Arctic-TILT encoder block combines *Contextualized Vision* from U-Net and *Textual Semantics* from input embeddings through *Fusion* (F) operation. The *Multi-Head Attention* is augmented with *1D and 2D positional biases* to capture spatial and sequential arrangement. This procedure is repeated in each layer (*Nx*), allowing to process integrated information further.

Max Input Length

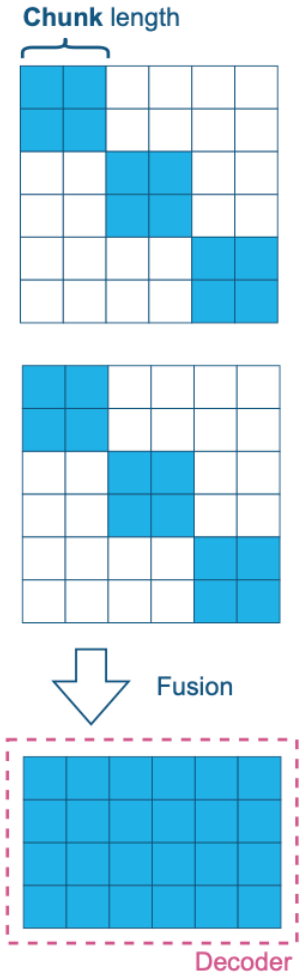
	Inference	Training
Vanilla TILT	9k	4k
+ attention sparsity	87k	41k
+ mixed precision	179k	51k
+ memory efficient attention	183k	56k
Inference-only optimizations		
+ no cross-attention KV cache	389k	
Training-only optimizations		
+ nested checkpointing		230k
+ CPU offloading		256k
+ random chunks		389k

Max Input Length

	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

Max Input Length

	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



Max Input Length

	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

Max Input Length

	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

Max Input Length

	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

Max Input Length

	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

Dataset	Industrial	Multipage	State-of-the-Art		Arctic-TILT
MP-DocVQA	✓	✓	GRAM	80.3	**81.2**
Kleister Charity	✓	✓	LAMBERT	83.6	**88.1**
Kleister NDA	✓	✓	ERNIE-Layout	88.1	**94.3**
DUDE	✓/ ✗	✓	GPT-4Vt + OCR	53.9	**58.1**
MMLongBench-Doc†	✓/ ✗	✓	GPT-4o	**42.8**	25.8
SlideVQA	✗	✓	GPT-4Vt + OCR	**57.3**	55.1
ArXiv-Lay	✗	✓	BigBird-Pegasus+Layout	41.2	**44.4**
PubMed-Lay	✗	✓	BigBird-Pegasus+Layout	42.1	**44.8**
DocVQA	✓	✗	InternVL 2.0 Pro	**95.1**	90.2
VQA-CD	✓	✗	QALayout	42.5	**90.7**
InfographicsVQA	✗	✗	InternVL 2.0 Pro	**86.8**	57.0

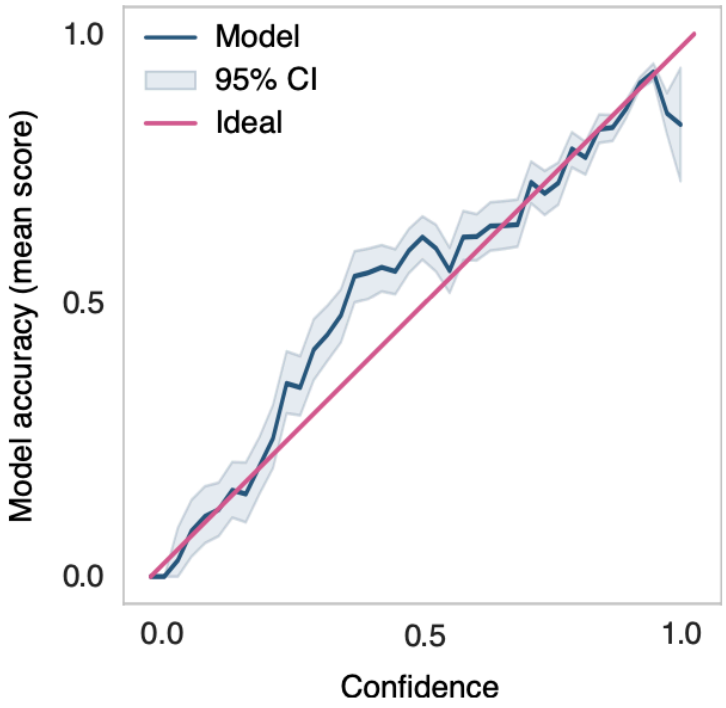
Table 3: Arctic-TILT compared to the previous state-of-the-art. Our model remains competitive despite having less than 1B parameters and excels when input is a long, business document. We use the original metrics for each dataset, i.e., F1 for Kleisters, Accuracy for MMLongBench-Doc, EM for SlideVQA, ROUGE-L for ArXiv-Lay and PubMed-Lay, and ANLS for the remaining tasks; † denotes zero-shot evaluation.

Arctic-TILT

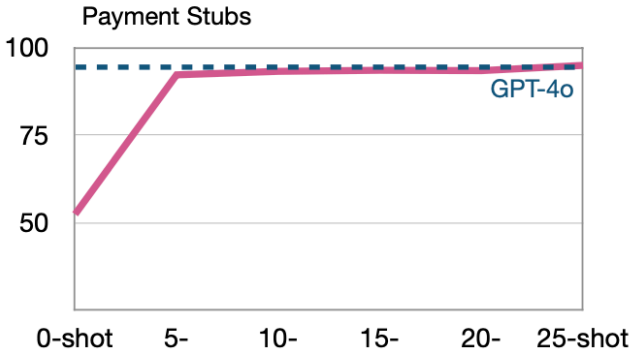
Document Understanding at Sub-Billion Scale

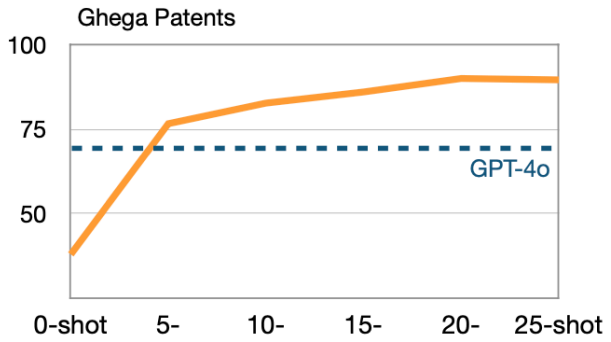
- Summarize Content
- Answer Questions
- Extract Key Information
- Finetune Model

24GB GPU **500 PAGES**


Model
95% CI
Ideal
1.0
0.5
0.0
Model accuracy (mean score)
0.0
0.5
1.0
Confidence

Improvement of Arctic-TILT zero-shot accuracy given fine-tuning on up to 25 annotated documents. Zero-shot performance of GPT-4o for comparison.





THANK YOU

WE ARE
OPEN SOURCE

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