Multimodal Structured Generation & CVPR's 2nd MMFM Challenge

By Franz Louis Cesista

franzlouiscesista@gmail.com

leloykun.github.io

Outline

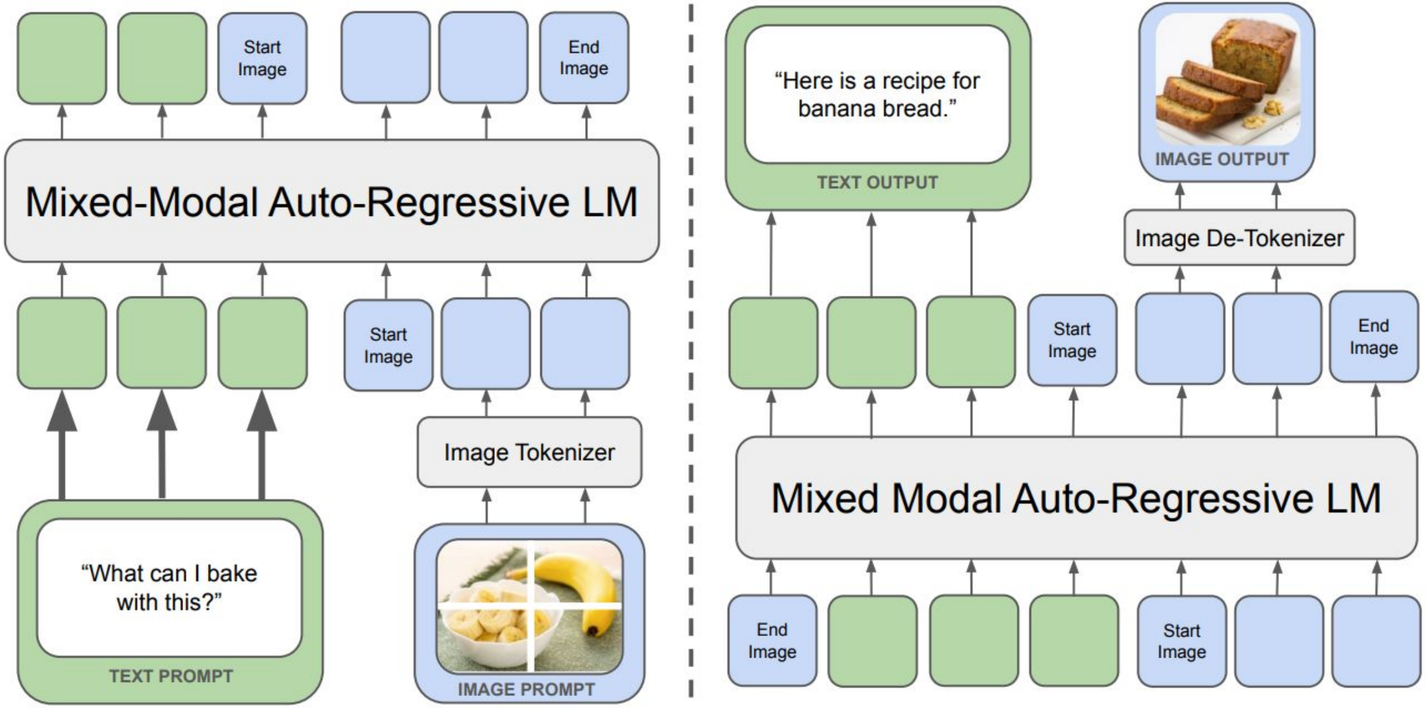
1. A brief overview of vision-language models (VLMs)
2. A brief description of CVPR's Multimodal Foundation Models (MMFM) Challenge
3. An overview of my approach, Multimodal Structured Generation
4. Results
5. Four possible reasons why current VLMs suck at doc-understanding tasks and what to do about them
6. Bonus demo: Interleaved Multimodal Structured Generation

Types of Vision-Language Models (VLMs)

Where does interaction between modalities happen?

Before Encoder	Chameleon
Within (layers of) Encoder	Llama 3.1
After Encoder	Clip/Llava

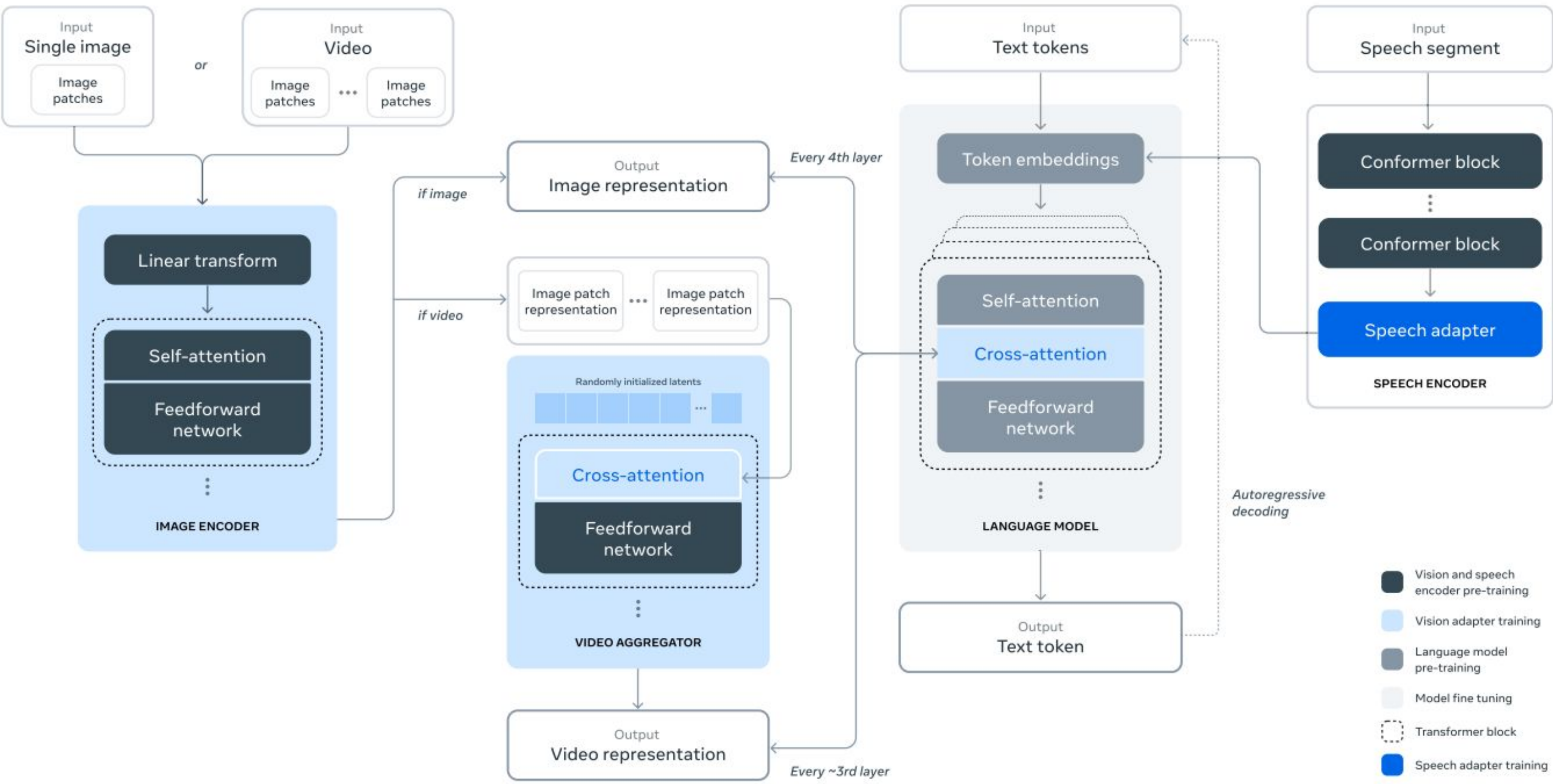
Early-interaction VLM: Chameleon



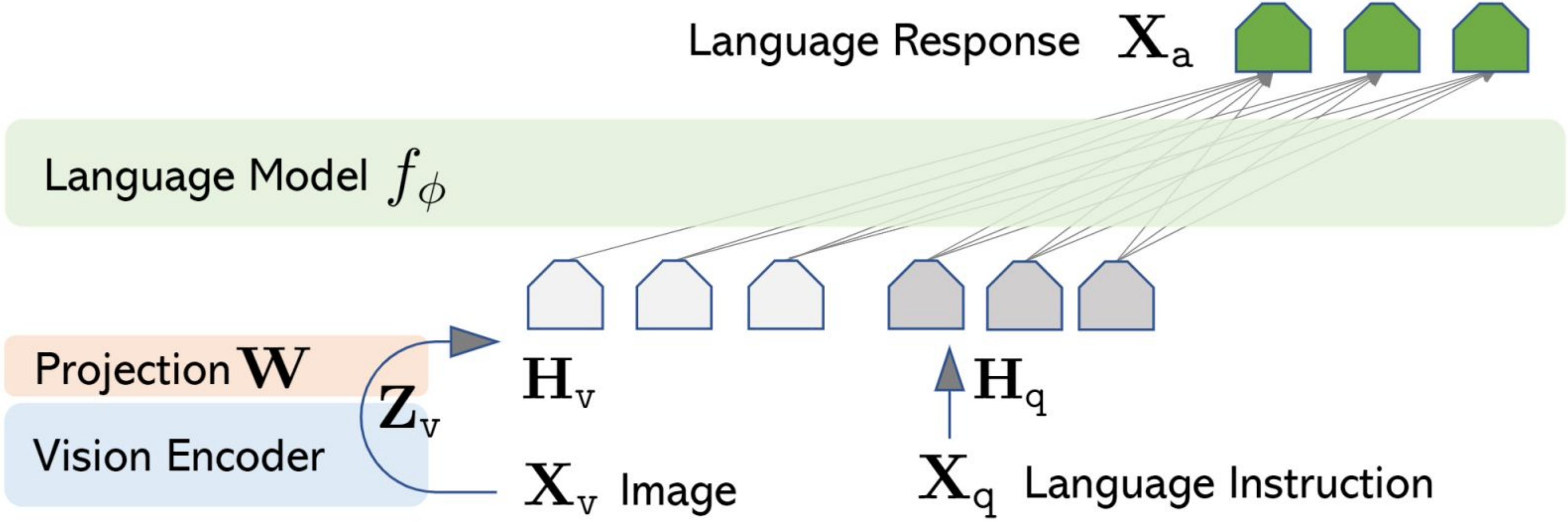
(a) Mixed-Modal Pre-Training

(b) Mixed-Modal Generation

Cross-interaction VLM: Llama 3.1



Late-interaction VLM: Llava



MMFM2: The 2nd Workshop on What is Next in Multimodal Foundation Models?

Tuesday, June 18

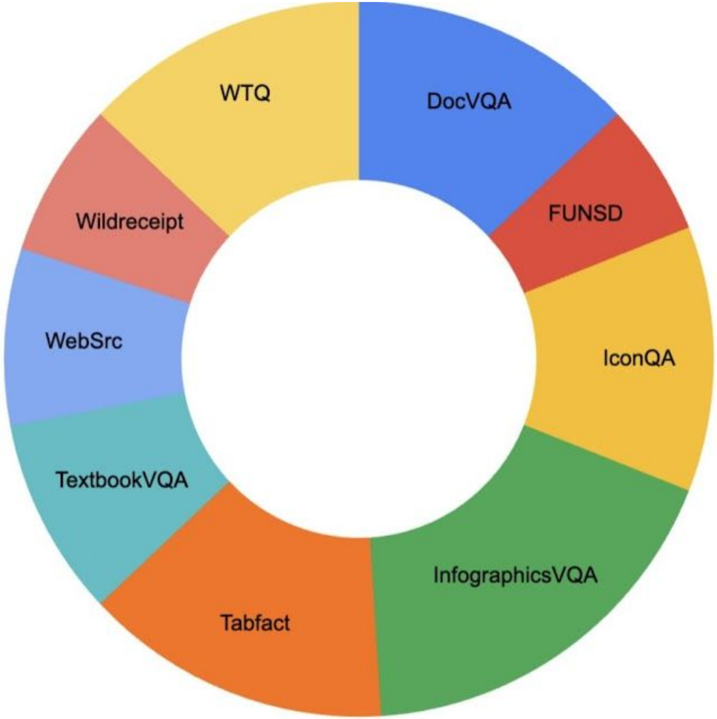
Summit 437-439

CVPR 2024 Workshop, Seattle, WA

Do Multimodal Foundation Models still suck at document understanding tasks?

Spoiler: kinda

Phase 1: 10 public document-understanding datasets



How many domain names are registered per minute?

Phase 2: 3 private test datasets

Phase 2: 3 private test datasets -- (1) MyDoc

Contract Data (Traffic) Report

SUMMARY FOR ORDER # 4074991

Traffic Order #	577569	
Order #	4074991	
Order Descrp	63145406_POL_Candidate_DONALD J TRUMP FOR PRES - S	
Client	AMP - DONALD J TRUMP FOR PRES -	
Start Date	12/18/2023	
End Date	12/31/2023	
# of Weeks	2	
SALES		
ActiveWeeks	2	
AE 1	NCC - SAV - DC	
AE 2		
Agency	AMP - STRATEGIC MEDIA SERVICES	15.00%
RepFirm	NCC	13.00%
Copy Instr ID		
Total Zones	1	
Zones	Savannah Interconnect	
Total Networks	1	
GENERAL COMMENTS		

Created On	12/21/2023 12:07:52 PM
Created By	NCC_Gateway_User
Updated On	12/21/2023 2:18:05 PM
Updated By	Smith, Brogan
Industry	Political-President
REFERENCES	
Primary	
Secondary	
Tertiary	
Quarternary	
TRAFFIC OPTIONS	
Address 1	AMP MEDIA
Address 2	
City, State, Zip	BLOOMFIELD, NJ
Zip	07003
Contact	
Phone	111-111-1111
Avail Tag	
Contract Type	Standard
Copy Group	
Division	
Reference #	

Order Status	Contract Confirmed
Gross $	2104.00
Net $	1514.88
Units	4
Credit Hold	NO
BILLING INFORMATION	
Purchase Order #	
Billing Schedule	EndOfFlight
EDI INFORMATION	
Product	932
Estimate	10954
Submit EDI Invoice?	Submit EDI Invoice

ORDER /INVOICE/TRAFFIC/REPORT NOTES/COMMENTS

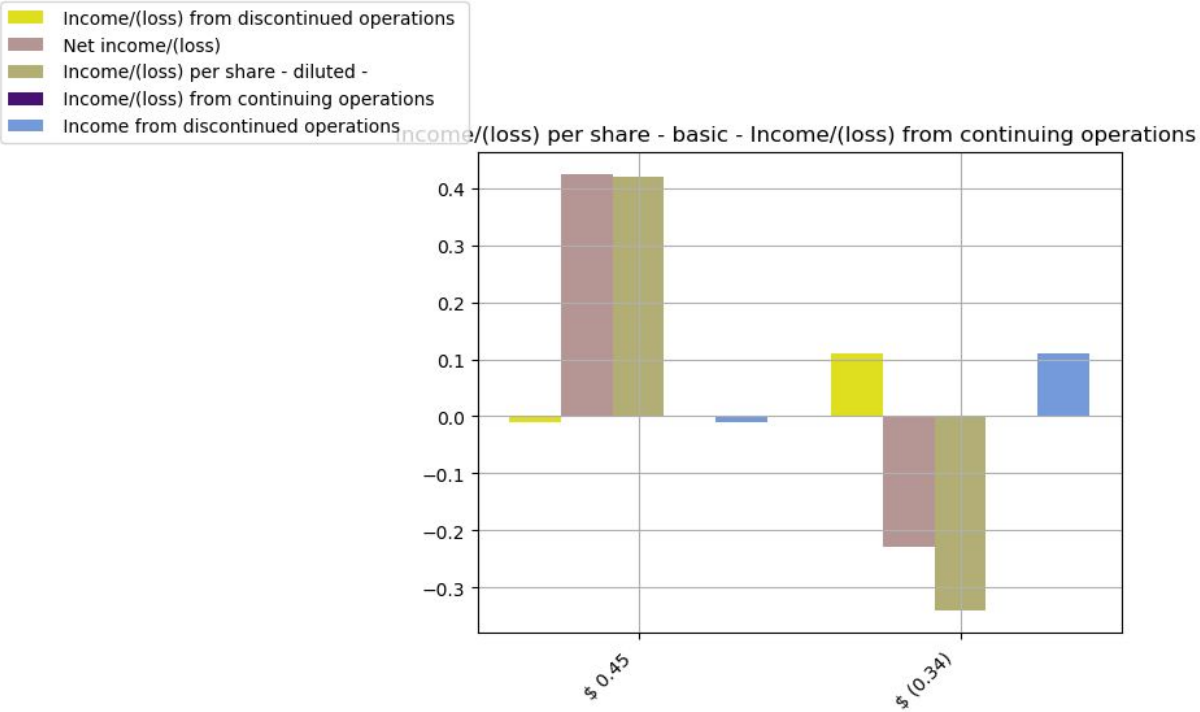
Savannah– PRIORITY CODE: NP=80, IP=74 – SEE KEY ON FCC SITE FOR NETWORK/ZONE INFORMATION

SYSCODE LIST

6996

<image> What is the address 1 in the image?

Phase 2: 3 private test datasets -- (2) MyChart



<image> Can you explain why the income from discontinued operations is (0.01)?

Phase 2: 3 private test datasets -- (3) MyInfographic

<image> Are there any icons or graphics that suggest a particular focus for the data?

My Approach: Multimodal Structured Generation

https://github.com/leloykun/MMFM-Challenge

Context

- I joined < 48 hours before the deadline
- I wasted 24+ hours working with commercial models (which weren't allowed)
- Laptop is 5 years old
- On student budget

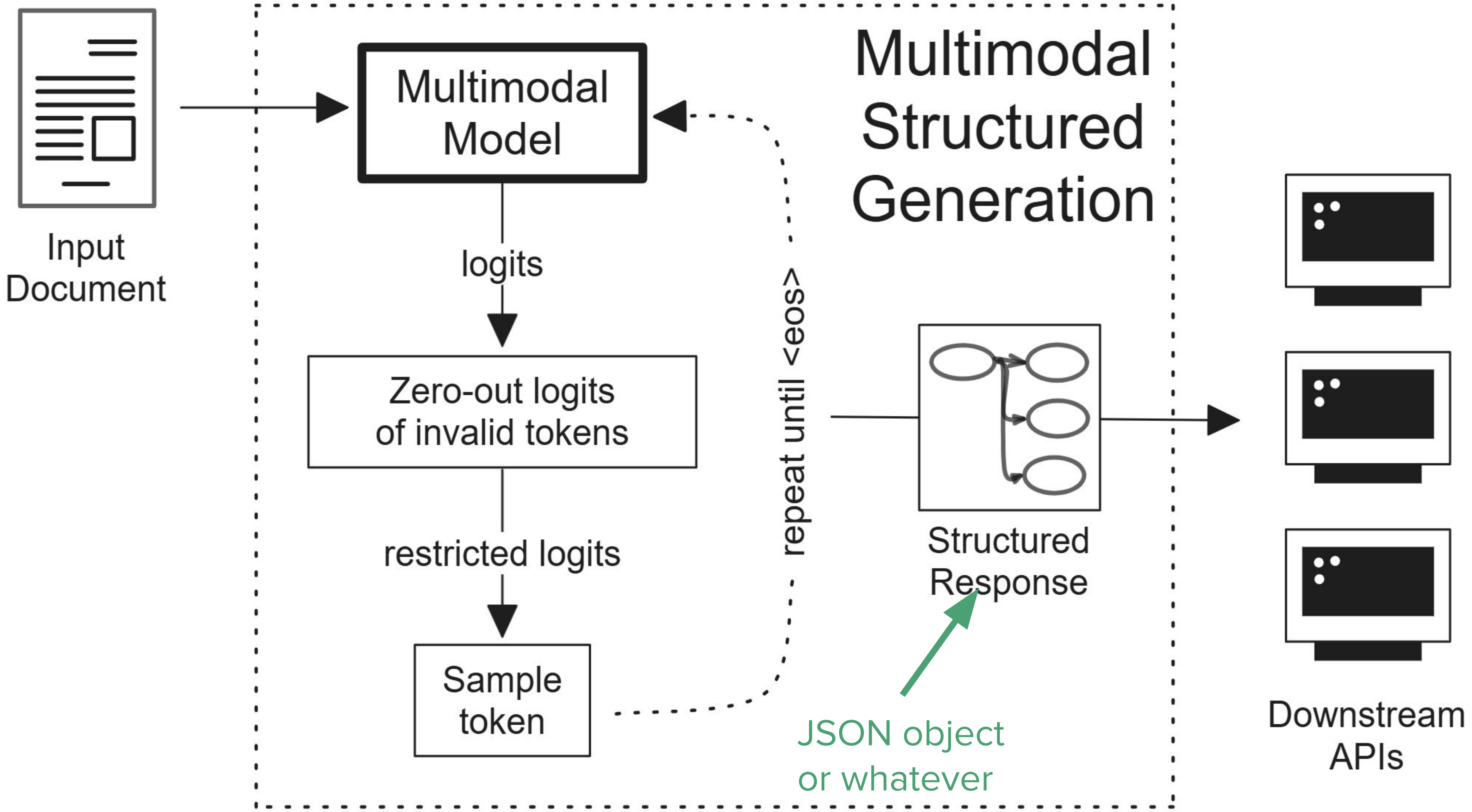

IN TERMS OF GPUs
WE HAVE NO GPUs

What I couldn't do with the constraints

- No Finetuning, because I didn't have *GPU*s
- No Retrieval Augmented Generation (RAG), because I didn't have the *time* to implement it

Yet, I managed to place 2nd in the hidden test set

So, how did I do it?



Input
Document
Multimodal
Model
logits
Zero-out logits
of invalid tokens
restricted logits
Sample
token
repeat until <eos>
Multimodal
Structured
Generation
Structured
Response
JSON object
or whatever
Downstream
APIs

To what end?

To force the models to *reason* before answering!

```
{
    "type": "object",
    "properties": {
        "1_reasoning": {"type": "string"},
        "2_answer": {
            "type": "string",
            "description": "Concise answer to the user question."
        },
    },
    "required": ["1_reasoning", "2_answer"],
}
```

Structured Generation with e.g. Outlines also gives us more control over how the models “think”!

```
{
      "type": "object",
      "properties": {
          "1_reasoning": {
              "type": "string",
              "minLength": 500,
          },
          f"2_{key}": {
              "type": "integer" if key == "page" else "string",
              "description": "The answer, exactly as it appears in the document.",
              "maxLength": 100,
          }
      },
      "required": ["1_reasoning", f"2_{key}"],
}
```

Controlled reasoning!

Hallucination-free outputs!

Folks at .TXT (Outlines) actually beat me to it:

Prompt Efficiency - Using Structured Generation to get 8-shot performance from 1-shot.

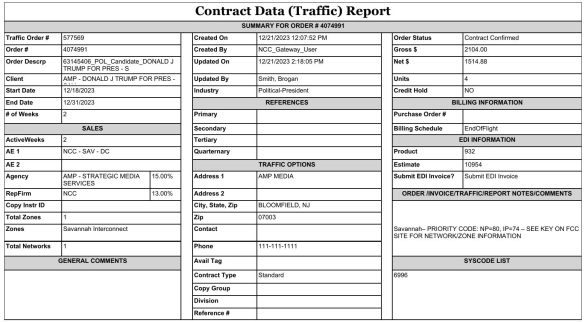
In this post we're going to explore a surprising benefit of structured generation that we've recently come across here at **.txt** we call *"prompt efficiency"*: For few-shot tasks, structured generation with **Outlines** is able to achieve superior performance in as little as **one example** than unstructured is with up to 8. Additionally we observed that 1-shot structured performance remains similar to higher shot structured generation, meaning 1-shot is all that is necessary in many cases for high quality performance. This is useful for a variety of practical reasons:

- **convenience:** For few-shot problems, examples can be difficult to come by and annotating examples that include a "Chain-of-Thought" reasoning step can be *very* time consuming and challenging.
- **speed:** Longer prompts mean more computation, so keeping prompt size smaller means faster inference.
- **context conservation:** Examples easily eat up a lot of context for models with limited context length.

We'll walk through the experiments we've run to show this property of structured generation.

Apr 24, 2024 -- https://blog.dottxt.co/prompt-efficiency.html

Llava-1.6 + Structured Generation performed the best for MyChart & MyInfographic…



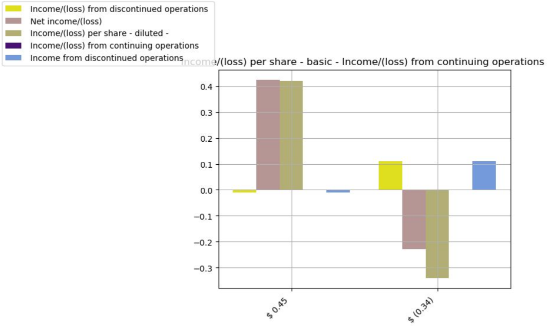
Contract Data (Traffic) Report

SUMMARY FOR ORDER # 4074991

Traffic Order #	577569
Order #	4074991
Order Descrp	63145406_POL_Candidate_DONALD J TRUMP FOR PRES - S
Client	AMP - DONALD J TRUMP FOR PRES -
Start Date	12/18/2023
End Date	12/31/2023
# of Weeks	2
SALES	
ActiveWeeks	2
AE 1	NCC - SAV - DC
AE 2	
Agency	AMP - STRATEGIC MEDIA SERVICES 15.00%
RepFirm	NCC 13.00%
Copy Instr ID	
Total Zones	1
Zones	Savannah Interconnect
Total Networks	1
GENERAL COMMENTS	

Created On	12/21/2023 12:07:52 PM
Created By	NCC_Gateway_User
Updated On	12/21/2023 2:18:05 PM
Updated By	Smith, Brogan
Industry	Political-President
REFERENCES	
Primary	
Secondary	
Tertiary	
Quarternary	
TRAFFIC OPTIONS	
Address 1	AMP MEDIA
Address 2	
City, State, Zip	BLOOMFIELD, NJ
Zip	07003
Contact	
Phone	111-111-1111
Avail Tag	
Contract Type	Standard
Copy Group	
Division	
Reference #	

Order Status	Contract Confirmed
Gross $	2104.00
Net $	1514.88
Units	4
Credit Hold	NO
BILLING INFORMATION	
Purchase Order #	
Billing Schedule	EndOfFlight
EDI INFORMATION	
Product	932
Estimate	10954
Submit EDI Invoice?	Submit EDI Invoice
ORDER /INVOICE/TRAFFIC/REPORT NOTES/COMMENTS	
Savannah– PRIORITY CODE: NP=80, IP=74 – SEE KEY ON FCC SITE FOR NETWORK/ZONE INFORMATION	
SYSCODE LIST	
6996	



MyDoc

MyChart

MyInfographic

(not so much)

✅

✅

For MyDoc, I had to revert to using an LLM…

Contract Data (Traffic) Report

SUMMARY FOR ORDER # 4074991

Traffic Order #	577569	
Order #	4074991	
Order Descrp	63145406_POL_Candidate_DONALD J TRUMP FOR PRES - S	
Client	AMP - DONALD J TRUMP FOR PRES -	
Start Date	12/18/2023	
End Date	12/31/2023	
# of Weeks	2	
SALES		
ActiveWeeks	2	
AE 1	NCC - SAV - DC	
AE 2		
Agency	AMP - STRATEGIC MEDIA SERVICES	15.00%
RepFirm	NCC	13.00%
Copy Instr ID		
Total Zones	1	
Zones	Savannah Interconnect	
Total Networks	1	
GENERAL COMMENTS		

Created On	12/21/2023 12:07:52 PM
Created By	NCC_Gateway_User
Updated On	12/21/2023 2:18:05 PM
Updated By	Smith, Brogan
Industry	Political-President
REFERENCES	
Primary	
Secondary	
Tertiary	
Quarternary	
TRAFFIC OPTIONS	
Address 1	AMP MEDIA
Address 2	
City, State, Zip	BLOOMFIELD, NJ
Zip	07003
Contact	
Phone	111-111-1111
Avail Tag	
Contract Type	Standard
Copy Group	
Division	
Reference #	

Order Status	Contract Confirmed
Gross $	2104.00
Net $	1514.88
Units	4
Credit Hold	NO
BILLING INFORMATION	
Purchase Order #	
Billing Schedule	EndOfFlight
EDI INFORMATION	
Product	932
Estimate	10954
Submit EDI Invoice?	Submit EDI Invoice
ORDER /INVOICE/TRAFFIC/REPORT NOTES/COMMENTS	
Savannah– PRIORITY CODE: NP=80, IP=74 – SEE KEY ON FCC SITE FOR NETWORK/ZONE INFORMATION	
SYSCODE LIST	
6996	

MyDoc

~~Vision Language Model~~
+ Structured Generation

✅ Large Language Model + Structured Generation

https://huggingface.co/leloy/Nous-Hermes-2-Pro-Docile-RASG-1ShotRetrieval-StructuredPrompt

Why? A brief review of literature…

There are three modalities of information you can extract from a document:

- Textual Information
- Visual Information (“what stuff are in the doc?”)
- Layout Information (“where are the stuff in the doc?”)

DocLLM: A layout-aware generative language model for multimodal document understanding

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, Xiaomo Liu

Enterprise documents such as forms, invoices, receipts, reports, contracts, and other similar records, often carry rich semantics at the intersection of textual and spatial modalities. The visual cues offered by their complex layouts play a crucial role in comprehending these documents effectively. In this paper, we present DocLLM, a lightweight extension to traditional large language models (LLMs) for reasoning over visual documents, taking into account both textual semantics and spatial layout. Our model differs from existing multimodal LLMs by avoiding expensive image encoders and focuses exclusively on bounding box information to incorporate the spatial layout structure. Specifically, the cross-alignment between text and spatial modalities is captured by decomposing the attention mechanism in classical transformers to a set of disentangled matrices. Furthermore, we devise a pre-training objective that learns to infill text segments. This approach allows us to address irregular layouts and heterogeneous content frequently encountered in visual documents. The pre-trained model is fine-tuned using a large-scale instruction dataset, covering four core document intelligence tasks. We demonstrate that our solution outperforms SotA LLMs on 14 out of 16 datasets across all tasks, and generalizes well to 4 out of 5 previously unseen datasets.

DocLLM has shown that **removing the vision encoder** and treating bounding boxes (i.e. layout information) as its own modality **does not harm performance** on doc understanding tasks...

There are three modalities of information you can extract from a document:

- Textual Information
- ~~Visual Information (“what stuff are in the doc?”)~~
- Layout Information (“where are the stuff in the doc?”)

Retrieval Augmented Structured Generation: Business Document Information Extraction As Tool Use

Franz Louis Cesista, Rui Aguiar, Jason Kim, Paolo Acilo

Business Document Information Extraction (BDIE) is the problem of transforming a blob of unstructured information (raw text, scanned documents, etc.) into a structured format that downstream systems can parse and use. It has two main tasks: Key-Information Extraction (KIE) and Line Items Recognition (LIR). In this paper, we argue that BDIE is best modeled as a Tool Use problem, where the tools are these downstream systems. We then present Retrieval Augmented Structured Generation (RASG), a novel general framework for BDIE that achieves state of the art (SOTA) results on both KIE and LIR tasks on BDIE benchmarks.
The contributions of this paper are threefold: (1) We show, with ablation benchmarks, that Large Language Models (LLMs) with RASG are already competitive with or surpasses current SOTA Large Multimodal Models (LMMs) without RASG on BDIE benchmarks. (2) We propose a new metric class for Line Items Recognition, General Line Items Recognition Metric (GLIRM), that is more aligned with practical BDIE use cases compared to existing metrics, such as ANLS*, DocILE, and GriTS. (3) We provide a heuristic algorithm for backcalculating bounding boxes of predicted line items and tables without the need for vision encoders. Finally, we claim that, while LMMs might sometimes offer marginal performance benefits, LLMs + RASG is oftentimes superior given real-world applications and constraints of BDIE.

Our previous work has shown that **removing layout information does not harm performance** on the Key-Information Extraction task either...

There are three modalities of information you can extract from a document:

- Textual Information
- ~~Visual Information (“what stuff are in the doc?”)~~
- ~~Layout Information (“where are the stuff in the doc?”)~~

(at least for Key-Information Extraction)

Final Results

Results for hidden test set

Method	Team	Acc
GPT4o	-	0.703
NBG-VL	**xray1112247**	0.565
Multimodal Structured Generation	**leloy**	0.505
Strong-DocFVLM	**necla**	0.470
Table Transformer	**MalumaDev**	0.293
LLaVA 1.6 13B	-	0.197
LLaVA 1.6 7B	-	0.184
LLaVA 1.5 13B finetuned on Phase-1 data	-	0.182
MoE LLaVA	**UTokyo-Nakayamalab**	0.173
LLaVA 1.5 13B	-	0.165
LLaVA 1.5 7B	-	0.144

mine

also mine

Results for hidden test set by task

Task	Best Approach	Score
MyDoc	Nous Hermes 2 Pro (LLM) + Structured Generation	62.25%
MyChart	LLava-1.6 (VLM) + Structured Generation	4.50%
MyInfographic	LLava-1.6 (VLM) + Structured Generation	60.98%

vs. 21% with LLava-1.6

Why did an LLM outperform a VLM on the MyDoc dataset?

Hypothesis 1: Visual and layout information are simply not important for Key-Information Extraction

Model	Key-Information Extraction F1 Score	Line Items Recognition GLIRM-F1 [2]
GPT-3.5	34.17%	28.31%
+ 1-Shot Retrieval	+ 22.08%	+ 20.67%
+ Supervised Finetuning	+ 22.31%	+ 17.73%
+ Structured Prompting	+ 4.96%	+ 19.42%
Hermes 2 Pro - Mistral 7B	13.55%	4.69%
+ 1-Shot Retrieval	+ 36.87%	+ 40.55%
+ Supervised Finetuning	+ 17.71%	+ 13.53%
+ Structured Prompting	+ 0.63%	+ 10.30%

* Benchmarks results ablating three components of Retrieval Augmented Structured Generation on Key-Information Extraction (KIE) & Line Items Recognition (LIR) tasks on the DocILE dataset [13]: (1) Retrieval Augmented Generation [4], (2) Supervised Finetuning, & (3) Structured Prompting [5]. Structured Generation was not included in the ablation benchmarks as it is a necessary component of RASG to ensure that the outputs are parseable by downstream APIs [3]. Results show that adding Structured Prompting, i.e. infusing layout information to the text prompt, only adds a marginal increase in performance.

Retrieval Augmented Structured Generation: Business Document Information Extraction as Tool Use
arXiv: 2405.20245

Hypothesis 2: LLMs can already infer the location of the words in the image from their index in the prompt

Transformer Language Models without Positional Encodings Still Learn Positional Information

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, Omer Levy

PDF

Cite

Search

Abstract

Causal transformer language models (LMs), such as GPT-3, typically require some form of positional encoding, such as positional embeddings. However, we show that LMs without any explicit positional encoding are still competitive with standard models and that this phenomenon is robust across different datasets, model sizes, and sequence lengths.Probing experiments reveal that such models acquire an implicit notion of absolute positions throughout the network, effectively compensating for the missing information.We conjecture that causal attention enables the model to infer the number of predecessors that each token can attend to, thereby approximating its absolute position.Our findings indicate that causal LMs might derive positional awareness not only from the explicit positioning mechanism but also from the effects of the causal mask.

What if this also applies in the 2D case?

Hypothesis 3: The vision-language models are simply at overcapacity.

- We already train LLMs to their full capacity according to Neural Scaling laws
- Grafting the Vision Encoders pushes them over the edge

Hypothesis 4: We are not using enough image tokens

# Tokens Per Grid	Approach	TextVQA	AI2D	ChartQA	DocVQA	MMBench	POPE	ScienceQA	MMMU
576	*SS*	64.53	64.83	59.28	75.40	66.58	87.02	72.29	34.3
	M^3	63.13	66.71	58.96	72.61	67.96	87.20	72.46	34.0
144	*SS*	62.16	65.77	55.28	67.69	67.78	87.66	72.15	36.4
	M^3	62.61	68.07	57.04	66.48	69.50	87.67	72.32	36.1
36	*SS*	58.15	65.90	45.40	56.89	67.01	86.75	71.87	36.2
	M^3	58.71	67.36	50.24	55.94	68.56	87.29	72.11	36.8
9	*SS*	50.95	65.06	37.76	44.21	65.29	85.62	72.37	36.8
	M^3	51.97	66.77	42.00	43.52	67.35	86.17	71.85	35.2
1	*SS*	38.39	63.76	28.96	33.11	61.43	82.83	72.32	35.3
	M^3	38.92	64.57	31.04	31.63	62.97	83.38	71.19	34.8
Oracle	# Tokens	31.39	11.54	41.78	64.09	8.90	6.08	7.43	22.85
	Performance	70.51	76.36	70.76	81.73	74.35	94.29	76.07	50.44

Figure 2: Comparison of approaches with the SS baseline and Matryoshka Multimodal Models (M^3) across various benchmarks under LLaVA-NeXT [28]. Here # Tokens denotes the number of visual tokens per image grid in LLaVA-NeXT. SS denotes the baseline model trained with a Specific Scale of visual tokens. M^3 is at least as good as SS, while performing better on tasks such as TextVQA, ChartQA, and MMBench. Oracle denotes the case where the best tradeoff between visual tokens and performance is picked.

Matryoshka Multimodal Models, arXiv: 2405.17430

Document Understanding requires MORE tokens

Demo: Interleaved Multimodal Structured Generation

github.com/leloykun/mmsg