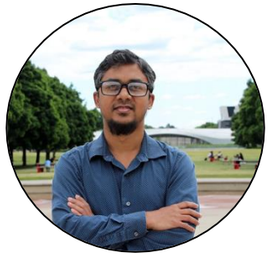
EACL 2026
RABAT · MOROCCO
Mars · March 24-29, 2026 · مارس

Multimodal Large Language Models for Human-AI Interaction

Half-day Tutorial

Shafiq Joty
Salesforce



Enamul Hoque
York University

Ahmed Masry
York University

Spandana Gella
ServiceNow

Samira Ebrahimi Kahou
University of Calgary

mllm4haii.github.io

Tutorial Resources

Multimodal Large Language Models for Human-AI Interaction

Foundations, Agents, and Inclusive Applications

Half-day tutorial at EACL 2026

Overview

Multimodal large language models (MLLMs) are redefining how humans communicate and collaborate with machines. They extend the capabilities of text-based LLMs to perceive, reason, and act across text, images, charts, forms, and graphical user interfaces (GUIs). These models are now capable of answering questions about charts, summarizing infographics, operating software through natural language, and supporting multilingual and accessible visualization.

This tutorial offers a concise, three-hour introduction to the foundations, agentic capabilities, and inclusive applications of MLLMs, with a focus on visually grounded and interactive language tasks. We will cover core architectural designs (encoders, connectors, fusion and decoding mechanisms), multimodal alignment and learning strategies, and reasoning techniques for structured visuals such as charts, forms, and infographics. The tutorial then examines multimodal and conversational agents that perform dialogue-driven reasoning and co-creative analysis in graphical user interfaces. We conclude with discussions on accessibility, multilingual communication, responsible deployment, and future challenges in building human-centered multimodal AI.

Schedule & Materials

Part I: Foundations of Multimodal Large Language Models [45 mins]

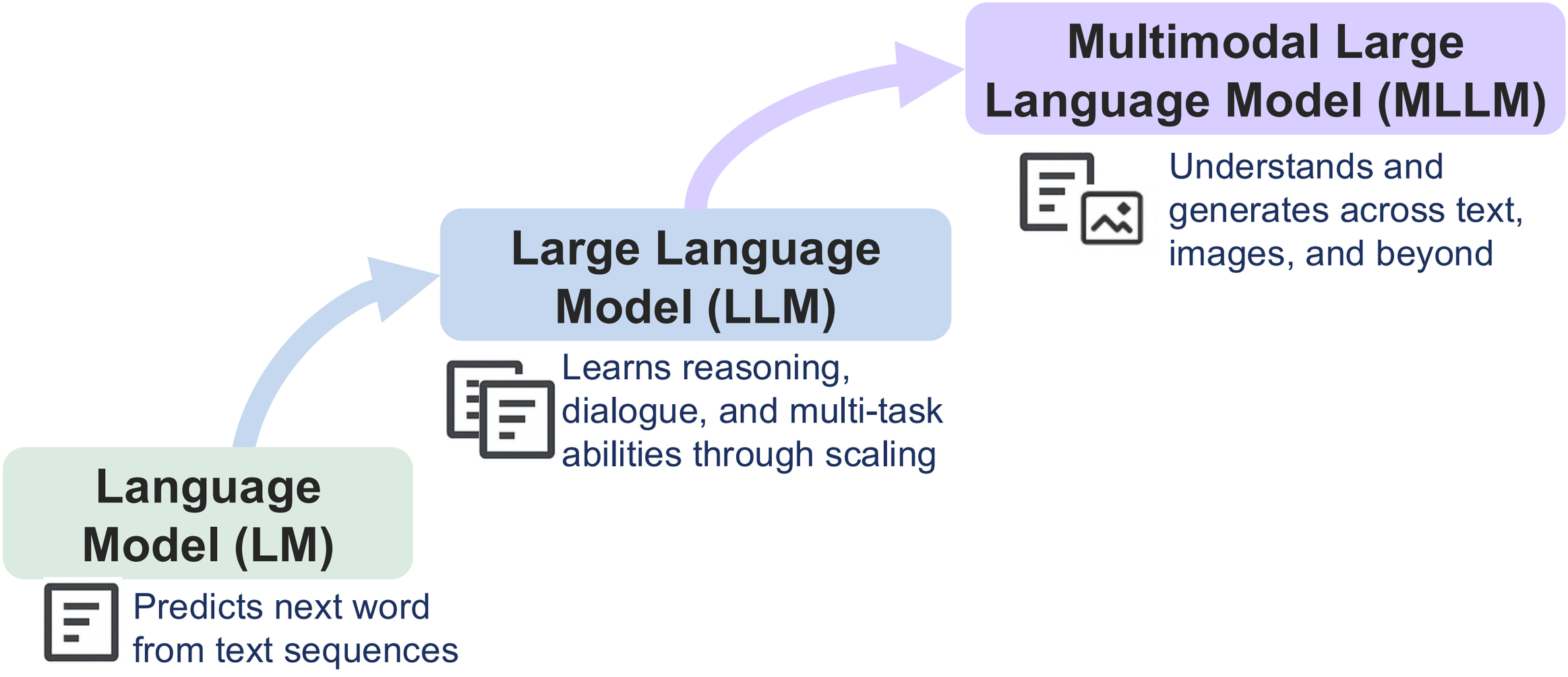
- **From Text to Multimodality:** evolution of LLMs from text-only to multimodal reasoning.
- **Architectural Principles:**
  - Vision encoders (ViT, Swin, CLIP, SigLIP) and language decoders (Llama, Qwen).
  - Connector modules for cross-modal fusion (linear projections, adapters, gated tokens).
  - Unified multimodal architectures: Flamingo, Kosmos, LLaVa, Pali, QwenVL.
- **Alignment and Reinforcement Learning:**
  - Instruction tuning and multimodal alignment.
  - Reinforcement learning for grounding and reward optimization: RLHF, GRPO, and visual reward shaping.
  - Case studies: ChartGemma (visual instruction tuning), BigCharts-R1 (RL for charts).

mllm4haii.github.io

- Tutorial Overview
- Slides
- Reading materials

The Rise of Multimodal AI

- Large language models are evolving into **multimodal systems** that understand **text, images, charts, documents, and GUIs**.



Applications & Impact

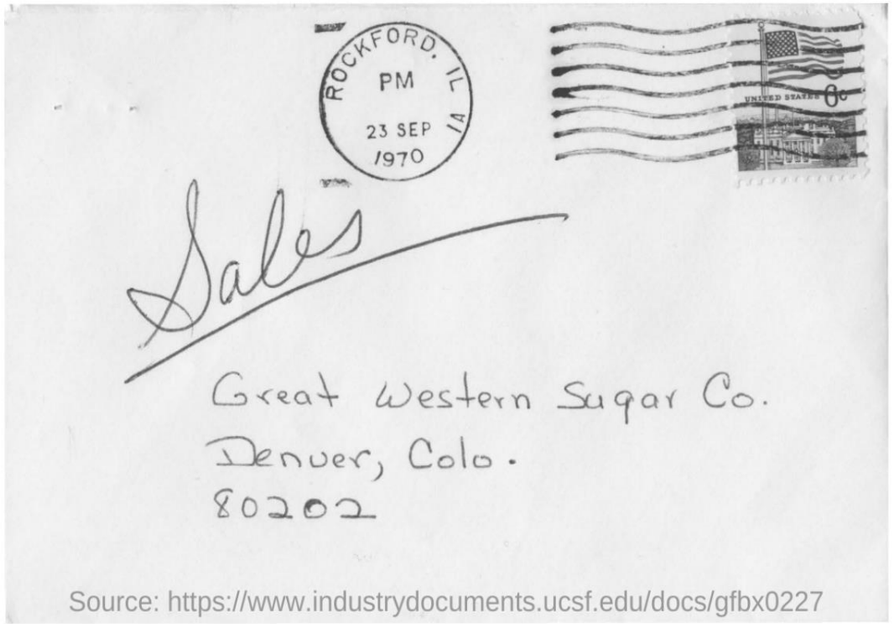
What Can Multimodal AI Enable?

- **Visual reasoning**
  - chart understanding, document QA, infographic analysis
- **Human–AI collaboration**
  - conversational data analysis
  - GUI and software interaction
- **Data communication**
  - automated chart explanation and storytelling
- **Inclusive AI**
  - accessible visualizations (alt-text, narration)
  - multilingual visual information systems

Example Tasks

How many slices of pizza are there?
Is this a vegetarian pizza?

Visual Question answering (Antol et al., ICCV 2015)



Source: https://www.industrydocuments.ucsf.edu/docs/gfbx0227

Q: Mention the ZIP code written?
A: 80202

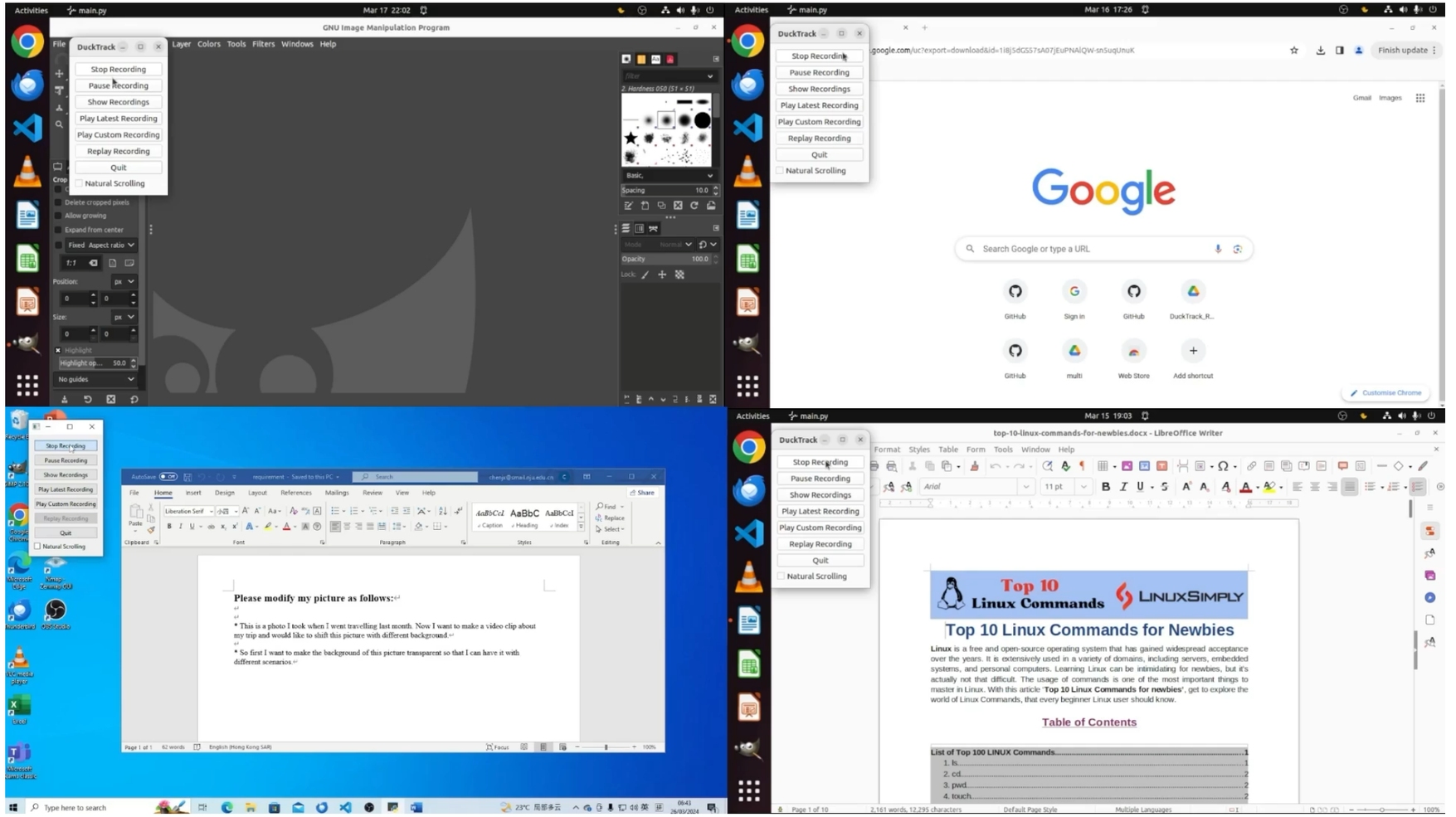
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970

Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

DocVQA (Mathew et al. 2021)

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. Vqa: Visual question answering. InProceedings of the IEEE international conference on computer vision 2015 (pp. 2425-2433).
Mathew, Minesh, Dimosthenis Karatzas, and C. V. Jawahar. "Docvqa: A dataset for vqa on document images." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

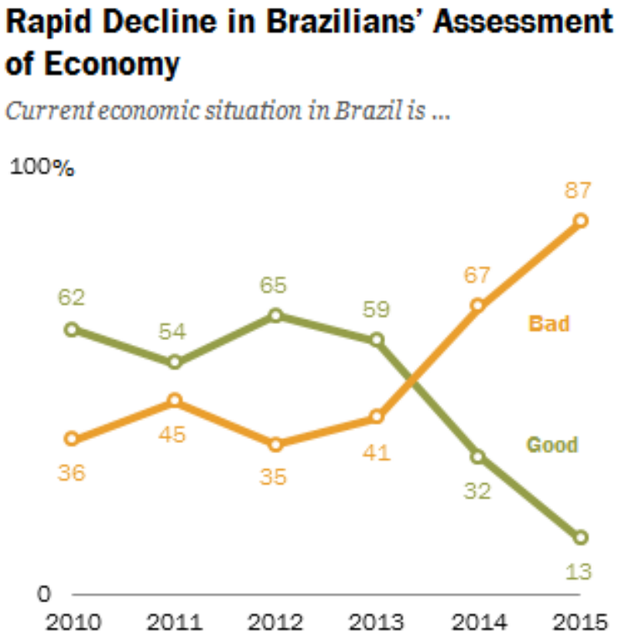
Example Tasks



OSworld (Tianbao et al., Neurips 2024)

Xie T, Zhang D, Chen J, Li X, Zhao S, Cao R, Hua TJ, Cheng Z, Shin D, Lei F, Liu Y. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems. 2024 Dec 16;37:52040-94.

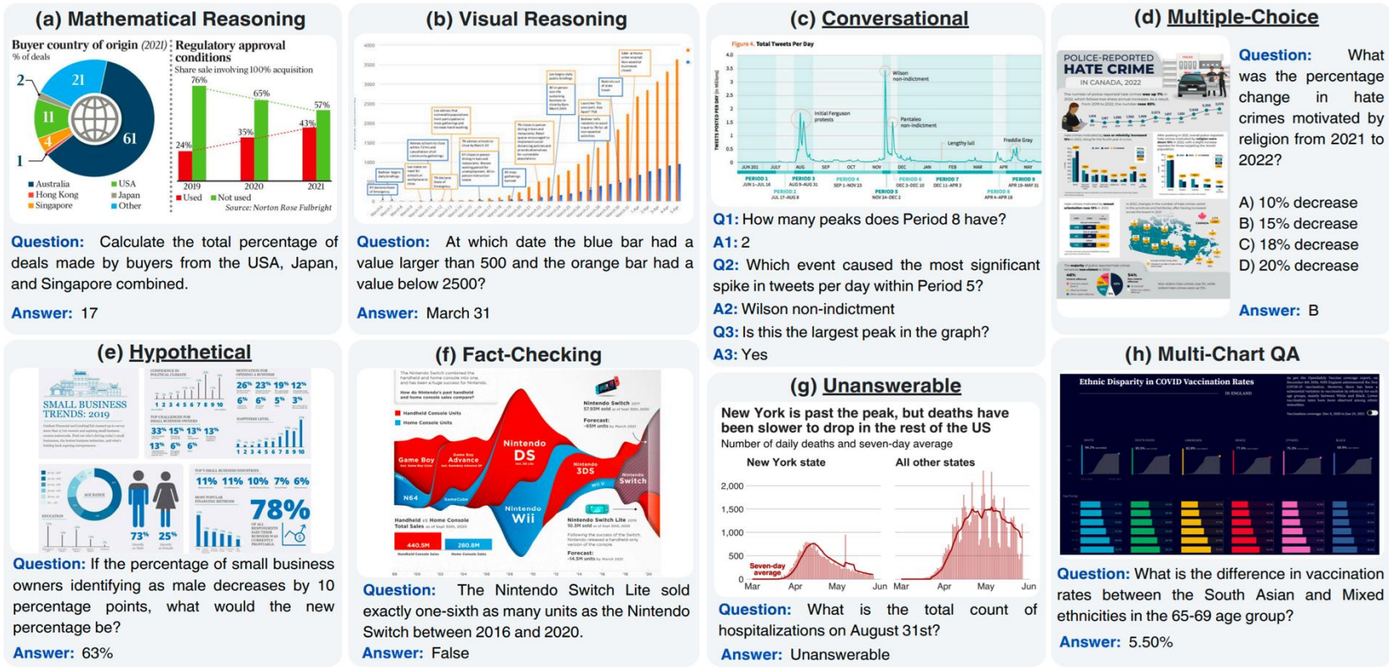
Example Tasks



Question: Which year has the most divergent opinions about Brazil's economy?
Answer: 2015

Question: What is the peak of the orange line?
Answer: 87

ChartQA (Masry et al., ACL 2022)



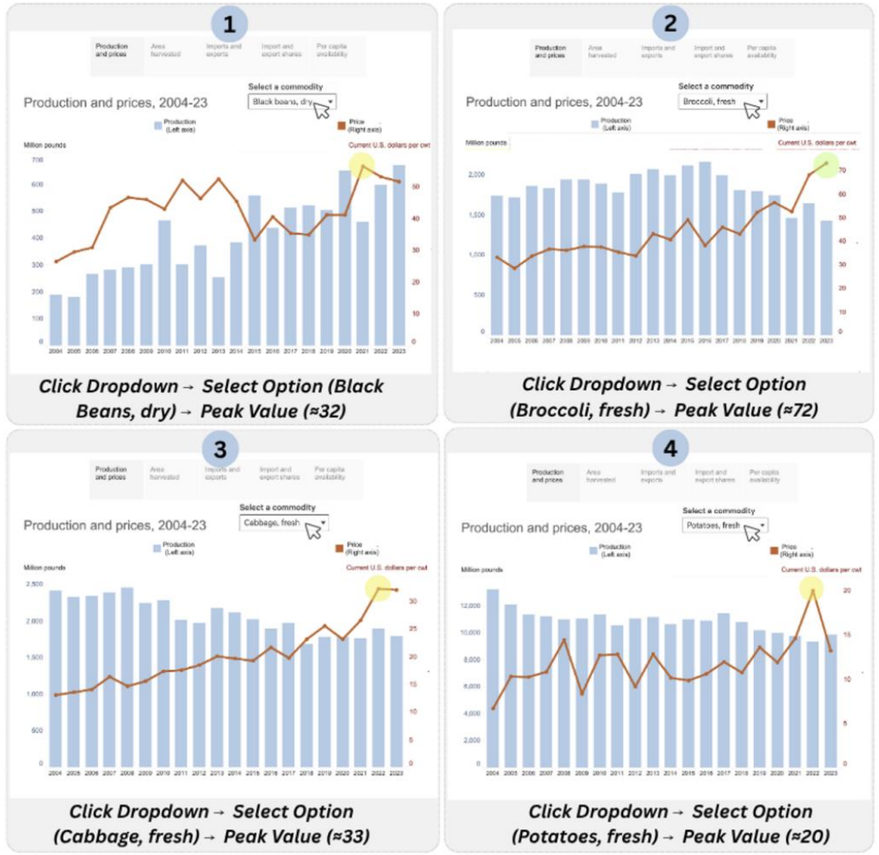
ChartQAPro (Masry et al., ACL 2025)

Masry A., Long d. X., Tan j. Q., Joty S. R., Hoque E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning, ACL 2022.

Masry, A., Islam, M.S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M.T.R., Rahman, M., Rahman, S., Shahmohammadi, M., Thakkar, M., Parvez, M.R., Hoque, E., Joty, S.: ChartQAPro: A more diverse and challenging benchmark for chart question answering, ACL 2025.

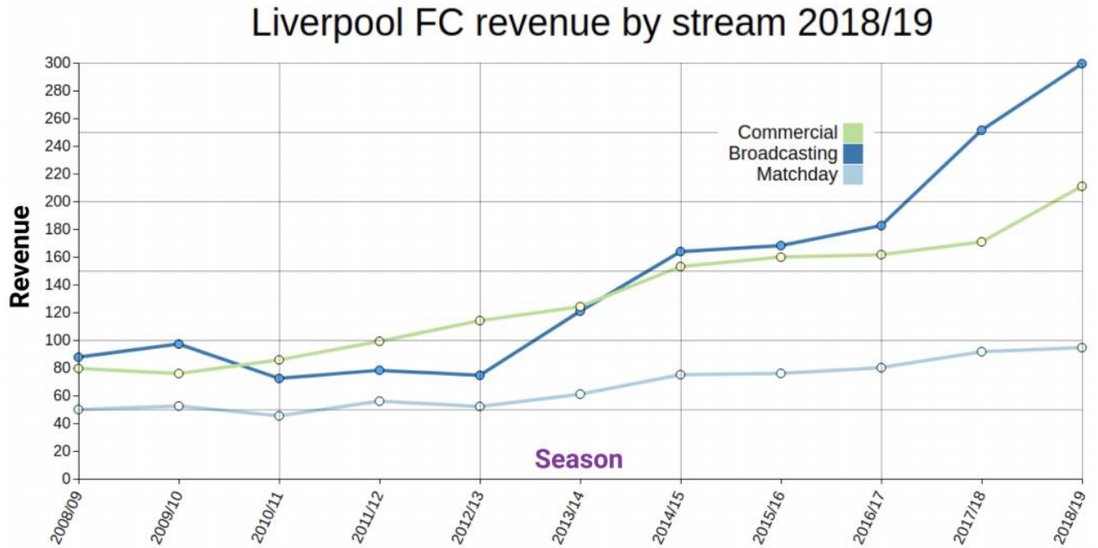
Example Tasks

Q: Among Black Beans (dry), Broccoli (fresh), Cabbage (fresh), and Potatoes (fresh), which crop had the highest peak price from 2004 to 2023?



Answer: Broccoli, fresh

DashboardQA (Kartha et al., EACL 2026)



Broadcasting is the largest source of revenue for Liverpool FC. In 2018/2019 , the club earned approximately 299.3 million euros from broadcasting, more than triple than in 2010/2011 . The second biggest revenue stream is the commercial one.

Chart-to-text (Kantharaj et al., ACL 2022)

Kartha et al, DASHBOARDQA: Benchmarking Multimodal Agents for Question Answering on Interactive Dashboards, ECAL 2026

Shankar K., Rixie tiffany ko l., Xiang l., Ahmed M., Megh T., Enamul H., Shafiq J.: Chart-to-text: A largescale benchmark for chart summarization. In proc ACL, 2022.

Visualization Generation

Query: What will be the next two days' closing prices for Apple stock using a 3-day Simple Moving Average? Also plot the trend in Apple's closing prices over the past 15 trading days.

Data Table (38 by 2)

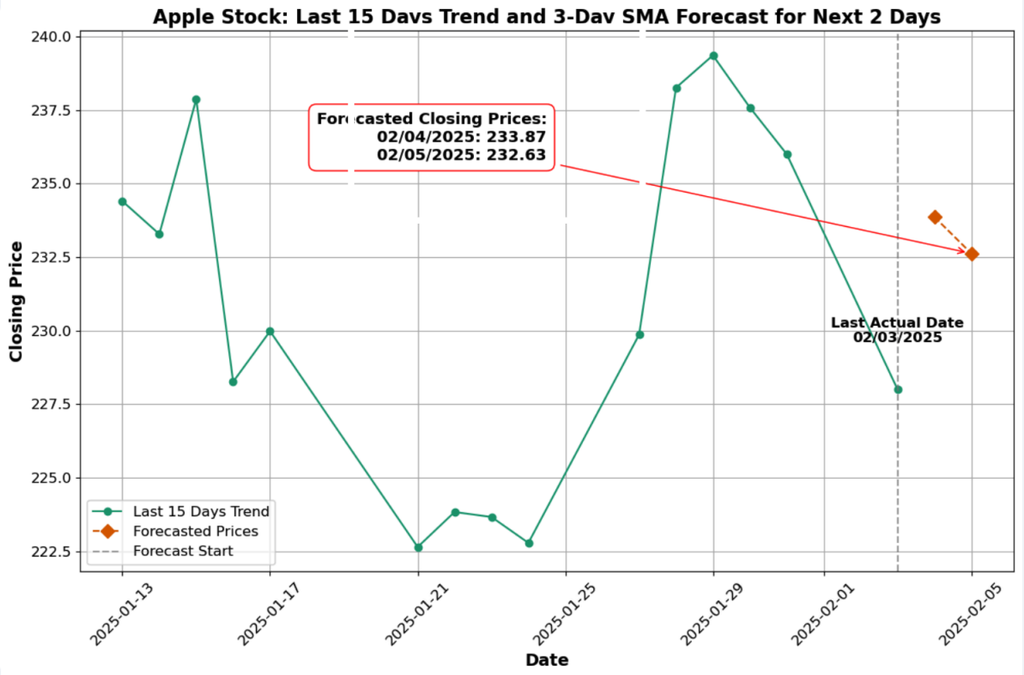
Date	Closing Price
01/13/2025	234.40
....	....
02/03/2025	228.01

Code

Answer: **233.87, 232.63**

Line Chart

```
# Import
import matplotlib.pyplot as plt

# Data information
dates = ["1/1/2025", "1/2/2025",...]
prices = [230, 235, ...]

# Plot stock trend
plt.plot(dates, prices, marker='o', \
    linestyle='-', label='Stock Trend')

# Annotation to the left
plt.annotate("Forecasted Prices"

# X and Y labels
plt.xlabel("Date")
plt.ylabel("Price")

# Add title and legend
plt.title("Stock Price Trend")
plt.legend()

# Show plot
plt.show()
```



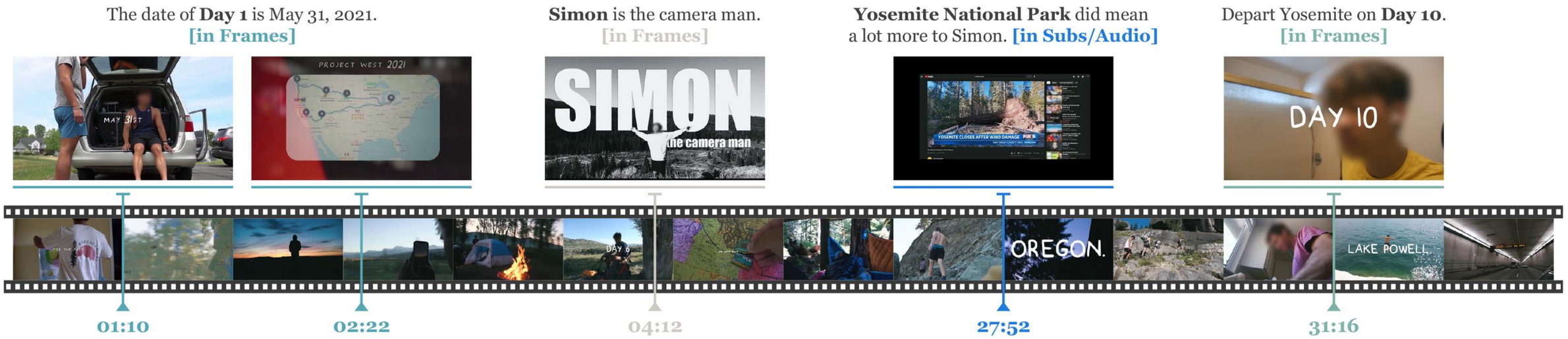
Mizanur Rahman, Md Tahmid Rahman Laskar, Shafiq Joty, Enamul Hoque, Text2Vis: A Challenging and Diverse Benchmark for Generating Multimodal Visualizations from Text, EMNLP 2025.

Example Tasks

Video-MME

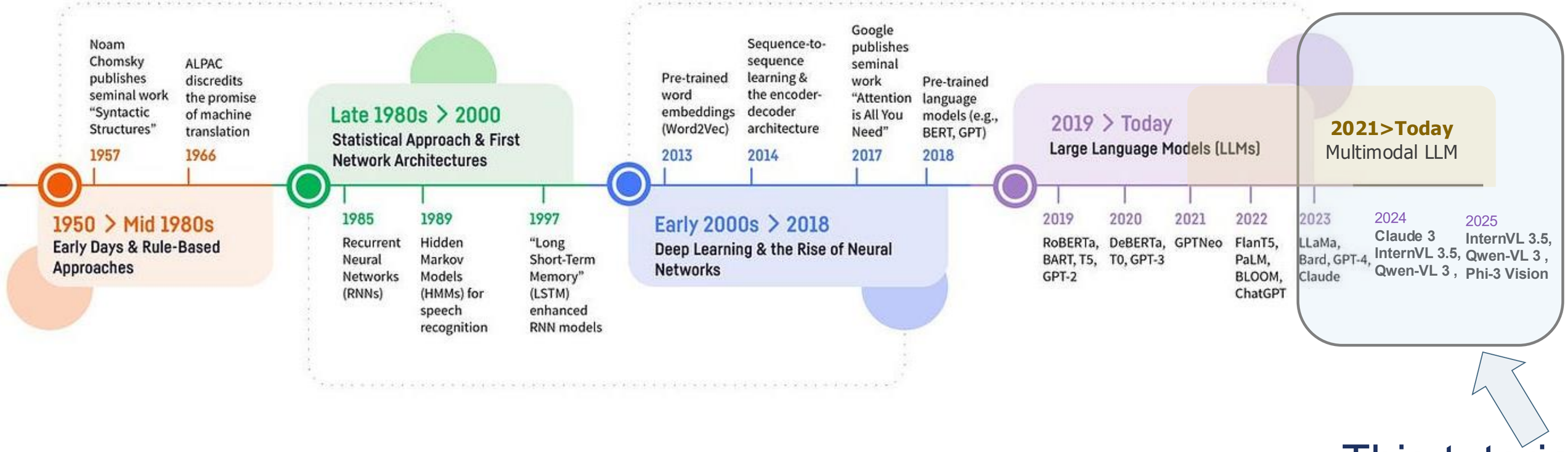
On what date did the individual in the video leave a place that Simon thought was very important to him?

A. May 31, 2022. B. June 9, 2021. C. May 9, 2021. D. June 31, 2021.



Fu et al., Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis
https://arxiv.org/pdf/2405.21075

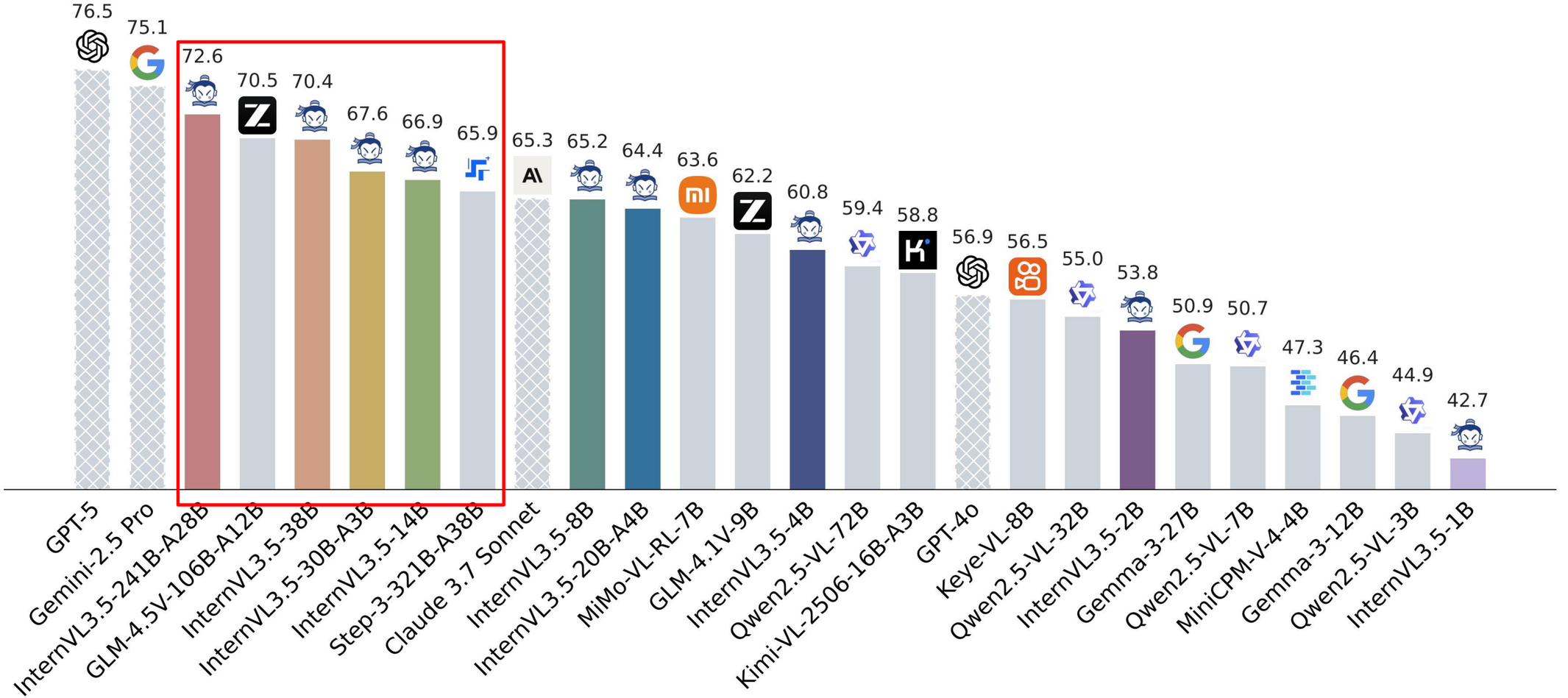
Evolution in LLMs & MLLMs



This tutorial

https://ai.plainenglish.io/evolution-of-language-representation-techniques-a-journey-from-bow-to-gpt-4b49b93e59bd

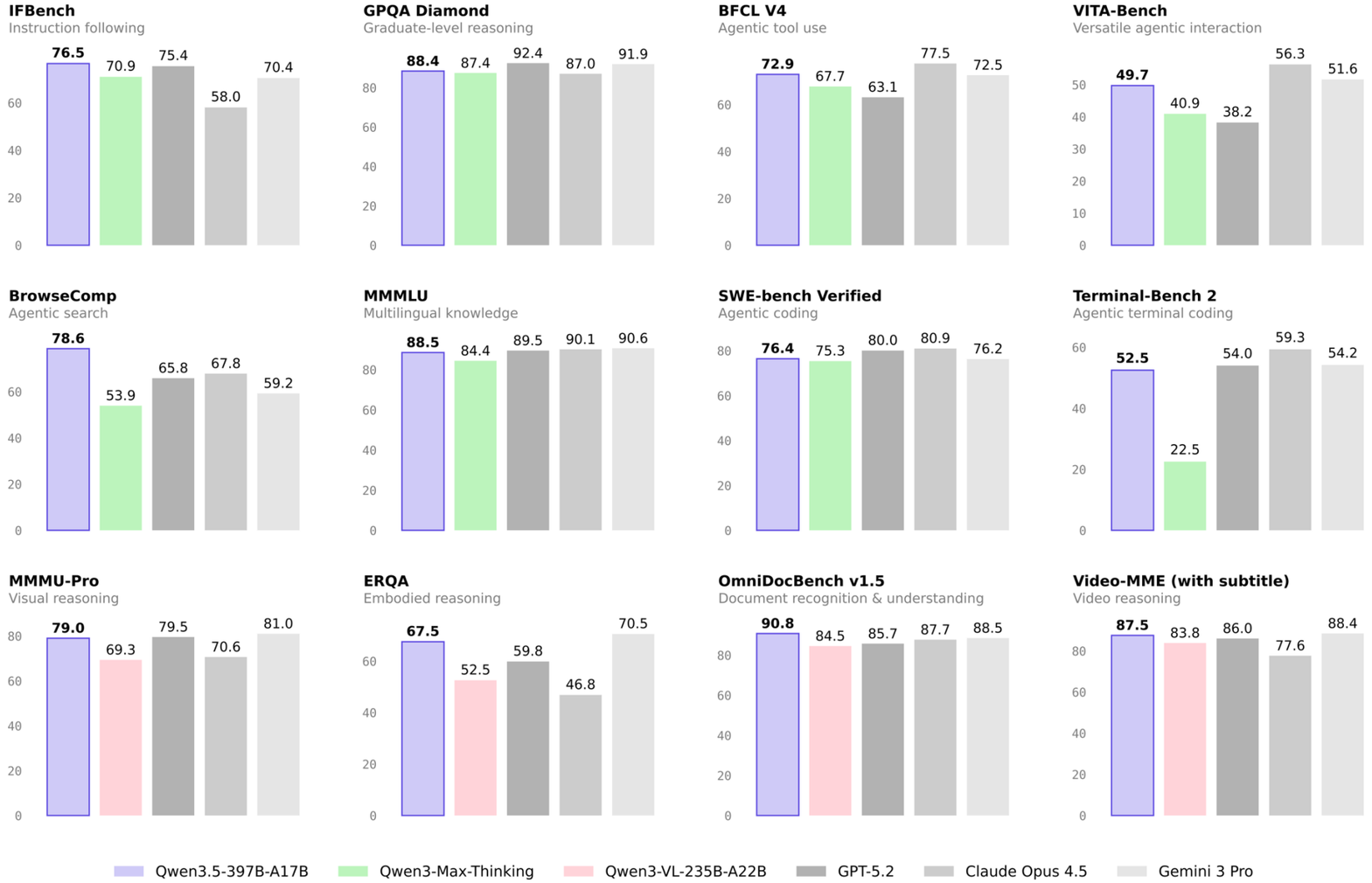
Evolution in LLMs & MLLMs



The performance gap between open and closed models is narrowing

Wang et al., InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025

Evolution in LLMs & MLLMs



Qwen Performance: https://qwen.ai/blog?id=qwen3.5

Key Research Questions

1. What are the **architectural** and **learning principles** that make multimodal large language models **effective** for grounded reasoning?
2. How can language interfaces **bridge** text and vision?
3. How can **AI agents interact** with software and visual environments?
4. How can **agents collaborate** with humans on visual analysis?
5. How do we ensure **accessibility, fairness, and responsible** deployment?

1. FOUNDATIONS OF MLLMs

Evolution of LLMs to multimodal models; architectures, training, and alignment

2. MULTIMODAL REASONING

Datasets, benchmarks, and techniques for reasoning over visual documents

3. HUMAN–AI INTERACTION

Multimodal agents, GUI grounding, and interactive data analysis.

4. RESPONSIBLE & INCLUSIVE MULTIMODAL AI

Accessibility, multilingual understanding, fairness, and hallucination risks

Future Challenges & Outlook

1. FOUNDATIONS OF MLLMs

Evolution of LLMs to multimodal models; architectures, training, and alignment

2. MULTIMODAL REASONING

Datasets, benchmarks, and techniques for reasoning over visual documents

3. HUMAN–AI INTERACTION

Multimodal agents, GUI grounding, and interactive data analysis.

4. RESPONSIBLE & INCLUSIVE AI

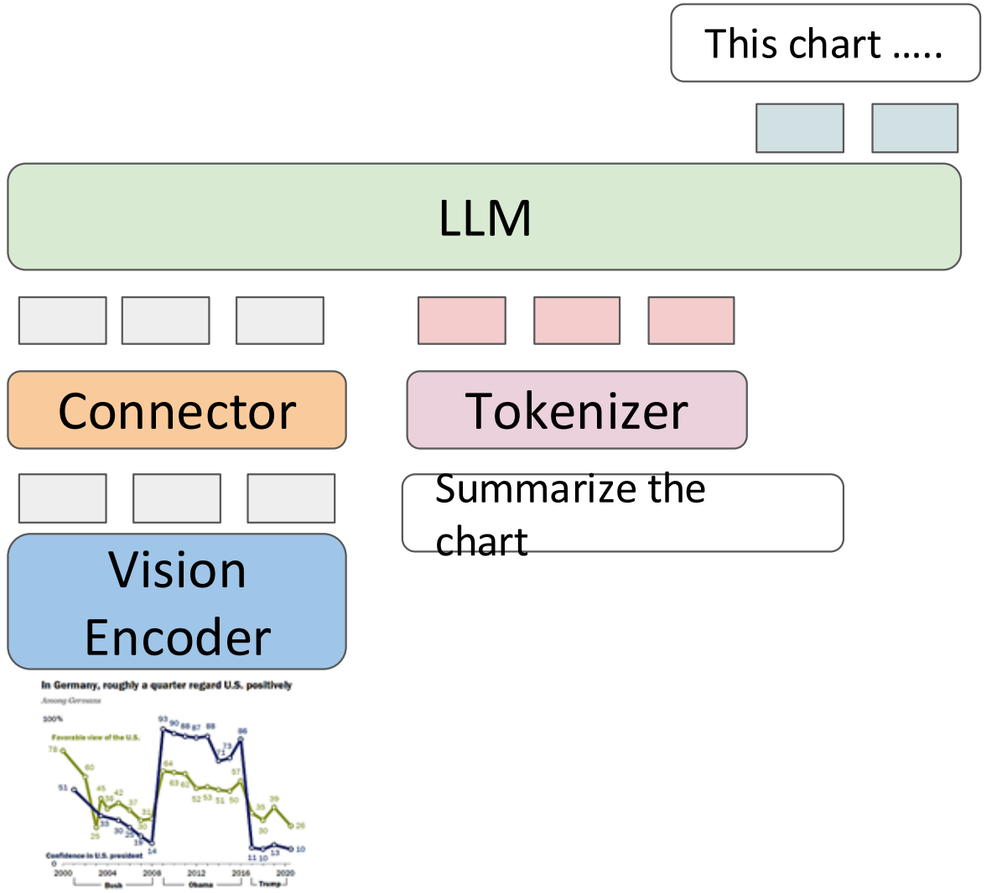
Accessibility, multilingual understanding, fairness, and hallucination risks

Future Challenges & Outlook

Architecture: Vision-Language Models

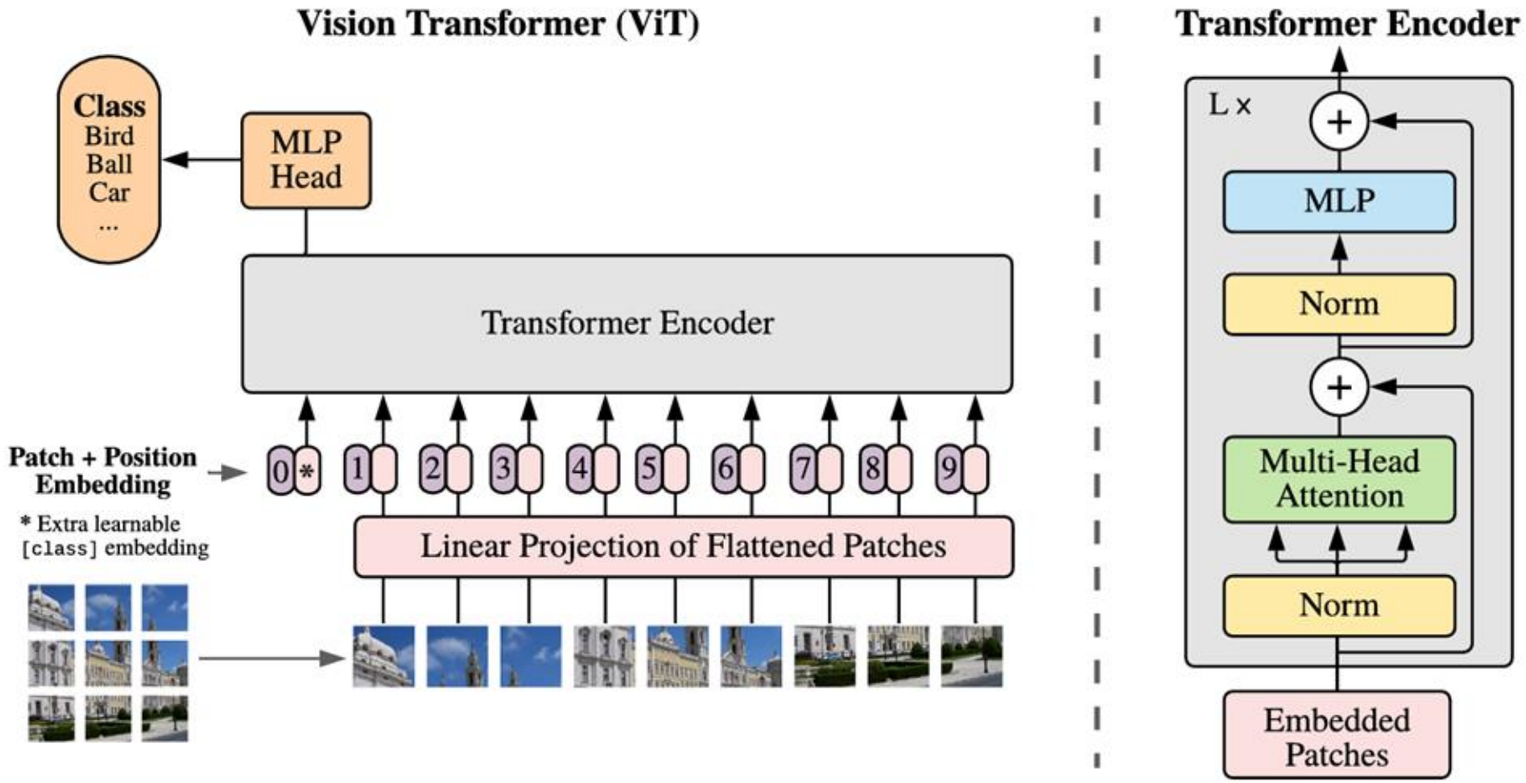
Vision Language Models (VLMs) typically consist of ***three*** components:

1. ***Vision Encoder***: pretrained on images.
2. **LLM**: pretrained on text.
3. **Connector** that maps visual features into the LLM's text space.



Architecture: Vision Encoders

How do vision transformers encode the image?



An Image is worth 16x16 words (Dosovitskiy et al. 2020)

Architecture: Vision Encoders

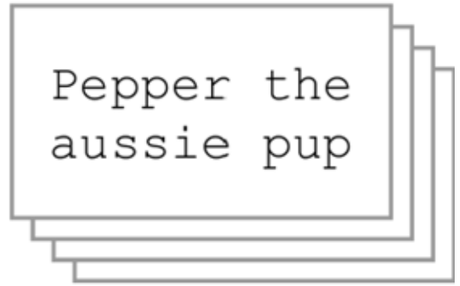
How to pretrain the ViT?

1. *Self-supervised Pretraining*
   a. *Pretrains on images only*

2. *Language-supervision*
   a. *Pretrains on Image-text pairs*

3. *Unified Recipes*

(Caron et al. 2021)

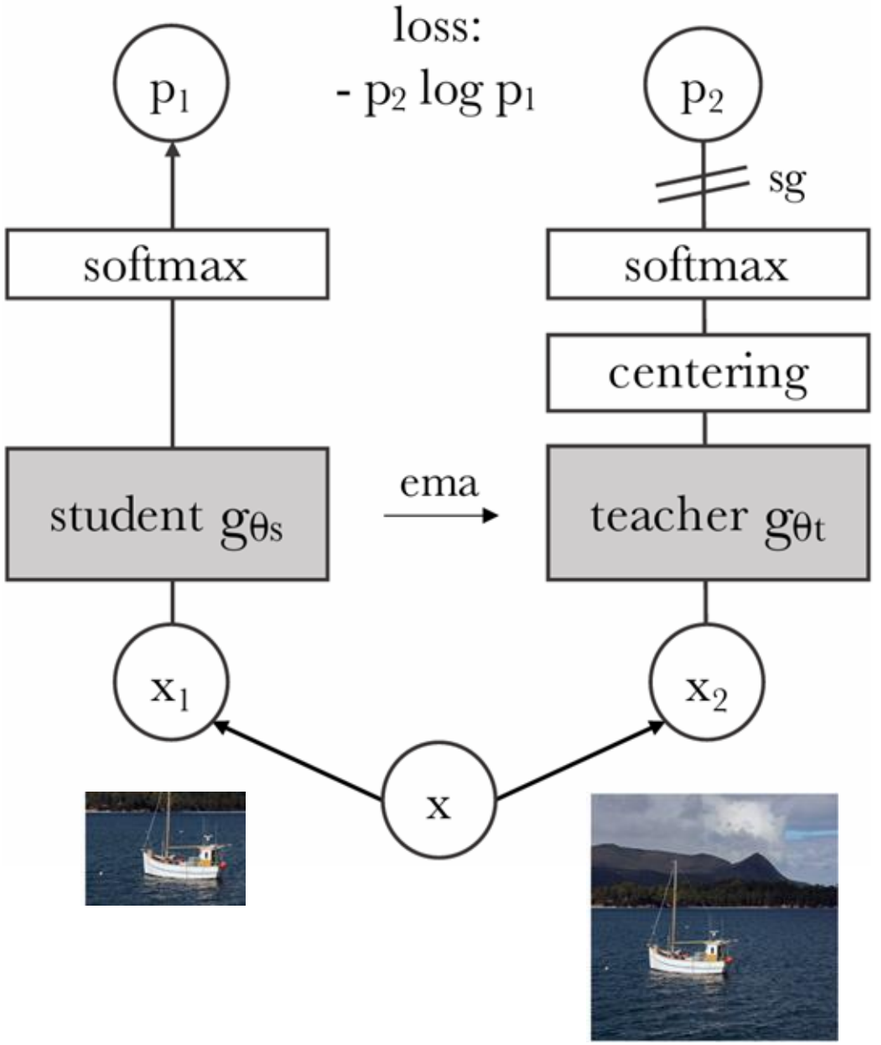


(Radford et al. 2021)

Architecture: Vision Encoders

How to pretrain the ViT?

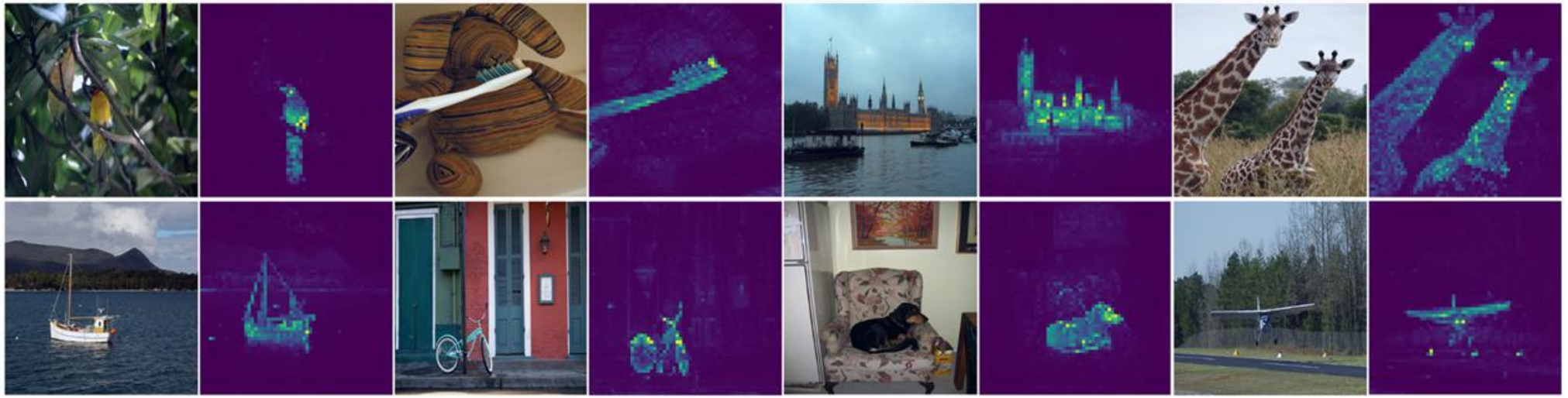
1. *Self-supervised Pretraining (DINO)*
    a. *Learns visual features by matching different views of the same image (no labels)*
    b. *Optimizes cross-entropy between teacher and student outputs*
    c. *Produces **invariant representations** useful for vision-centric downstream tasks*



(Caron et al. 2021)

Architecture: Vision Encoders

1. *Self-supervised Pretraining (DINO)*
    a. *The model learns **low-level dense features** without any supervisions*
    b. *Useful for **vision-centric** tasks such as **object detection** and **segmentation***



(Caron et al. 2021)

Architecture: Vision Encoders

2. *Language Supervision*

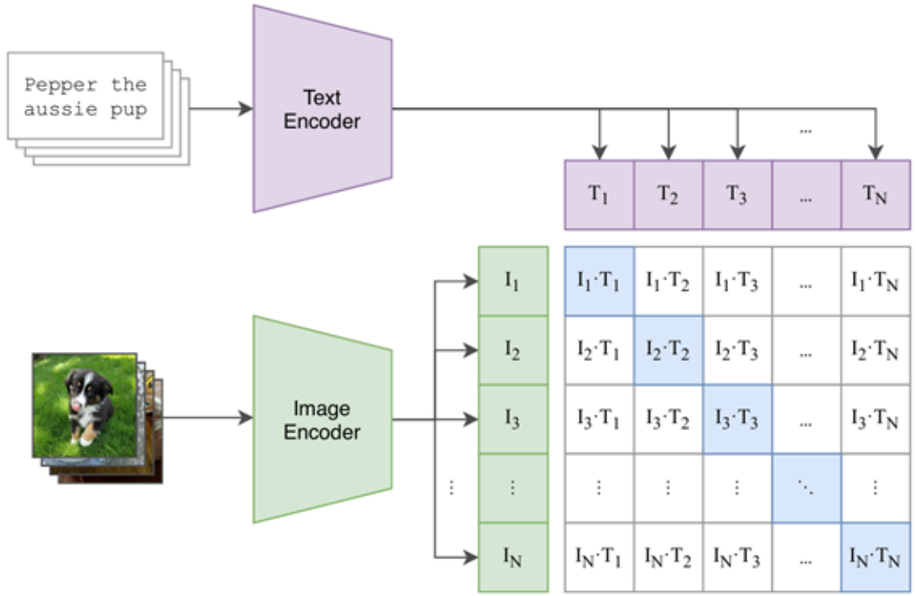
    1. ***CLIP*** *(Contrastive Language–Image Pretraining)*

        a. *Learn aligned embeddings for images and text using paired data*

        b. *Train with contrastive loss:*
            i. *match correct image-text pairs*
            ii. *push apart mismatched ones*

        c. *The model learns high-level semantics (object categories, text)*

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$



(Radford et al. 2021)

Architecture Principles: Vision Encoders

2. Language Supervision

2. SigLIP (Sigmoid Loss)

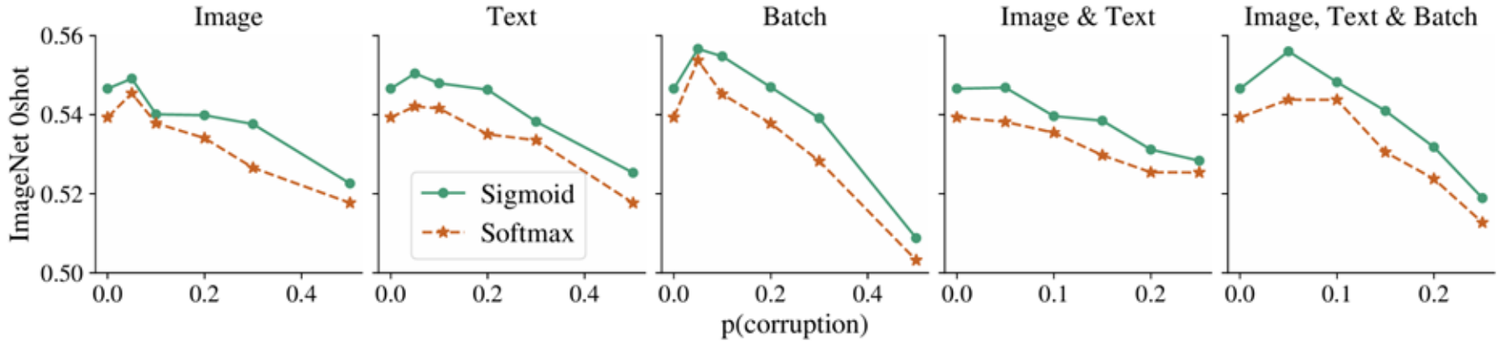
a. Replaces softmax contrastive loss with ***sigmoid (binary) loss***

b. Treats each image-text pair as **positive or negative independently** (no global normalization)

c. More scalable training to larger batch size.

d. More robust to noise and data corruption.

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

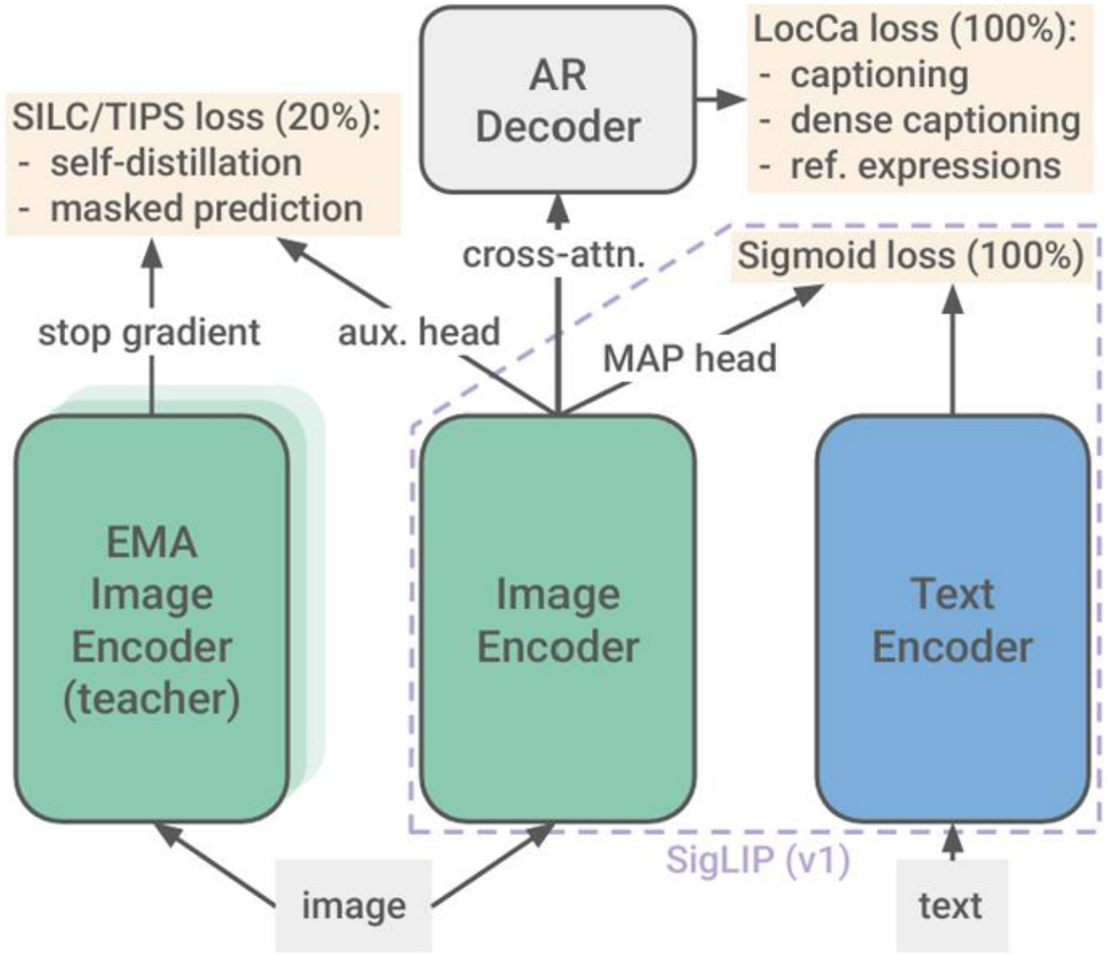


(Zhai et al. 2023)

Architecture: Vision Encoders

*3. Unified Recipe (**SigLIP2**)*

- Combines Sigmoid, self-distillation, and captioning losses.



(Tschannen et al. 2025)

Architecture: Vision Encoders

Which ViTs are preferred for VLMs?

1. *Cross-modal alignment is critical*
    a. ***Vision** features must be compatible with **language.***
    b. *Enables seamless fusion in LLMs*

2. *Language-supervised encoders are preferred*
    a. *Learn a shared vision-text embedding space*

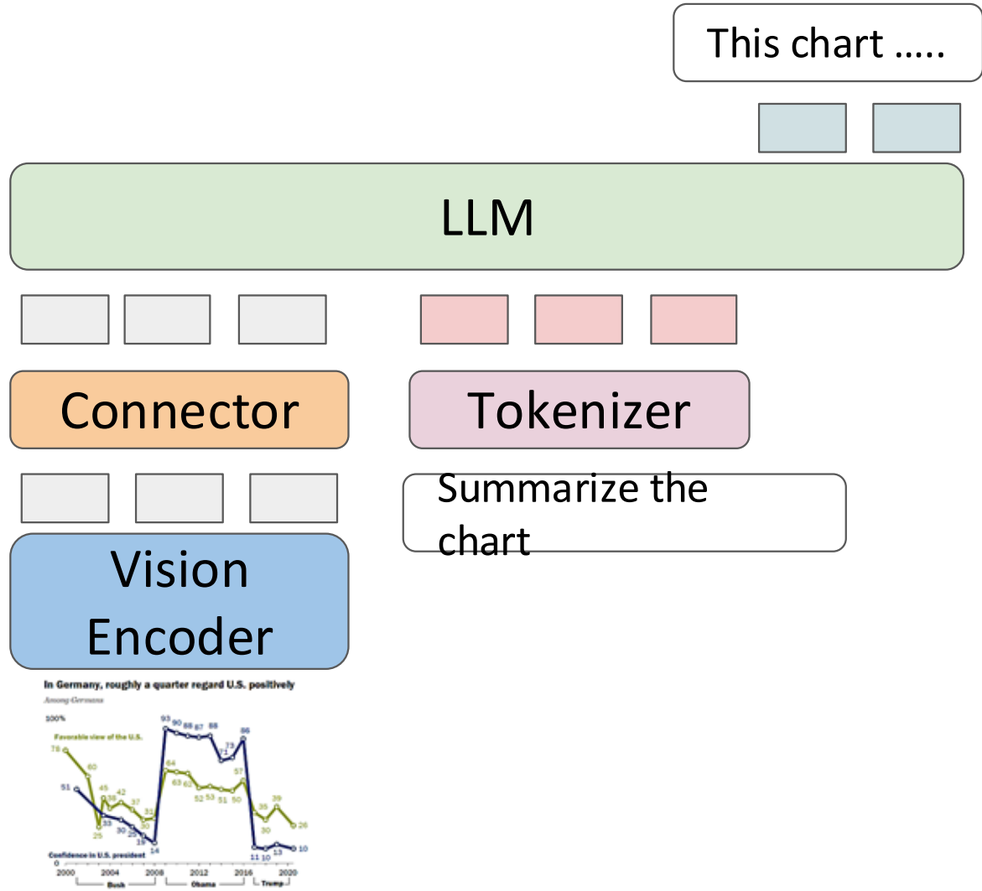
Architecture: Connectors

How do we connect vision encoders with LLMs?

Two class of connectors:

1. ***Deep Fusion***
   a. *Modifies the LLM's architecture.*
      i. *Adds new cross-attention layers.*
   b. *Injects visual features across multiple layers*
      i. *Enables deeper vision–language interaction*

2. ***Shallow Fusion***
   a. *Projects visual features into the LLM input space*
   b. *No changes to LLM architecture*
      i. *Simpler and more efficient*



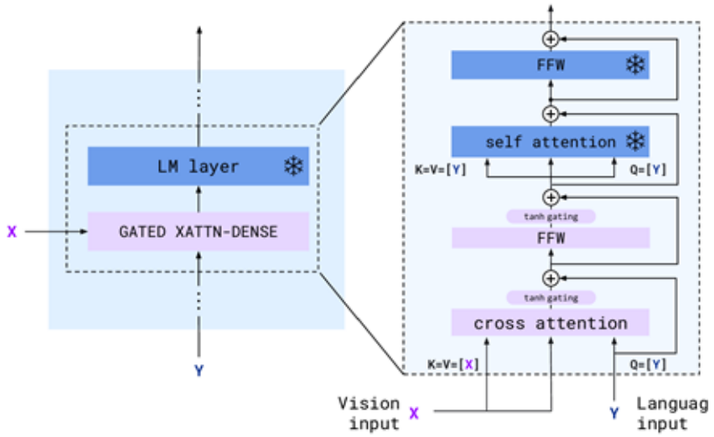
Architecture: Connectors

1. Deep Fusion

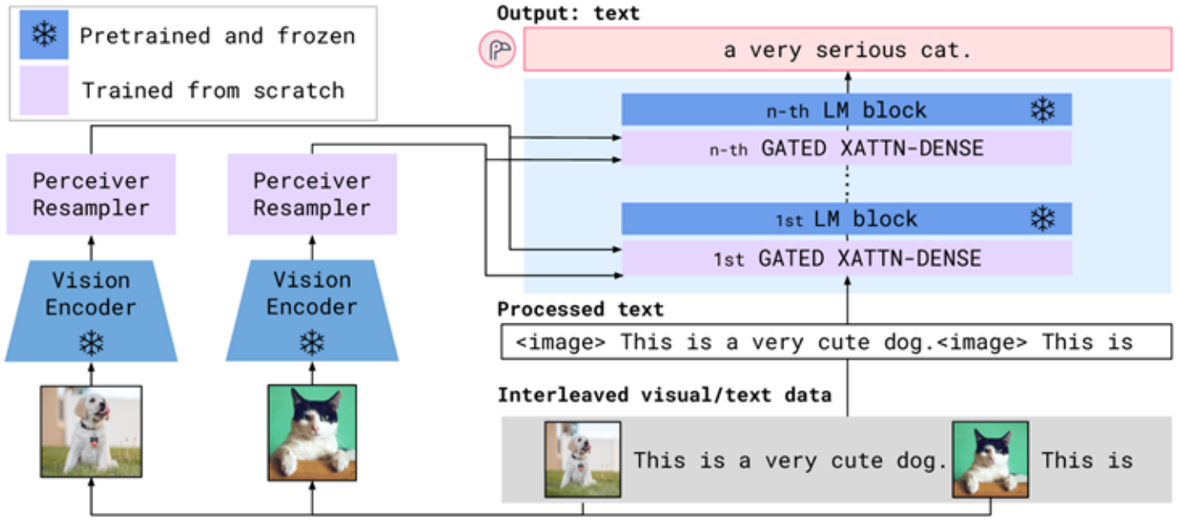
Flamingo *(Alyarac et al. 2022)*

1. *Perceiver Resampler:*
   a. *Uses learnable latent queries to attend over visual features*
   b. *Produces fixed size of visual tokens.*

2. *Gated attention layers*
   a. *Combines text and visual features.*



GATED XATTN-DENSE (Alayrac et al. 2022)



Flamingo (Alayrac et al. 2022)

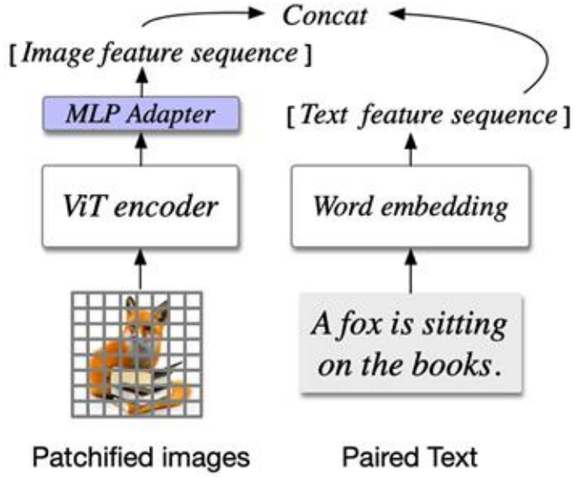
Architecture: Connectors

1. Deep Fusion

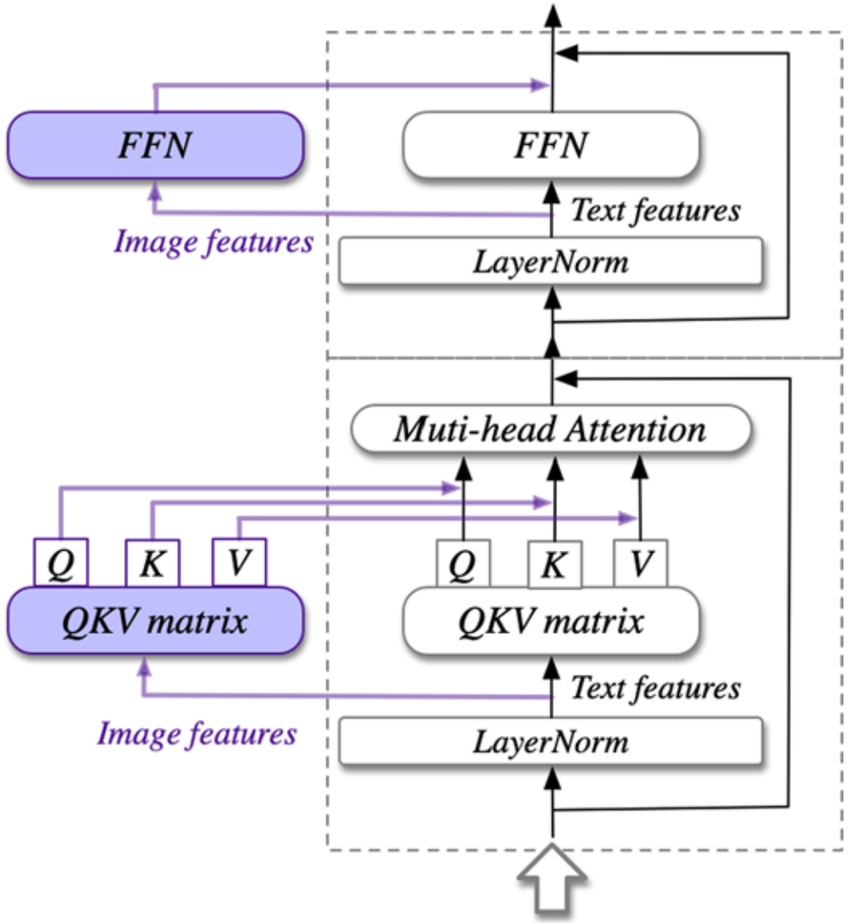
CogVLM *(Wang et al. 2023)*

1. *Image is encoded by a pretrained ViT and projected into the same embedding space as text using an MLP.*

2. *Within the Transformer block, image features use separate QKV projections and FFN from text features*



CogVLM Inputs (Wang et al. 2023)



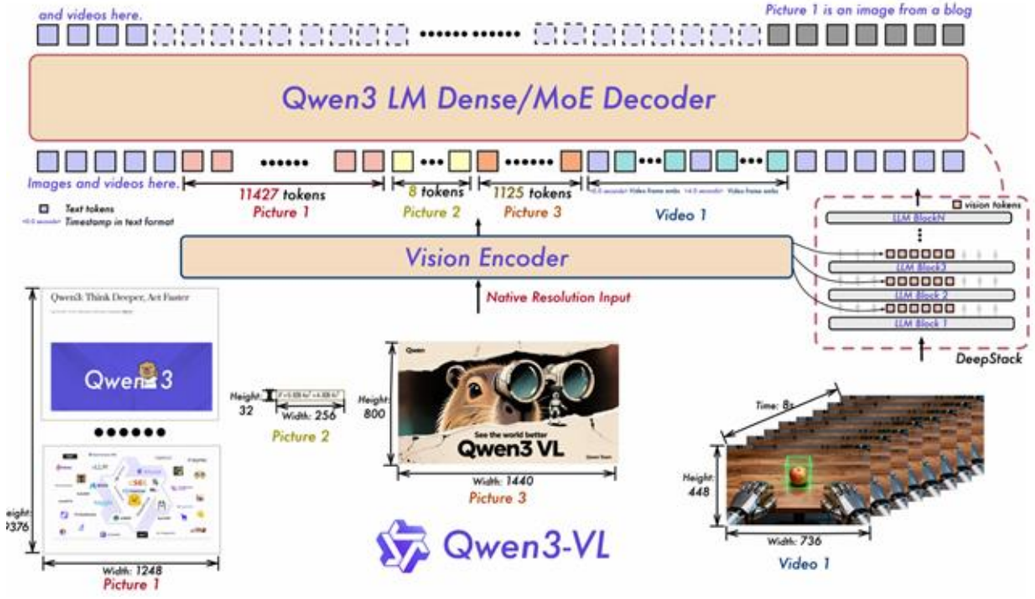
CogVLM (Wang et al. 2023)

Architecture: Connectors

1. Deep Fusion

Qwen3-VL *(Qwen Team 2025)*

1. *Injects multi-level visual features into multiple LLM layers (DeepStack)*
2. *Preserves both **low** and **high-level** visual information (from intermediate ViT layers)*



Qwen3-VL (Qwen Team 2025)

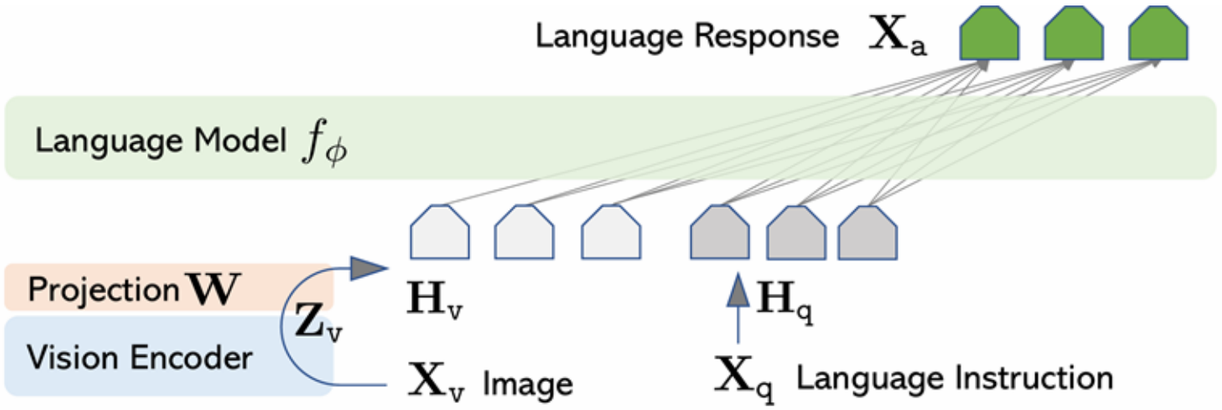
Architecture: Connectors

2. Shallow Fusion

Llava *(Liu et al. 2023)*

1. *Image is encoded by a pretrained ViT and projected into the same embedding space as text using either a linear layer or MLP.*

2. *Visual features are concatenated with text features as input to the LLM.*



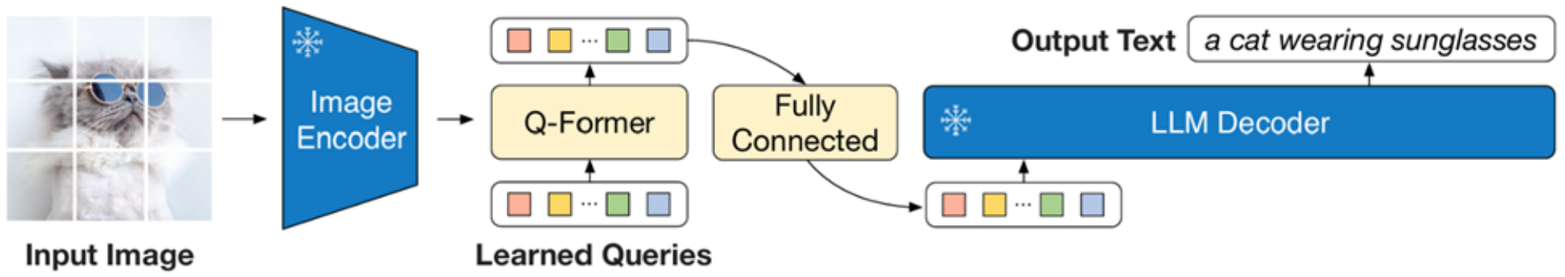
Llava (Liu et al. 2023)

Architecture Principles: Connectors

2. Shallow Fusion

Q-former *(Li et al. 2023)*

1. *Uses learnable query tokens to attend over image features.*
2. *Visual features are concatenated with text features as input to the LLM.*



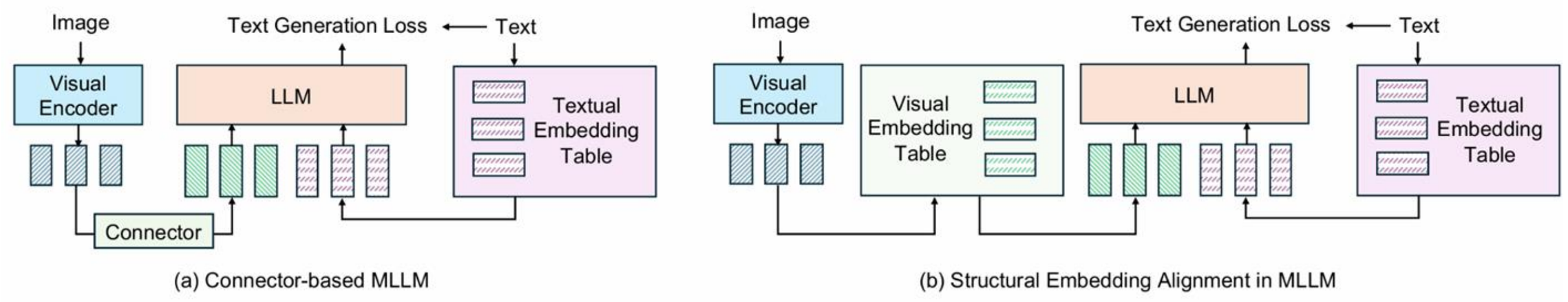
BLIP-2 (Li et al. 2023)

Architecture: Connectors

2. Shallow Fusion

Ovis *(Lu et al. 2024)*

1. Aligns vision and text structurally by introducing a learnable visual embedding table.
2. Mirrors how text tokens use embedding lookups

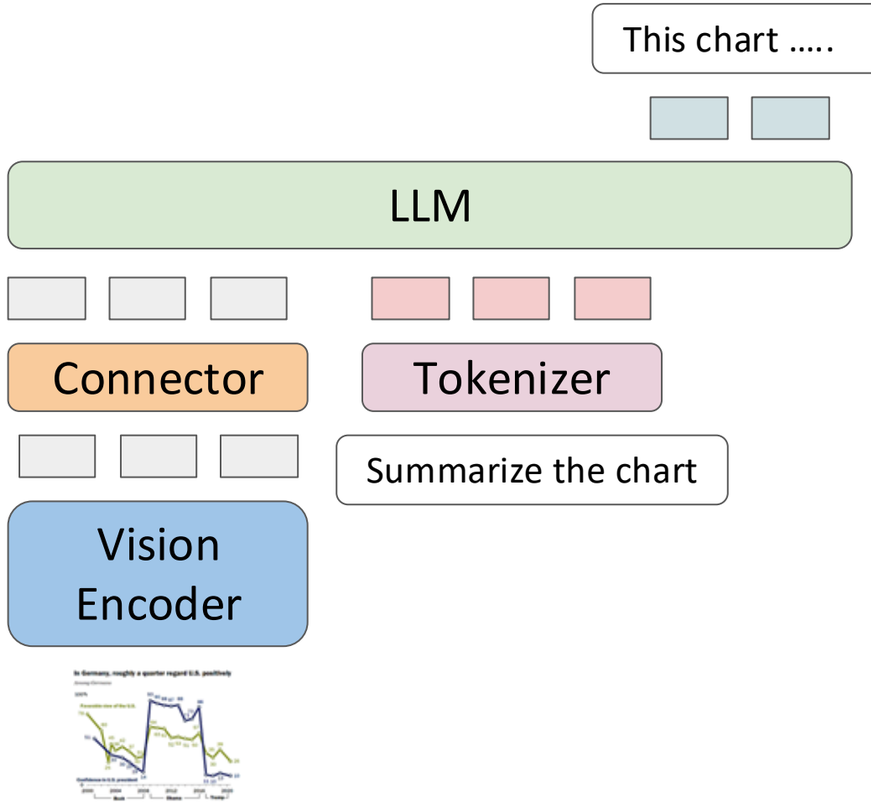


Ovis (Lu et al. 2024)

AlignVLM: Bridging Vision and Language Latent Spaces for Multimodal Document Understanding

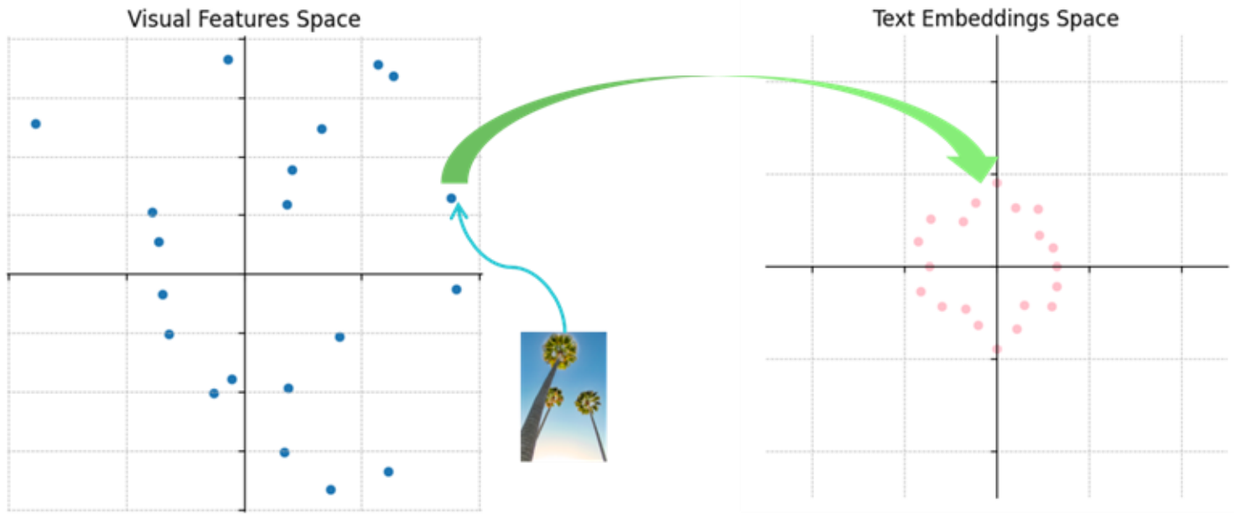
[NeurIPS 2025]

AlignVLM Case Study: Motivation

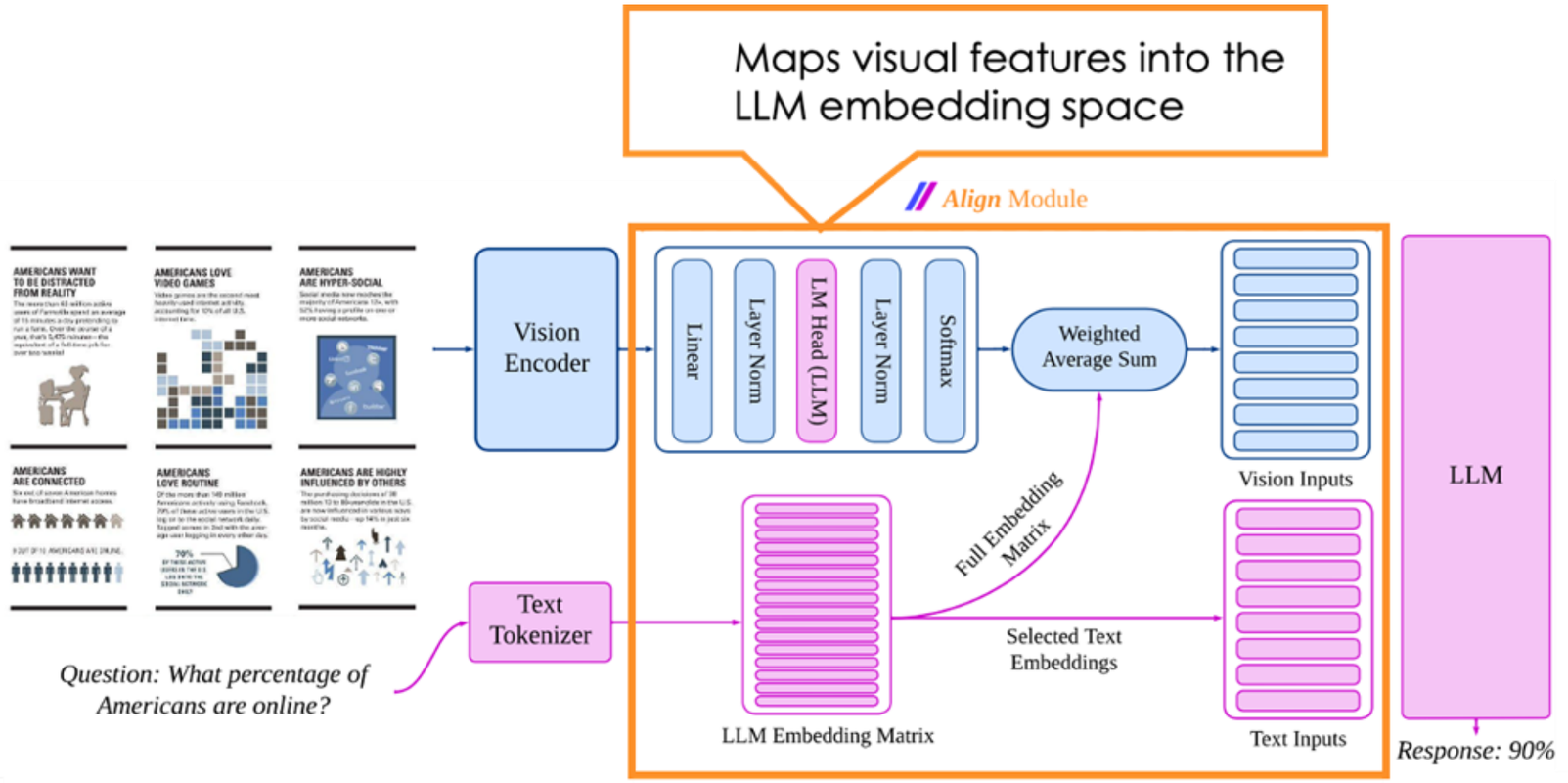


- LLM is pretrained to process a discrete set of embeddings
- Existing connector (e.g., MLP) produce continuous visual features
    - **Out-of-distribution (OOD):** making the connector data hungry!
    - **Unconstrained Mapping:** They do not enforce any hard constraints which makes them *prone to noise.*

Can we exploit the LLM's inductive bias by aligning visual features directly with text embeddings?



AlignVLM Case Study: Solution



- Map visual features to a prob distribution over LLM token embeddings.

$$\mathbf{P}_{\text{vocab}} = \text{softmax}(\text{LayerNorm}(\mathbf{W}_2\,\text{LayerNorm}(\mathbf{W}_1\mathbf{F}))) \quad (1)$$

- Computes final features as a weighted average of text embeddings.

$$\mathbf{F}'_{\text{align}} = \mathbf{P}_{\text{vocab}}^{\top}\mathbf{E}_{\text{text}}.$$

- Constraints visual inputs to the convex hull of the LLM's embedding space, making them familiar to the LLM.

AlignVLM Case Study: Training Setup

Training Stages

Stage 1: Natural Image Understanding
Data: CC-12M (Image-Caption)

Stage 2: Document Understanding
Data: BigDocs-7.5M

Stage 3: Instruction Tuning for downstream tasks
Data: BigDocs-Docdownstream

Model components

- **LLM:** Llama 3.2 Family (1B, 3B, 8B)
- **Vision Encoder:** SigLip-400m

Evaluation Benchmarks

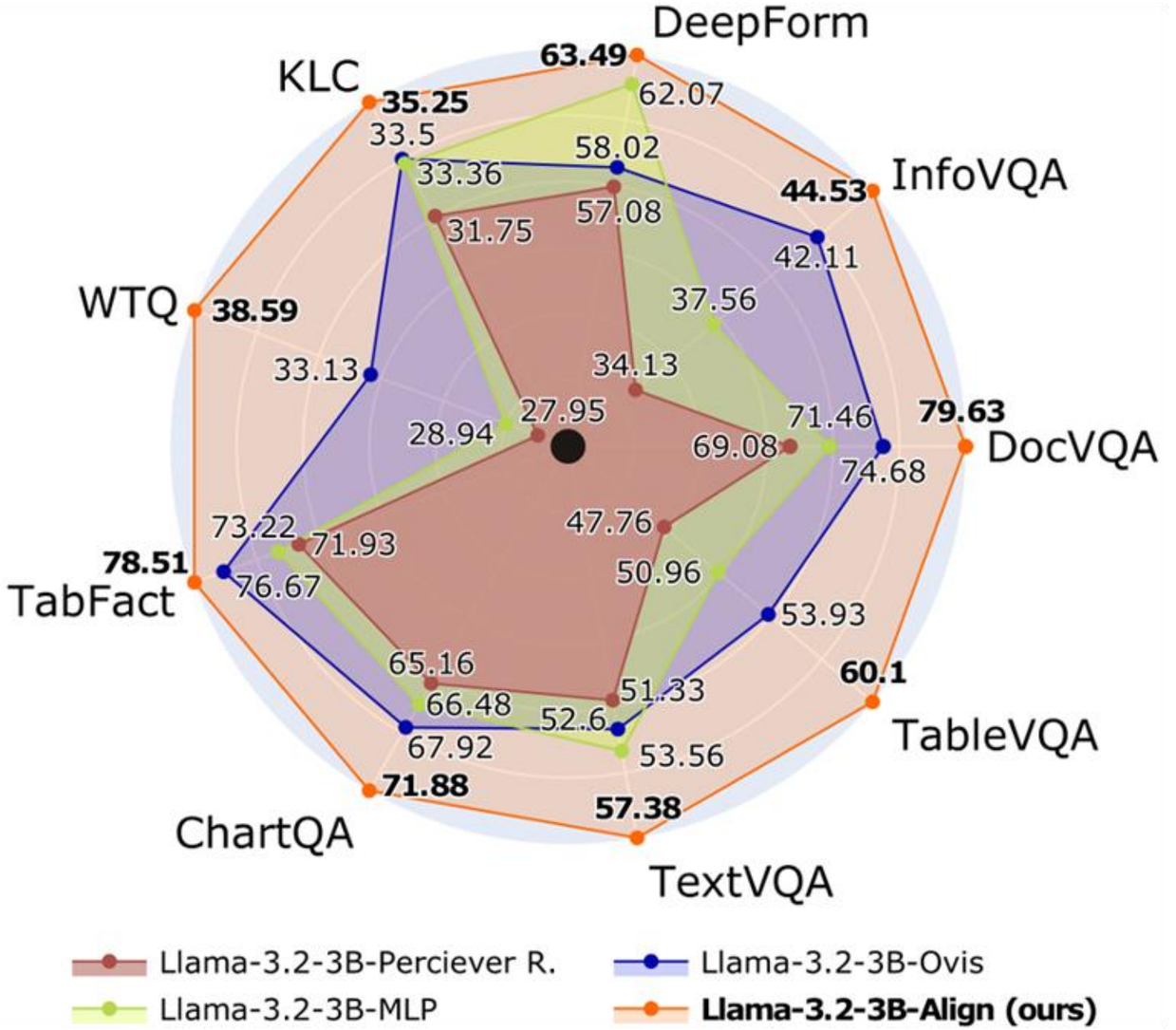
Nine document benchmarks, including:

- DocVQA, InfoVQA, ChartQA, TableVQA, etc.

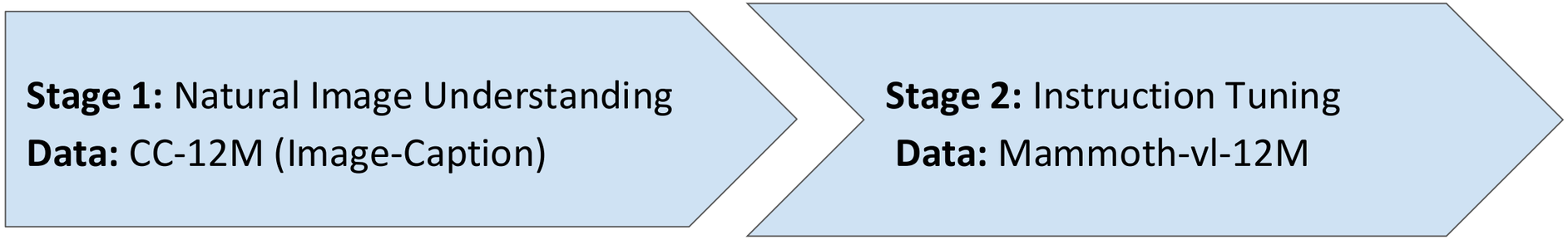
AlignVLM Case Study: Results on Document Understanding

- We compare **our Align Module** against common connectors:
  - MLP, Perceiver Resampler, Ovis
- Trained under similar configurations to ensure a fair comparison.

The **Align Module** outperforms them all and achieves better accuracy on diverse document understanding tasks.

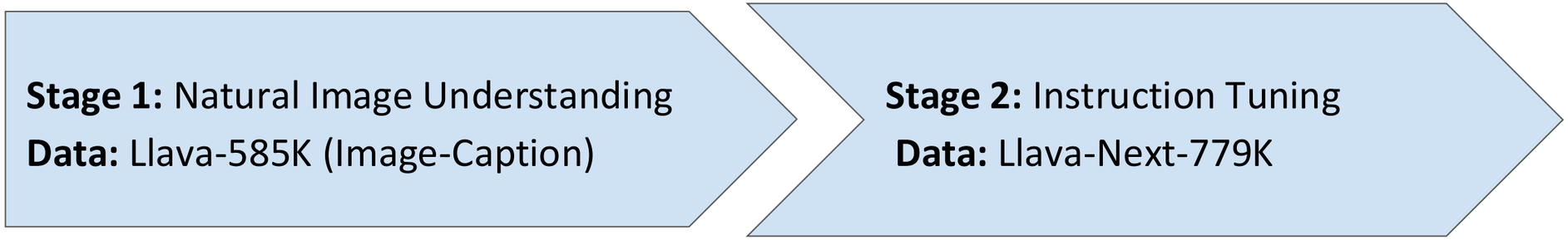


AlignVLM Case Study: Results on General Vision Tasks



Model	MMMU (dev)	SeedBench	MMVet	POPE	GQA
Llama-3.2-3B-MLP	35.66	71.68	44.95	84.11	37.07
Llama-3.2-3B-ALIGN (ours)	**38.66**	**72.87**	**47.75**	**84.73**	**42.77**

AlignVLM Case Study: Results in low-resource setup



Model	DocVQA	InfoVQA	ChartQA	TextVQA	Average	Δ
LLama-3.2-3B-MLP (Llava Next)	42.11	19.93	48.44	51.97	40.61	
LLama-3.2-3B-Align (Llava Next)	71.43	30.50	69.72	65.63	59.32	+18.71
LLama-3.2-3B-MLP (BigDocs)	71.46	37.56	66.48	53.56	57.26	
LLama-3.2-3B-Align (BigDocs)	79.63	44.53	71.88	57.38	**63.35**	+6.09

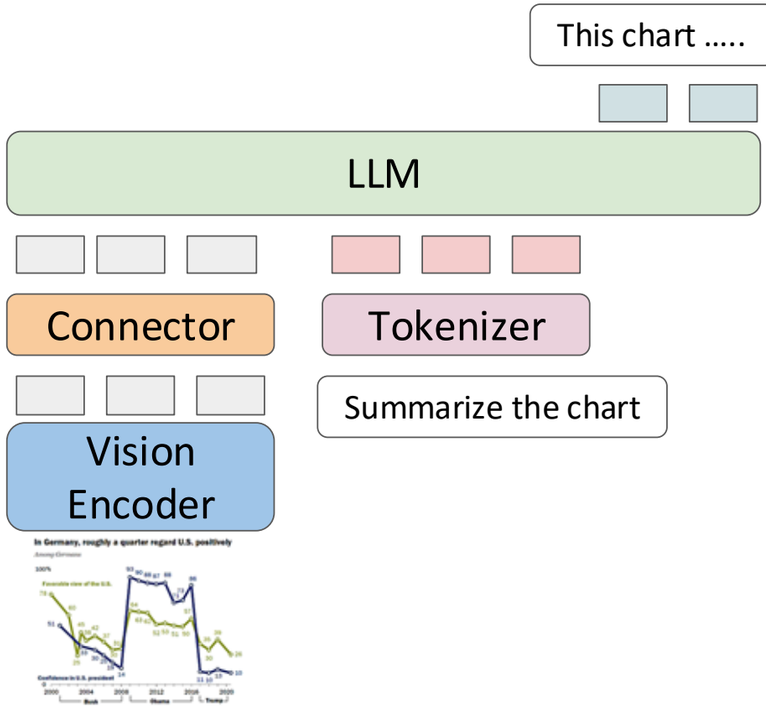
Training: Alignment and Instruction Tuning

How do we train the VLMs?

***Two-stage** training process:*

1. **Alignment:**
   a. Image-text pairs (e.g., charts + captions).
   b. Freeze vision & LLM. Train the connector only!

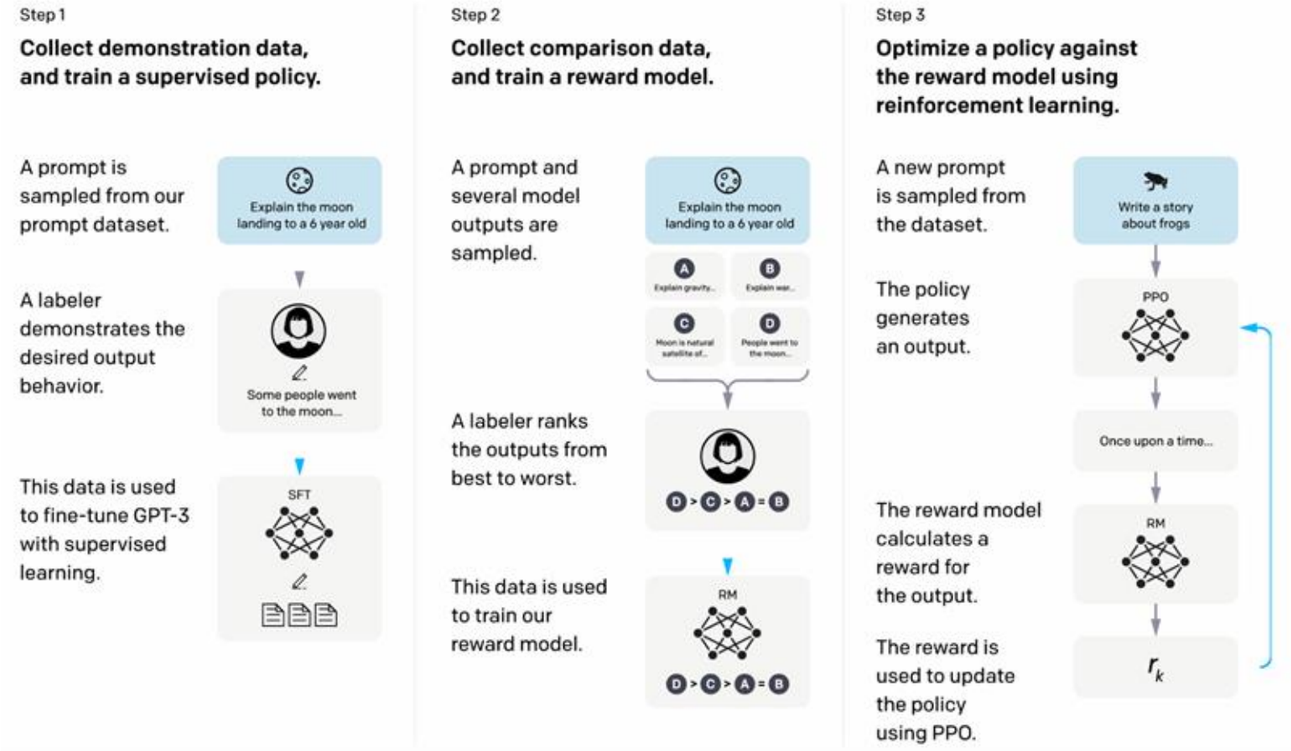
2. **Instruction-tuning:**
   a. Instruction following (e.g., QA).
   b. Train LLM & Connector. Freeze vision.



Training: Reinforcement Learning with Human Feedback

Reinforcement Learning with Human Feedback (RLHF)

- **Goal:** Align model outputs with human preferences (e.g., removes toxicity)
- **Key Idea:** Learn a reward function from humans, then optimize the model to maximize it.



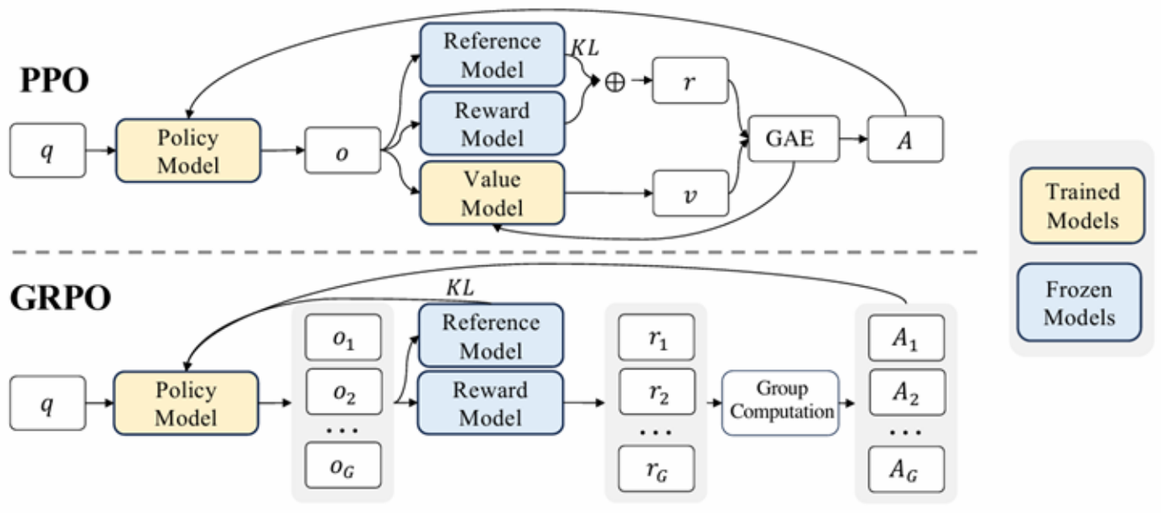
RLHF Process (Ouyang et al. 2022)

Training: Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR)

- **Goal:**
  - Improve **reasoning** capabilities using **objective, programmatic rewards**

- **Key Idea:**
  - Replace human feedback with **automatically verifiable signals**
  - Reward correctness using **rules, programs, or ground truth checks**
    - Works for Math, coding, symbolic reasoning, ..etc.
    - Struggles with subjective tasks (writing, open-ended generation)



PPO vs. GRPO (Deepseek AI 2025)

Training: How to obtain the data for training?

Human Annotation

Pros

- High-quality and linguistically rich annotations.
- Complex reasoning and real-world knowledge

Cons

- Very costly.
- Scaling is challenging

Semi-automatic Annotation

Pros

- More scalable than human annotations.
- Preserve accuracy, nuance from human oversight

Cons

- Diversity is limited by the generating model
- Scaling is still challenging

Training: How to obtain data for training?

Synthetic Generation

Pros

- **Highly scalable:** LLM/VLM generation can produce millions
- **Low annotation cost:** Reduces reliance on expensive human labeling.

Cons

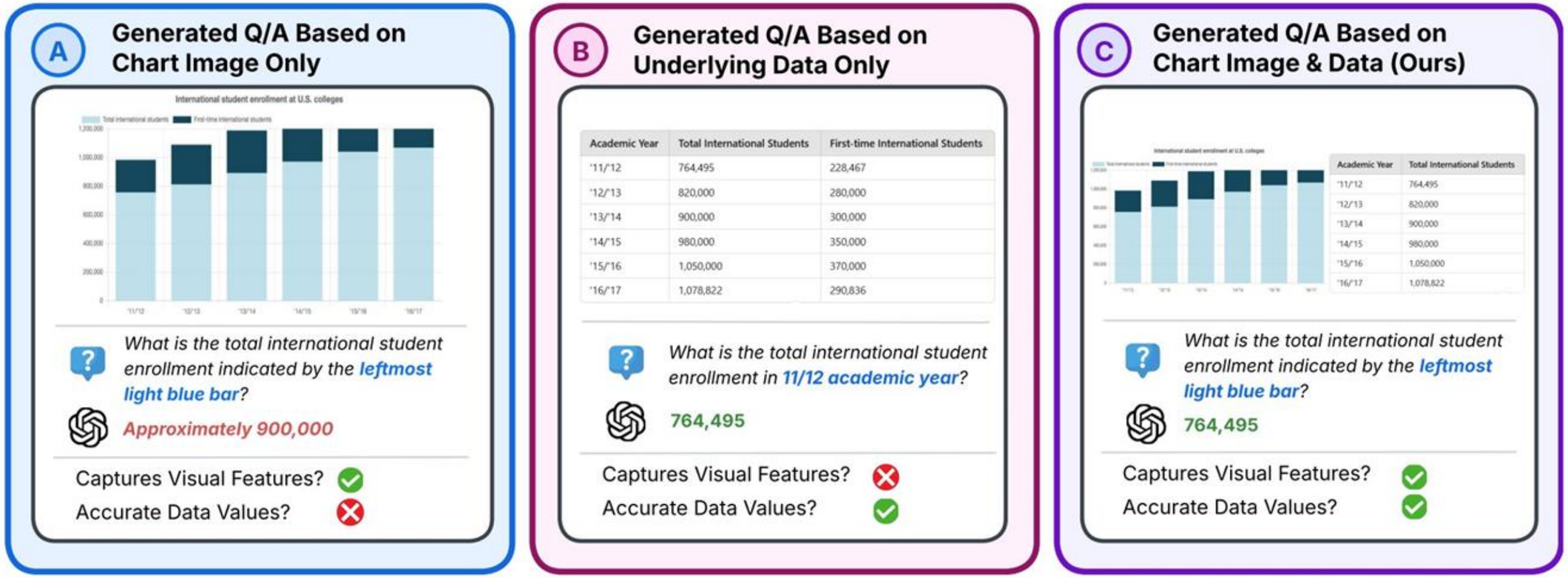
- **Quality and collapse issues:** Synthetic labels may suffer from errors

BigCharts-R1: Enhanced Chart Reasoning with Visual Reinforcement Finetuning

[COLM 2025]

BigCharts-R1 Case Study: Motivation

- Synthetic data for charts are generated from a single modality.
  - Chart image or underlying data table.

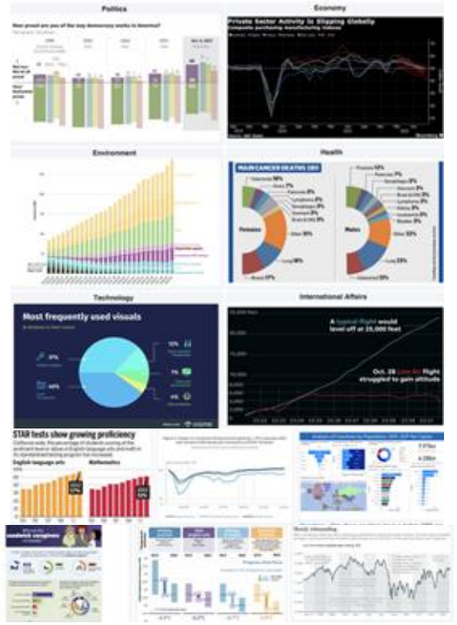


BigCharts-R1 Case Study: Motivation

- Chart images that have accompanying data tables/metadata come from few sources:
  - Lack of Diversity
  - Homogenous.
- Most charts on the web do not have any associated data tables/codes
  - Visually diverse

*How can we obtain chart images that are **both** visually diverse and provide underlying metadata?*

Real-world chart images
(Diverse Styles)

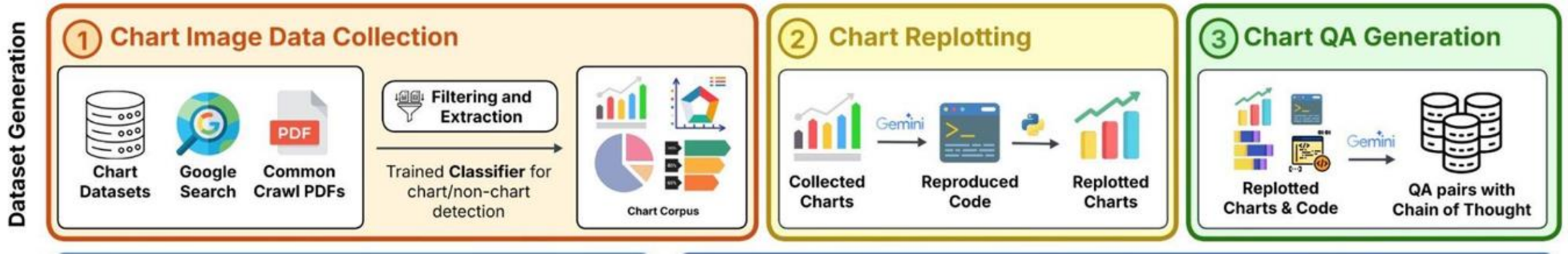


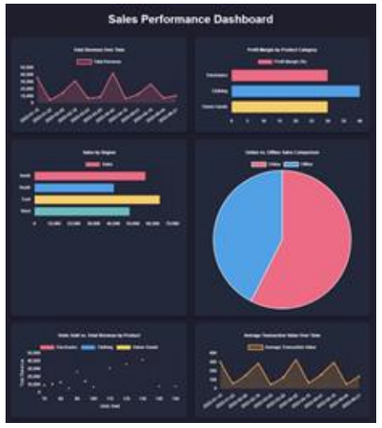
☹ **Homogenous Chart Images**
(Consistent Style, Real & Synthetic)

BigCharts-R1 Case Study: Dataset Pipeline

A dataset creation pipeline that:

- Source real-world charts from multiple online platforms.
- Generates visually diverse chart images by “replotting” real-world chart images.



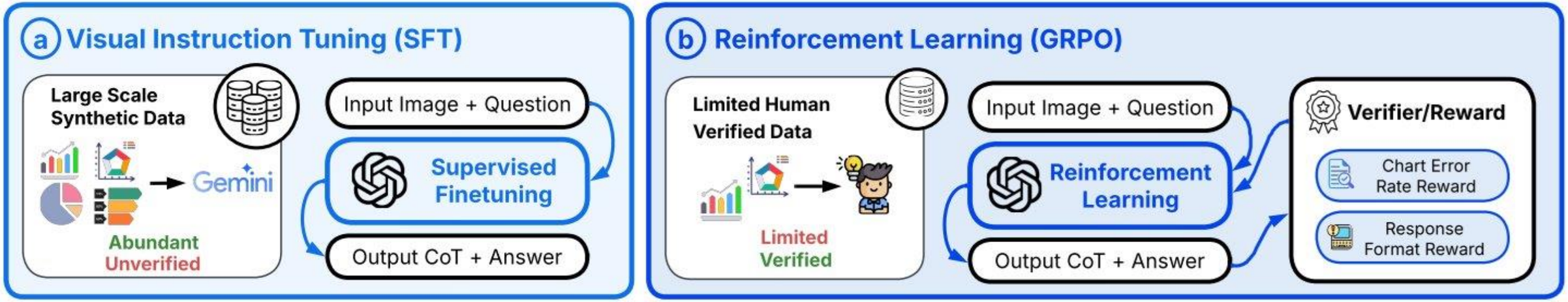


Original Image

Code & Data

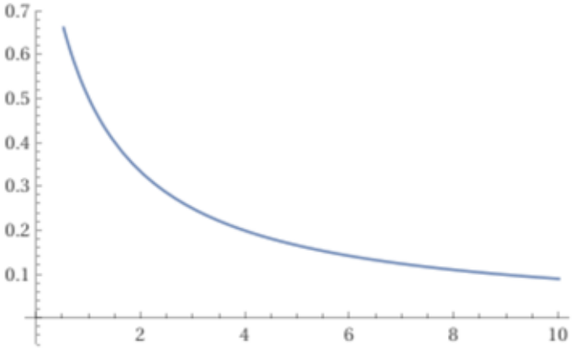
Replotted Image

BigCharts-R1 Case Study: Training Framework



RL (GRPO) on human-labeled data to enhance chart visual math reasoning with verifiable rewards

$$\mathrm{ER}(\hat{y}, y) = \frac{|\hat{y} - y|}{|y|}, \qquad R_{\mathrm{CERM}}(\hat{y}, y) = \begin{cases} \dfrac{1}{1 + \mathrm{ER}(\hat{y}, y)}, & \text{if both } \hat{y} \text{ and } y \text{ are numeric,} \\ 1, & \text{if non-numeric and } \hat{y} = y, \\ 0, & \text{otherwise.} \end{cases}$$

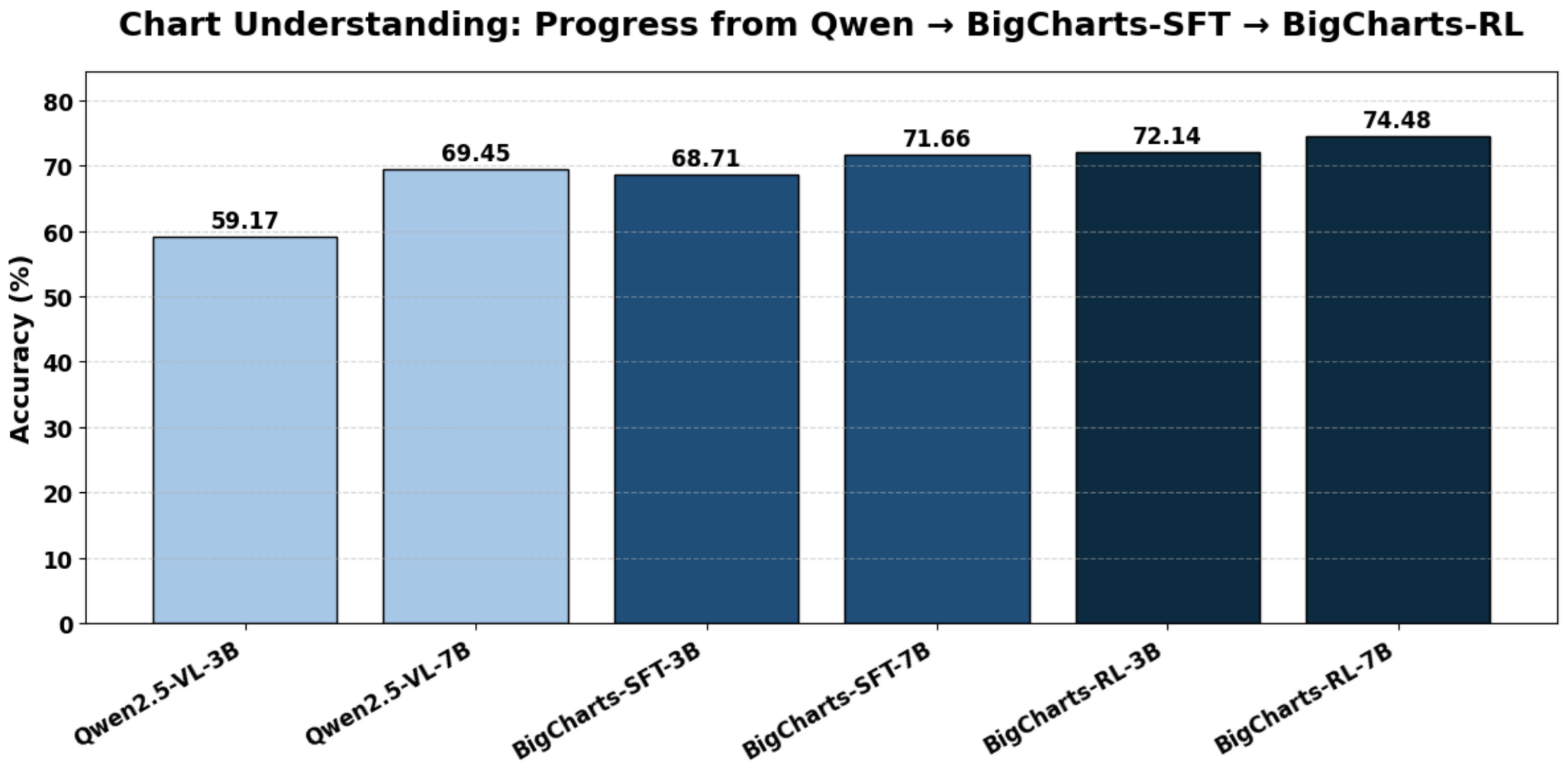


$$R_{\mathrm{Fmt}} = \begin{cases} 1 & \text{if the model follows the required response structure,} \\ 0 & \text{otherwise.} \end{cases}$$

$$R_{\mathrm{total}} = R_{\mathrm{CERM}} + R_{\mathrm{Fmt}}$$

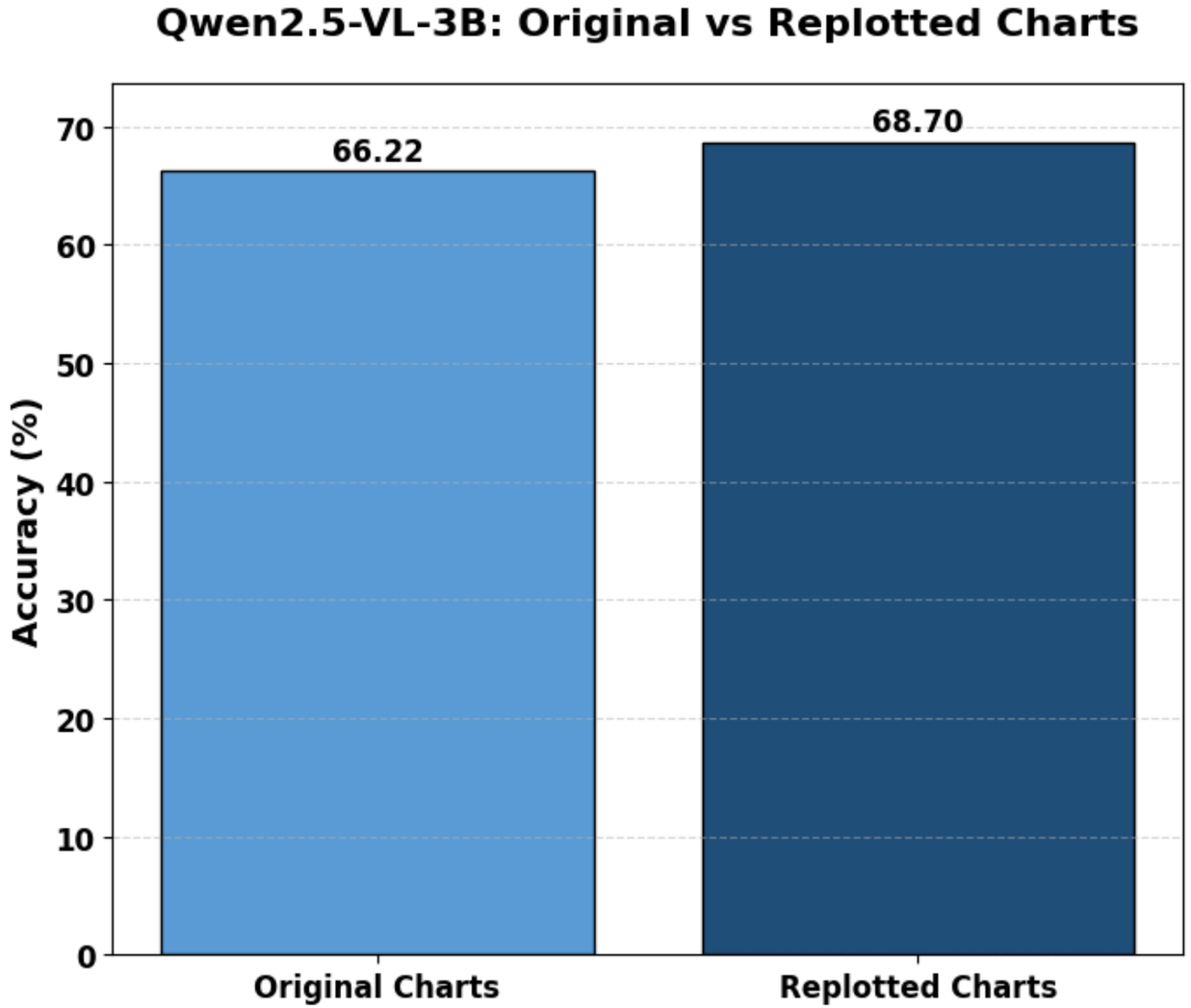
BigCharts-R1 Case Study: Results

Evaluation: average score across five benchmarks **(**FigureQA, DVQA, PlotQA, ChartQA, CharXiv)



BigCharts-R1 Case Study: Results

Comparison between finetuning on Q/A generated from Original Images vs Replotted Images

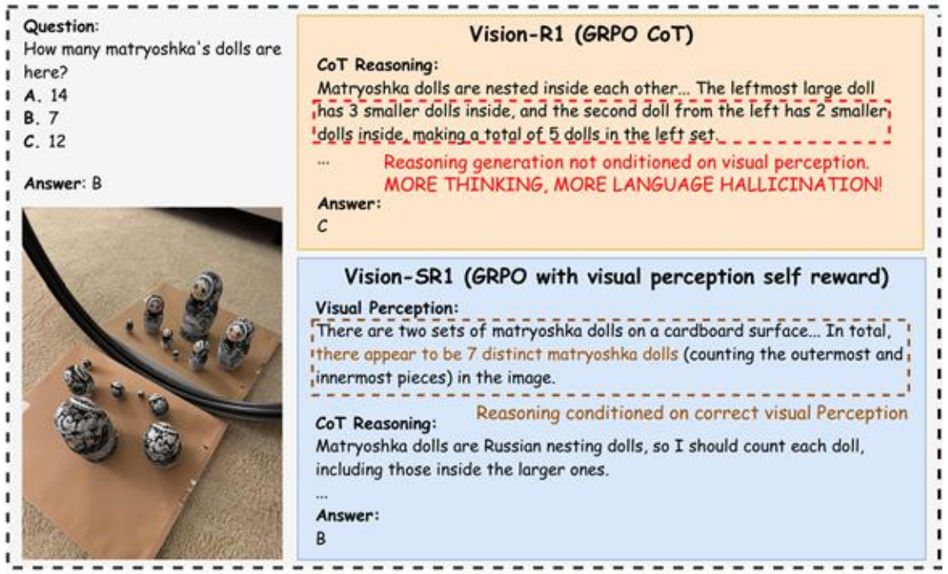


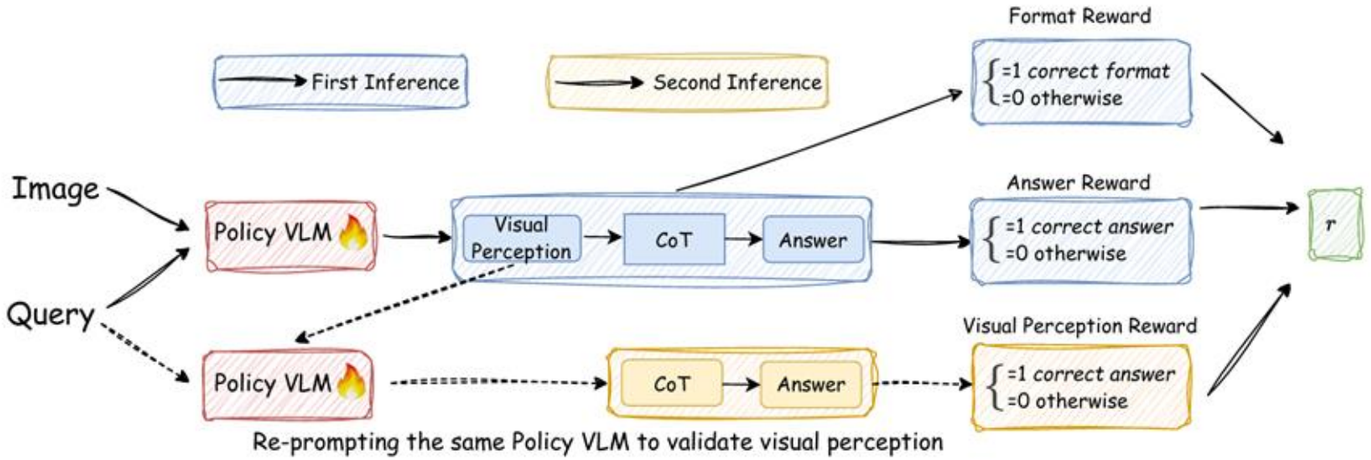
Training: Can RLVR improve perception capabilities?

Can we use RLVR to improve perception as well?

Vision-SR1

- *Decomposes reasoning into visual perception + language reasoning*
- *Uses self-rewarding via re-prompting (perception -> reasoning)*





Vision-SR1 (Li et al. 2025)

Training: Can RLVR improve perception capabilities?

Can we use RLVR to improve perception as well?

Vision-SR1

- *Decomposes reasoning into visual perception + language reasoning*
- *Uses self-rewarding via re-prompting (perception -> reasoning)*

Methods	General Visual Understanding					Visual Math & Hallucination			Avg.
	MMMU -Pro	MMMU	MM -Vet	RealWorld QA	VisNum Bench	Math Verse	MATH -Vision	Hallusion Bench	
Vision-SR1 (3B)	40.8	49.6	69.7	66.1	41.9	48.5	38.5	68.3	52.9
⊢ w/o self-reward	40.0	48.0	67.4	62.6	41.6	47.7	38.9	65.8	51.5
Vision-SR1 (7B)	49.1	57.2	76.2	71.6	42.6	56.5	46.7	69.8	58.8
⊢ w/o self-reward	48.8	55.3	78.4	70.9	41.4	54.8	45.3	66.4	57.7

Vision-SR1 (Li et al. 2025)

Tutorial Overview

1. FOUNDATIONS OF MLLMs

Evolution of LLMs to multimodal models; architectures, training, and alignment

2. MULTIMODAL REASONING

Tasks, benchmarks, and techniques for reasoning over visual documents

3. HUMAN–AI INTERACTION

Multimodal agents, GUI grounding, and interactive data analysis.

4. RESPONSIBLE & INCLUSIVE AI

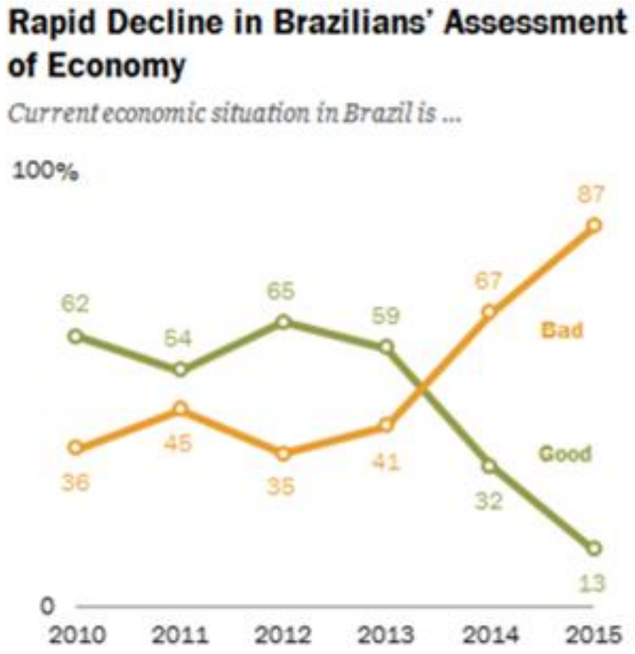
Accessibility, multilingual understanding, fairness, and hallucination risks

Future Challenges & Outlook

Tasks & Benchmarks

ChartQA (Masry et al., 2022)

- *Real-world charts crawled from 4 online sources*
- *9.6k human-authored and 23.1K Machine-generated question*
  - *Saturated with Clause Sonnet 3.5 achieving +90%*



Question: Which year has the most divergent opinions about Brazil's economy?
Answer: 2015

Question: What is the peak of the orange line?
Answer: 87

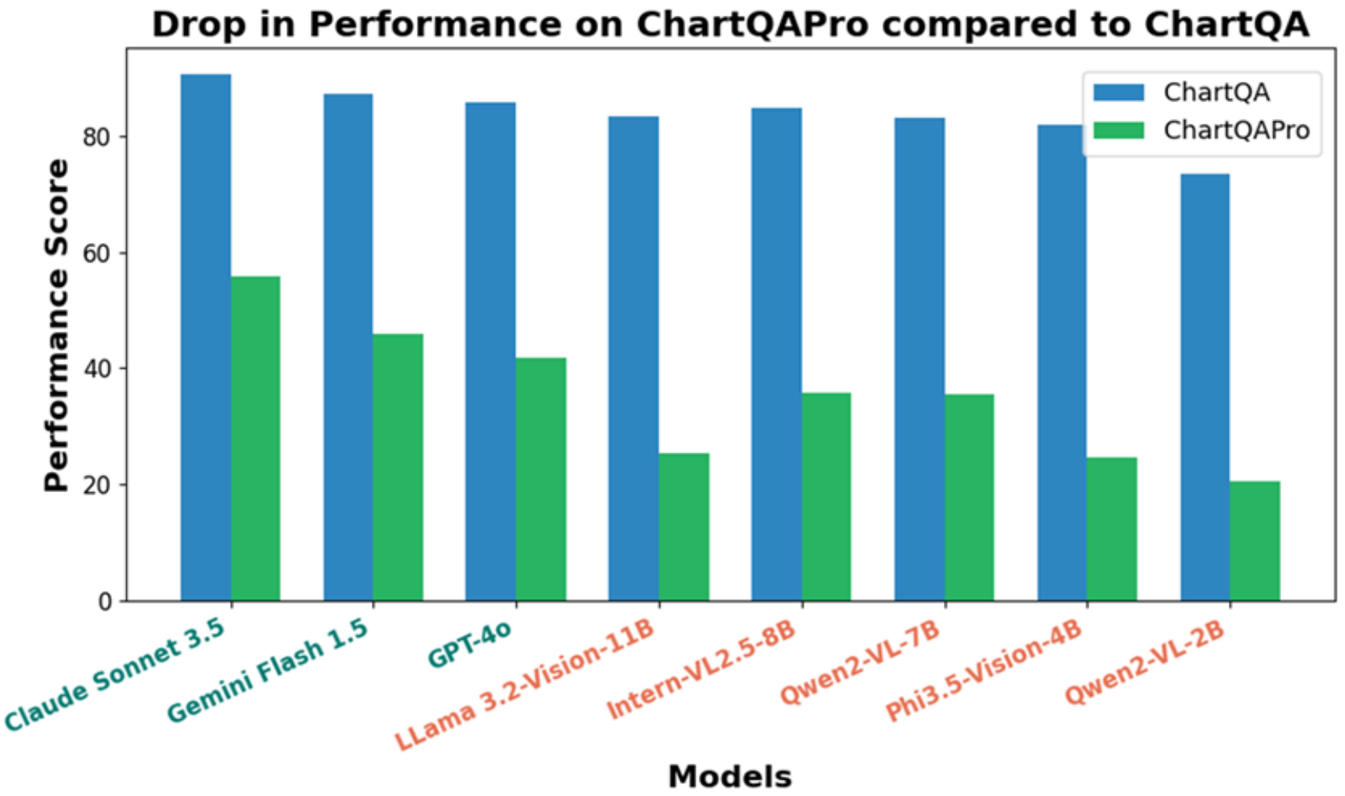
Tasks & Benchmarks

ChartQAPro *(Masry et al. 2025):*

- *1,341 charts and from 99 diverse sources*
- *More diverse chart types including infographics and dashboards*
- *1,948 questions with 8+ question types*

Tasks & Benchmarks

- *Drop in Performance on ChartQAPro*
- *Chart Question Answering is far from solved!*

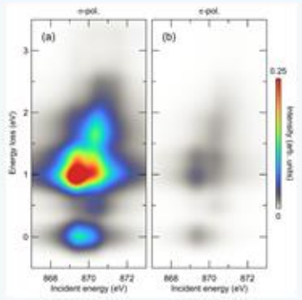


Tasks & Benchmarks

CharXiv (Wang et al. 2024)

- *Scientific Charts Question Answering with 2.3K QA pairs*

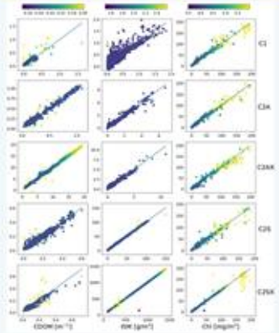
Example



Question: For the subplot at row 1 and column 1, what are the names of the labels in the legend?
* You should write down the labels from top to bottom, then from left to right and separate the labels with commas. Your final answer should account for only labels relevant to the plot in the legend, even if the legend is located outside the plot.
* If the plot does not have a legend or no legend is not considered relevant to this plot, answer "Not Applicable".

Answer: Not Applicable

Example



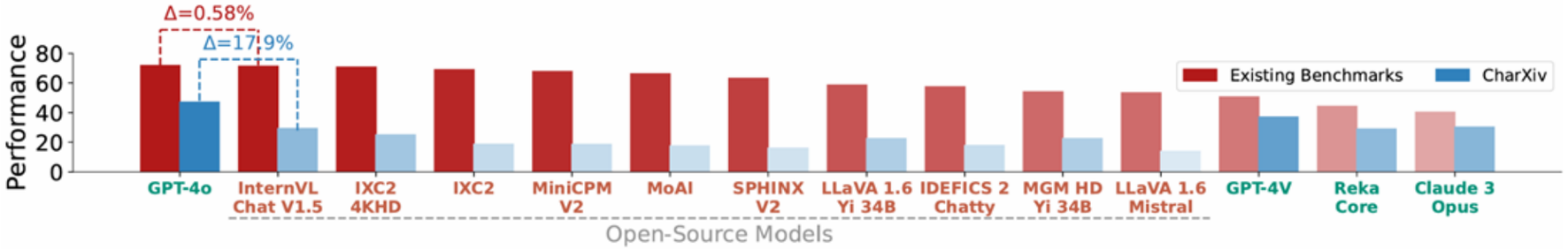
Question: For the subplot at row 5 and column 2, what is the difference between the maximum and minimum values of the tick labels on the continuous legend (i.e., colorbar)?
* You should remove the percentage sign (if any) in your answer.
* If the plot does not have an explicit colorbar-based continuous legend or the legend is not considered relevant to this subplot, answer "Not Applicable".

Answer: 0.8

Example

Question: For the bottom right subplot, what is the difference between the maximum and minimum values of the tick labels on the continuous legend (i.e., colorbar)?
* You should remove the percentage sign (if any) in your answer.
* If the plot does not have an explicit colorbar-based continuous legend or the legend is not considered relevant to this subplot, answer "Not Applicable".

Answer: Not Applicable

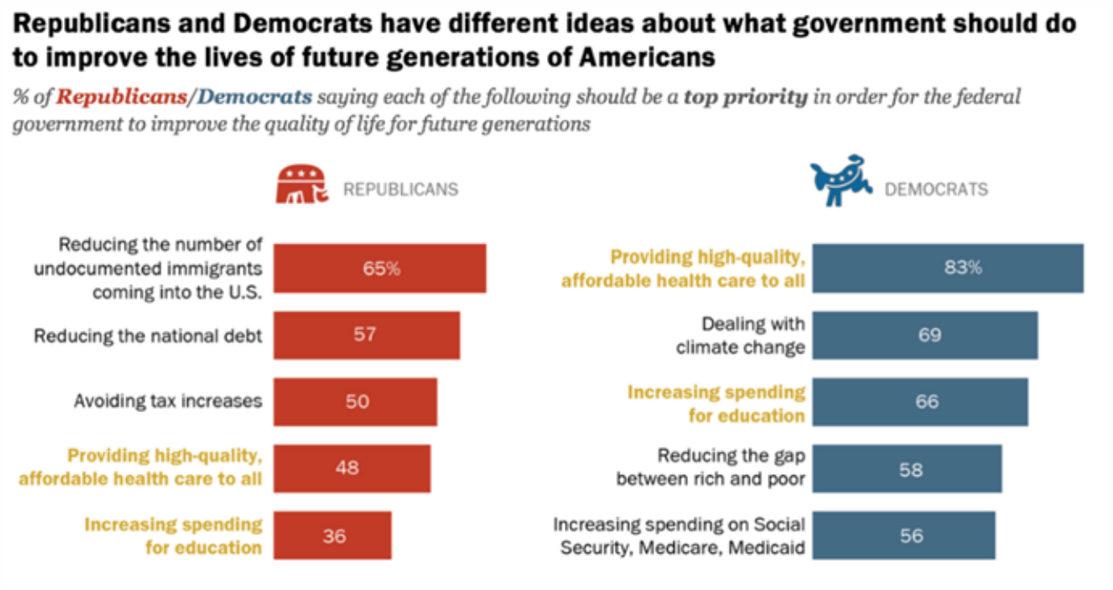


Tasks & Benchmarks: Open-ended Question Answering

OpenCQA (Kantharaj et al., 2022)

- *7,724 human-written questions about charts and the associated descriptive answers*

Question: *Compare the Democrats and Republicans views about providing health care to the population?*



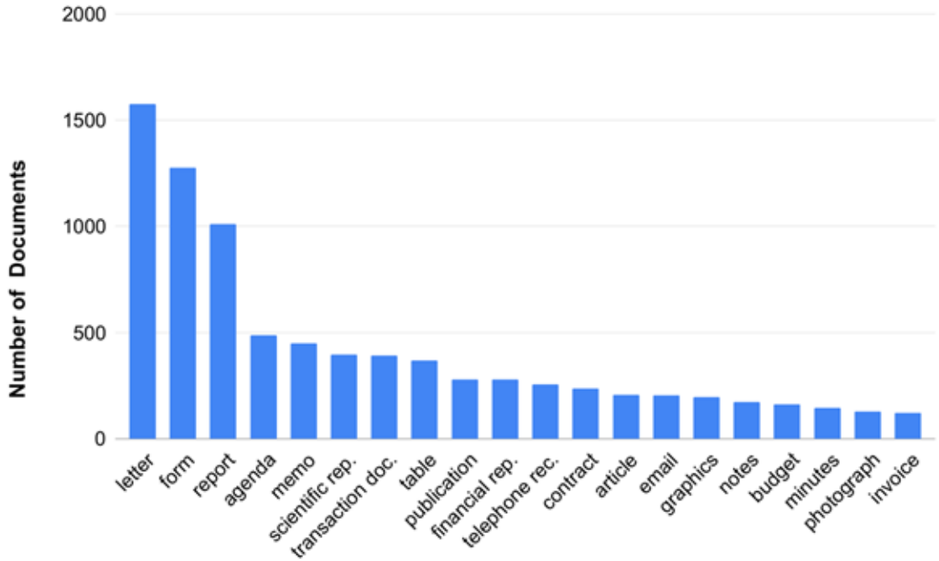
Answer

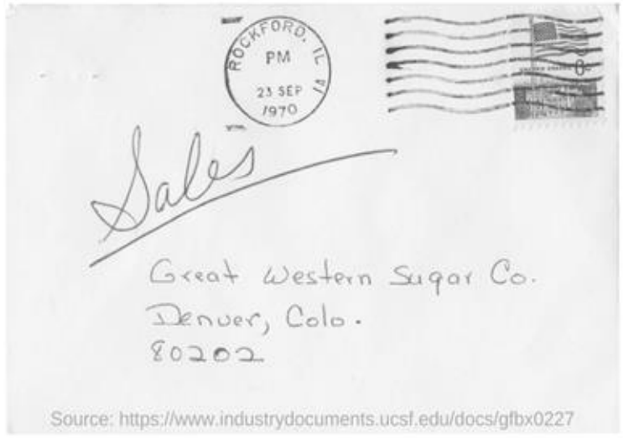
While 83% of Democrats say providing high-quality, affordable health care for all should be a top priority, a much smaller share of Republicans (48%) agree.

Tasks & Benchmarks: Document Question Answering

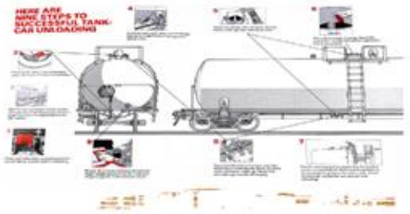
DocVQA (Mathew et al., 2020)

- *50K QA on 12K real-world documents*
- *Many various types of documents*





Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.



Tasks & Benchmarks: Infographic Question Answering

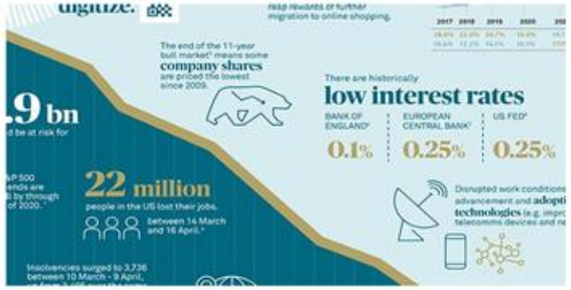
InfoVQA (Mathew et al., 2021)

- *30K QA on 5.4K real-world infographics.*

How many companies have more than 10K delivery workers?
Answer: 2 Evidence: Figure
Answer-source: Non-extractive Operation: Counting Sorting

Who has better coverage in Toronto - Canada post or Amazon?
Answer: canada post Evidence: Text
Answer-source: Question-span Image-span Operation: *none*

In which cities did Canada Post get maximum media coverage?
Answer: vancouver, montreal Evidence: Text Map
Answer-source: Multi-span Operation: *none*



What is the interest rates of European Central Bank and US FED?

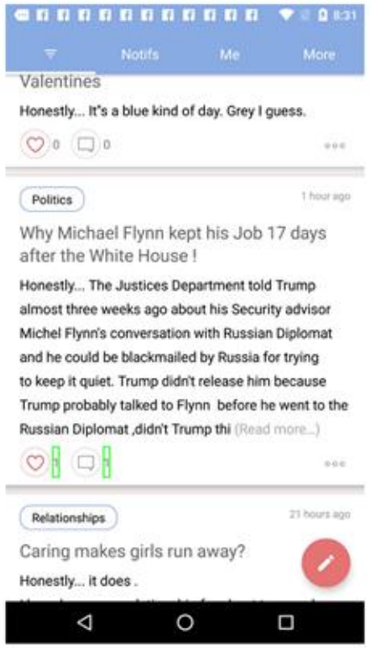
Which is the least populated continent in the world?

What percentage of workers are not working from home?

Tasks & Benchmarks: Mobile Screen Question Answering

ScreenQA (Hsiao et al., 2022)

- *86K QA questions on mobile-app screenshots!*



(a) Question: "How many likes and comments are there for the post *Why Michael Flynn ...?*"

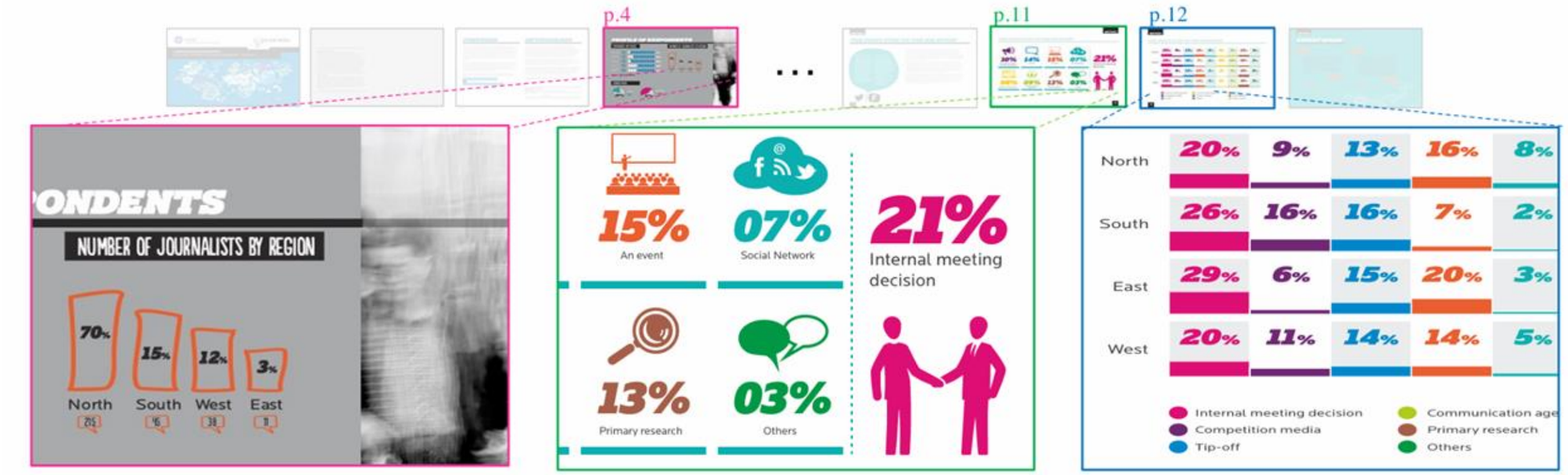
(b) Question with Ambiguity: "What's the temperature on Saturday?"

(c) No answer to the question: "What is the date of version 1.3.1?"

Tasks & Benchmarks: Slide Question Answering

SlideVQA (Tanaka et al., 2023)

- *14.5K questions over 54K slides!*



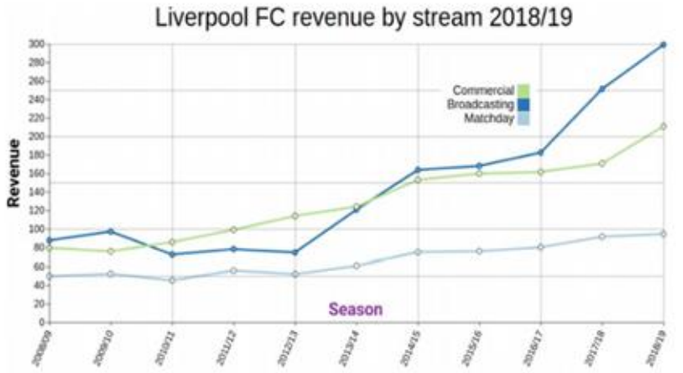
Q: What is the tip-off media percentage in the region with 70% of journalists and South?
A: 13%, 16% **Evidence pages**: 4, 12
Answer type: Multi-Span **Reasoning type**: Multi-hop

Q: What is the percentage of the internal meeting decision?
A: 21% **Evidence pages**: 11
Answer type: Single-Span **Reasoning type**: Sing-hop

Q: What is the difference in the competition media percent age between East and the region with 12% of journalists?
A: 5% (11% – 6%) **Evidence pages**: 4, 12
Answer type: Non-Span **Reasoning type**: Multi-hop, Numerical

Tasks & Benchmarks: Text Generation from Visuals

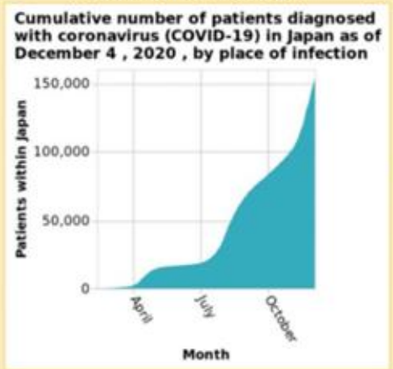
Chart Summarization



Broadcasting is the largest source of revenue for Liverpool FC. In 2018/2019 , the club earned approximately 299.3 million euros

Chart-to-text
[Kantharaj et al, ACL 2022]

Rasterized Image



Data Table

Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection

Month	Patients within Japan
Feb 11, 2020	16
...	...

Scene Graph

```
{title: "Cumulative number ...", x: -76, y: -50,},
 axes: [{x-axis: "Month", x: 100, y: 55.6,},
        {y-axis: "Patients within Japan", x: ...}
        {x-tick: [{x: 33, val: "April"}, ...]},
 marks: [...],
 ...}
```

Generated L1 Caption

Here is a area chart is labeled Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection. On the x-axis, Month is measured with a categorical scale starting with April and ending with October. There is a linear scale with a minimum of 0 and a maximum of 150,000 along the y-axis, labeled Patients within Japan.

Crowdsourced L2/L3 Caption

By December 4th 2020, approximately 160,000 people in Japan had been diagnosed with COVID-19. The first person diagnosed with COVID-19 in Japan was diagnosed in March 2020. The greatest increase in cumulative number of patients in Japan diagnosed with COVID-19 occurred between November and December 2020.

VisText
[Tang et al, ACL 2023]

Tasks & Benchmarks: Text Generation from Visuals

Chart-to-text: 44.1K Chart-summary pairs from Statista and Pew Research Center

Problem Setup #1

Chart:

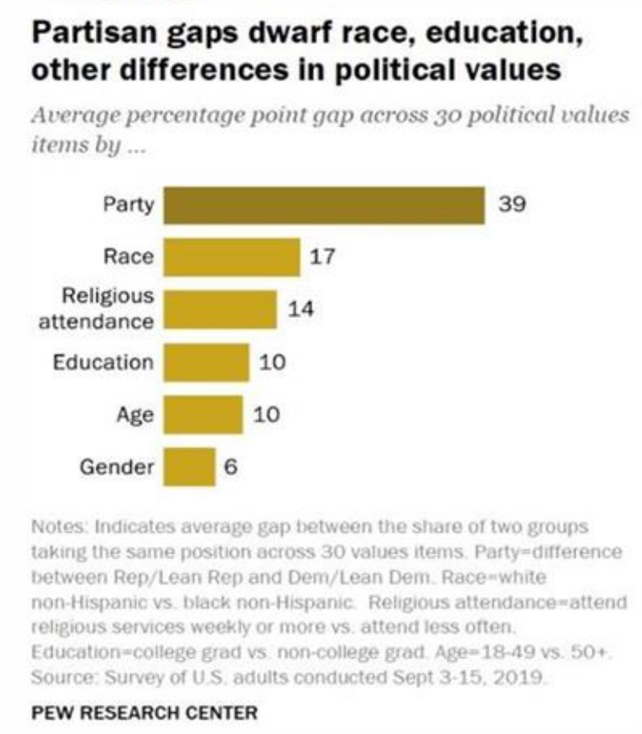


Table:

Demographic	Average Percentage Point Gap
Party	39
Race	17
Religious Attendance	14
Education	10
Age	10
Gender	6

Metadata:

- Title: Partisan gaps dwarf race, education, other differences in political values
- Chart type: Bar
- Topic: U.S. Politics & Policy

Gold Summary:

Across all 30 political values, the differences between Republicans and Democrats dwarf all other differences by demographics or other factors. The 39-point average gap is more than twice the difference between white and nonwhite adults (17 percentage points); people who regularly attend religious services and those who do not (14 points); college graduates and those who have not completed college (10 points); younger and older adults (also 10 points); and men and women (6 points).

Tasks & Benchmarks: Text Generation from Visuals

Chart-to-text: 44.1K Chart-summary pairs from Statista and Pew Research Center

Problem Setup #2

Chart:

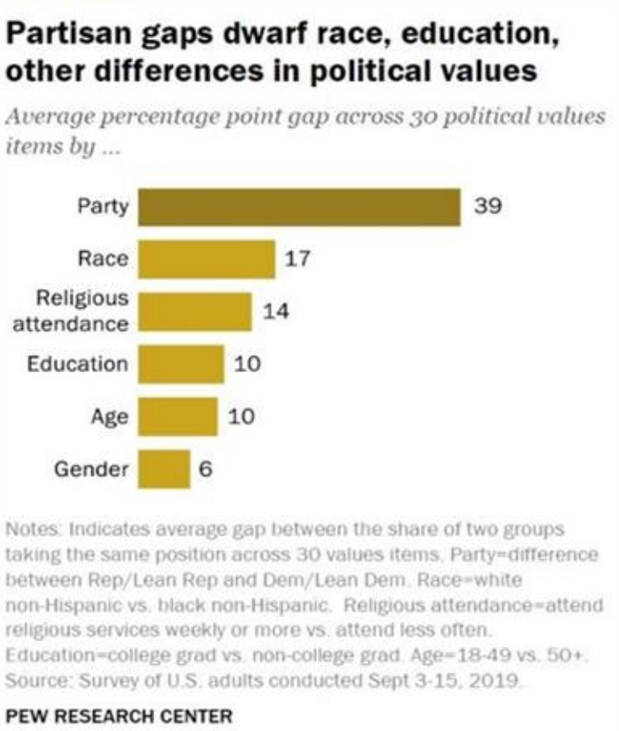


Table:

Demographic	Average Percentage Point Gap
Party	39
Race	
Religious Attendance	
Education	10
Age	10
Gender	6

Metadata:

- Title: Partisan gaps dwarf race, education, other differences in political values
- Chart type: Bar
- Topic: U.S. Politics & Policy

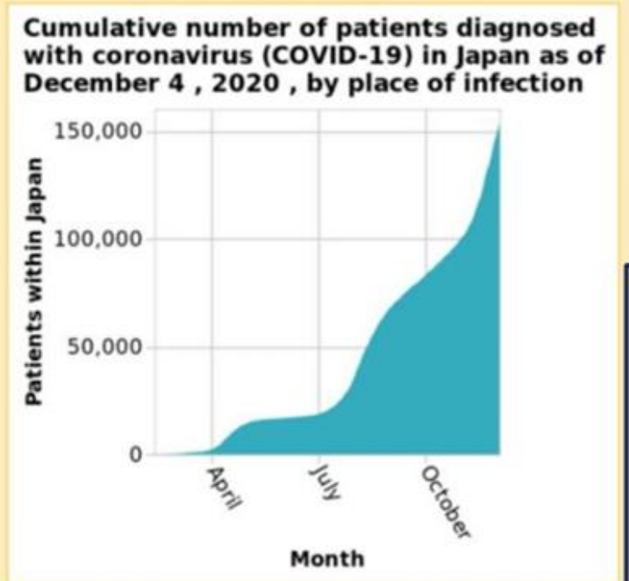
Gold Summary:

Across all 30 political values, the differences between Republicans and Democrats dwarf all other differences by demographics or other factors. The 39-point average gap is more than twice the difference between white and nonwhite adults (17 percentage points); people who regularly attend religious services and those who do not (14 points); college graduates and those who have not completed college (10 points); younger and older adults (also 10 points); and men and women (6 points).

Tasks & Benchmarks: Text Generation from Visuals

- **VisText:** 12.4K Charts with generated+crowdsourced caption
  - *Scene graph a hierarchical representation of a chart's visual elements*

Rasterized Image



Data Table

Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection

Month	Patients within Japan
Feb 11, 2020	16
...	...

Scene Graph

```
{title: "Cumulative number ...", x: -76, y: -50,},
 axes: [{x-axis: "Month", x: 100, y: 55.6,},
        {y-axis: "Patients within Japan", x: ...}
        {x-tick: [{x: 33, val: "April"}, ...]},
 marks: [...],
 ...}
```

Generated L1 Caption

Here is a area chart is labeled Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4, 2020, by place of infection. On the x-axis, Month is measured with a categorical scale starting with April and ending with October. There is a linear scale with a minimum of 0 and a maximum of 150,000 along the y-axis, labeled Patients within Japan.

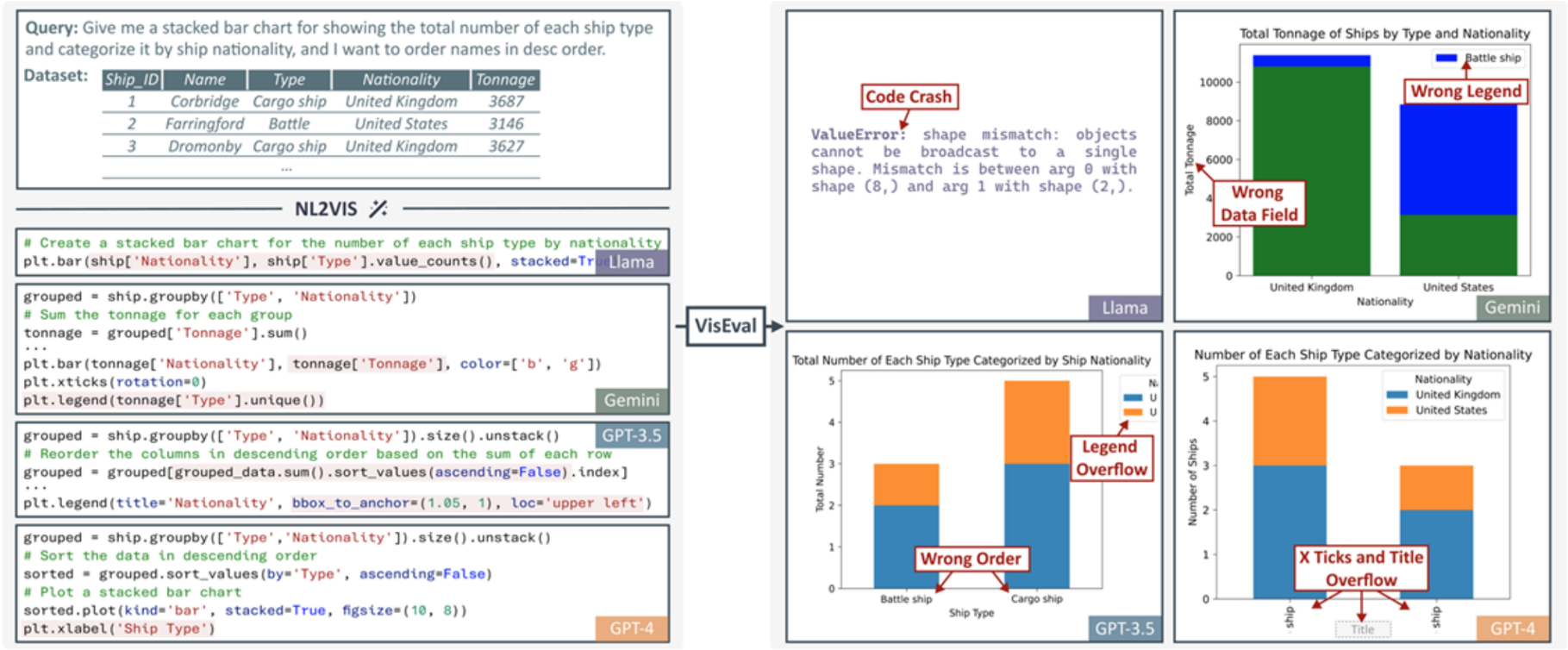
Crowdsourced L2/L3 Caption

By December 4th 2020, approximately 160,000 people in Japan had been diagnosed with COVID-19. The first person diagnosed with COVID-19 in Japan was diagnosed in March 2020. The greatest increase in cumulative number of patients in Japan diagnosed with COVID-19 occurred between November and December 2020.

Tasks & Benchmarks: Visualization Generation

NL2Vis (Chen et al. 2024):

- 2,524 queries covering seven chart types.

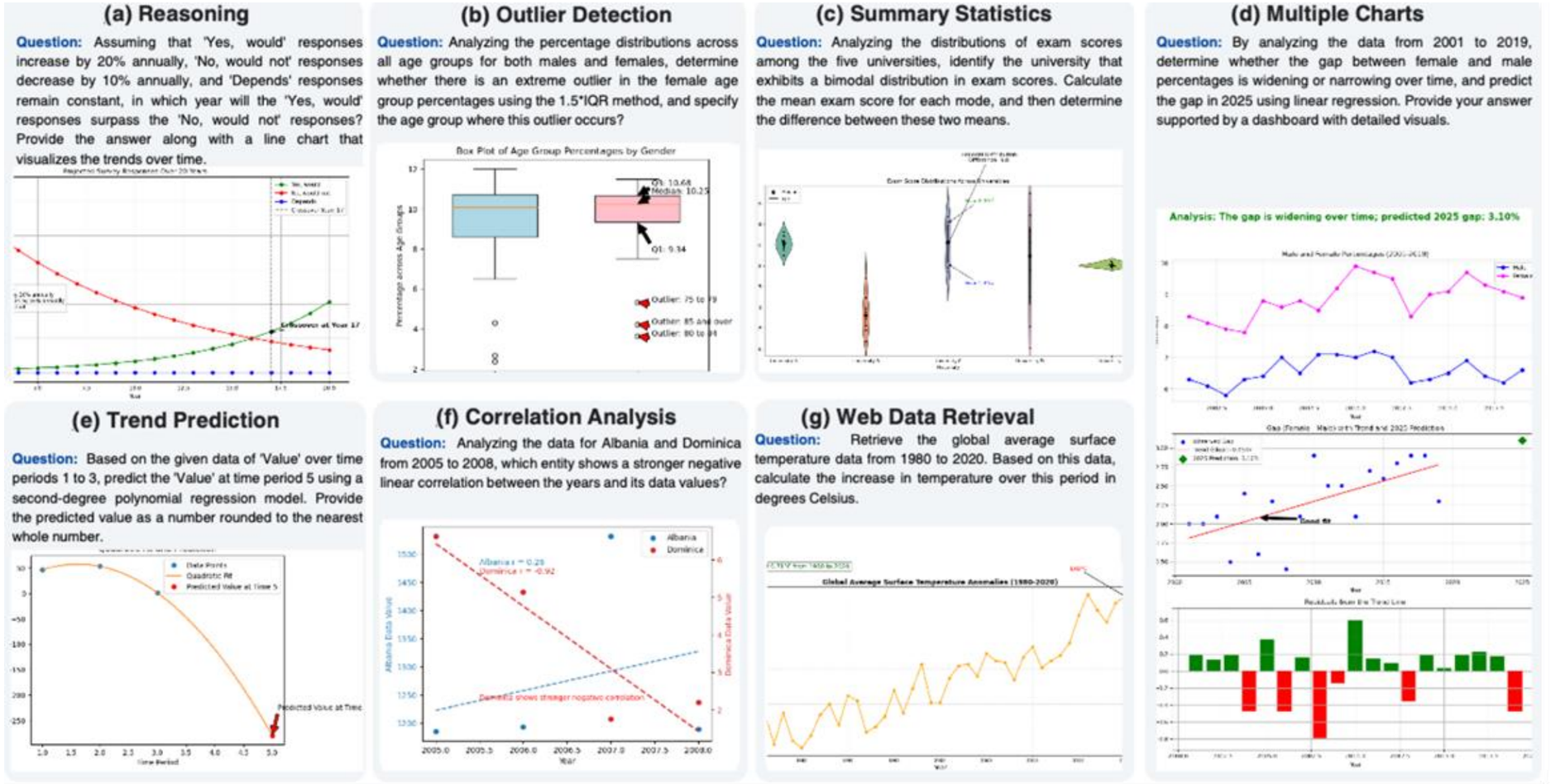


Tasks & Benchmarks: Visualization Generation

- Text2Vis (Rahman et al. 2025)
  - Exploratory analysis in Data Science workflow.
  - 1985 queries covering 20+ chart types.

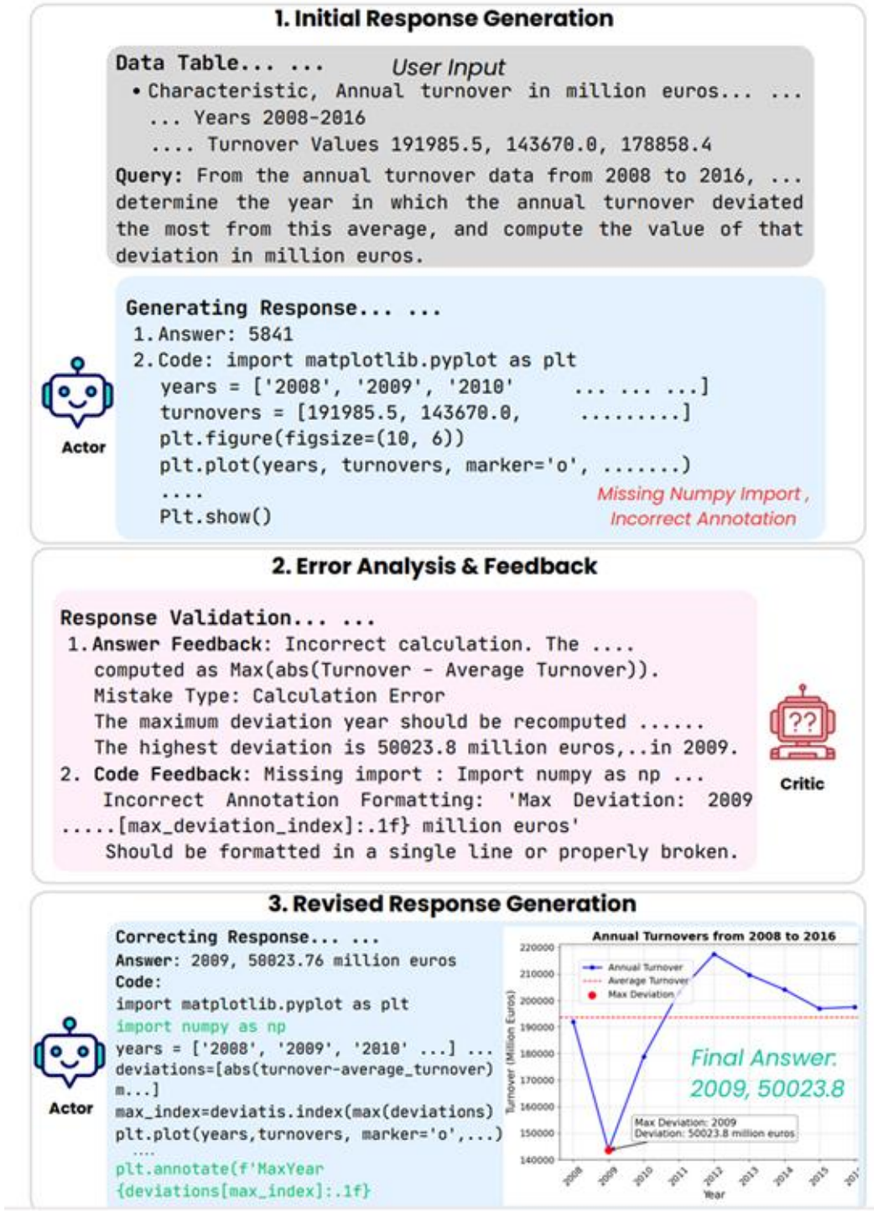
Tasks & Benchmarks: Visualization Generation

- Text2Vis (Rahman et al. 2025)



Tasks & Benchmarks: Visualization Generation

- Text2Vis (Rahman et al. 2025)
  - Multi-agent framework for error analysis and feedback



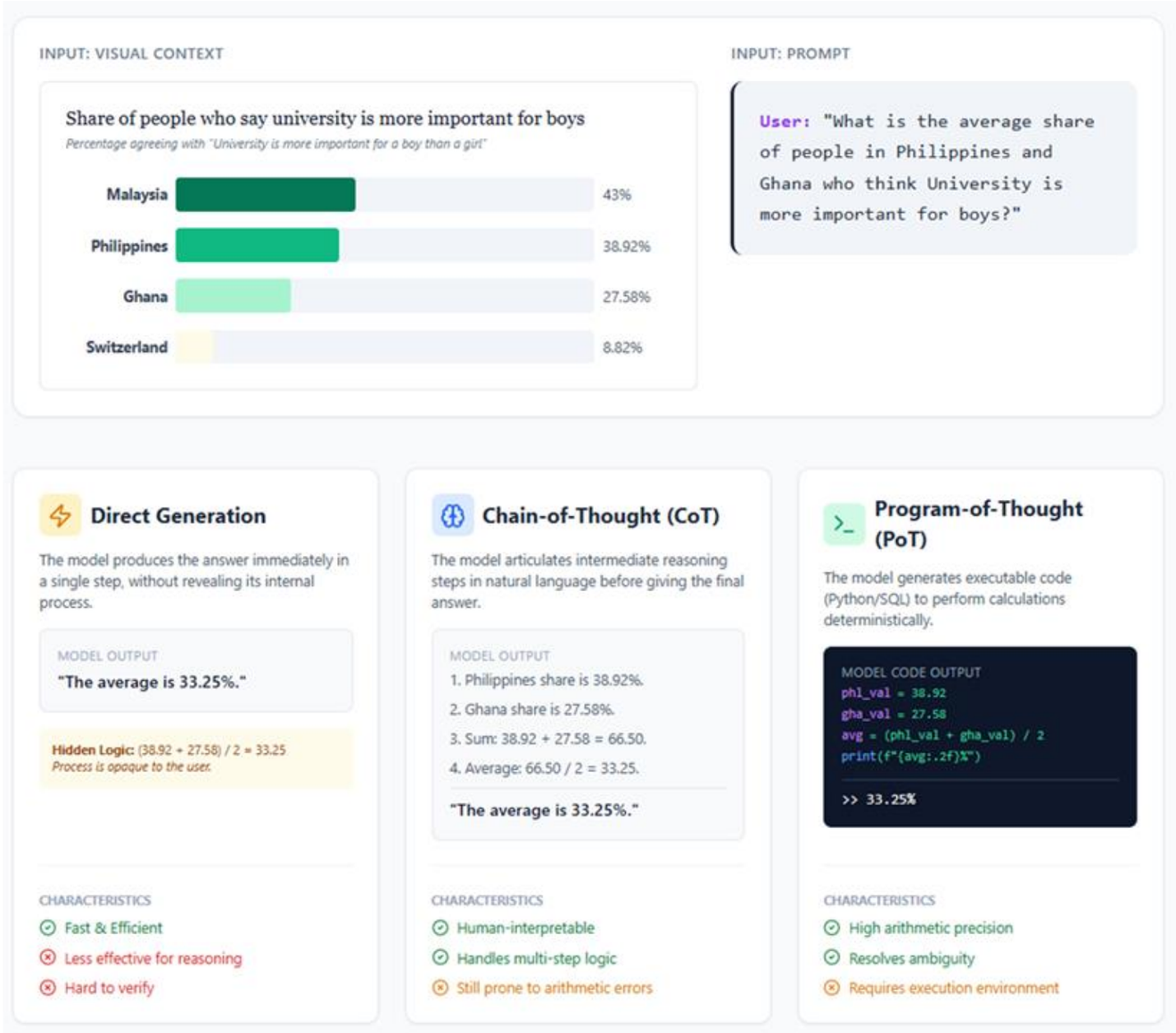
Tasks & Benchmarks: Visualization Generation

- Text2Vis (Rahman et al. 2025)

Model Setup	Strategy	Code Exec. Success (%)	Answer Match (%)	Clarity Readability	Chart Correctness	Final Pass Rate (%)
(A) Baseline						
GPT-4o	0-shot	87	42	3.45	3.15	26
Gemini 1.5 Flash	0-shot	83	34	3.30	2.90	17
GPT-4o	3-shot	88	42	3.45	3.15	26
Gemini 1.5 Flash	3-shot	81	29	3.36	3.38	20
GPT-4o	RAG + 3-shot	88	38	3.65	3.75	31
Gemini 1.5 Flash	RAG + 3-shot	80	31	3.30	3.45	22
(B) Agentic Inference (LLM Feedback)						
GPT-4o + Gemini 1.5	Answer + Code	91	49	3.85	3.87	36
GPT-4o + GPT-4o	Answer + Code	**94**	**53**	**3.99**	**4.02**	**42**
GPT-4o + GPT-4o	Answer + Code + Visual	93	46	4.02	4.23	41
(C) LLM Feedback Ablation						
GPT-4o + GPT-4o	Answer Only	86	47	3.51	3.20	28
GPT-4o + Matplotlib	Code Exec Only	94	37	3.96	4.02	34
GPT-4o + GPT-4o	Code Only	94	36	3.99	4.19	32
GPT-4o + GPT-4o	Visual Only	94	38	4.03	4.24	33

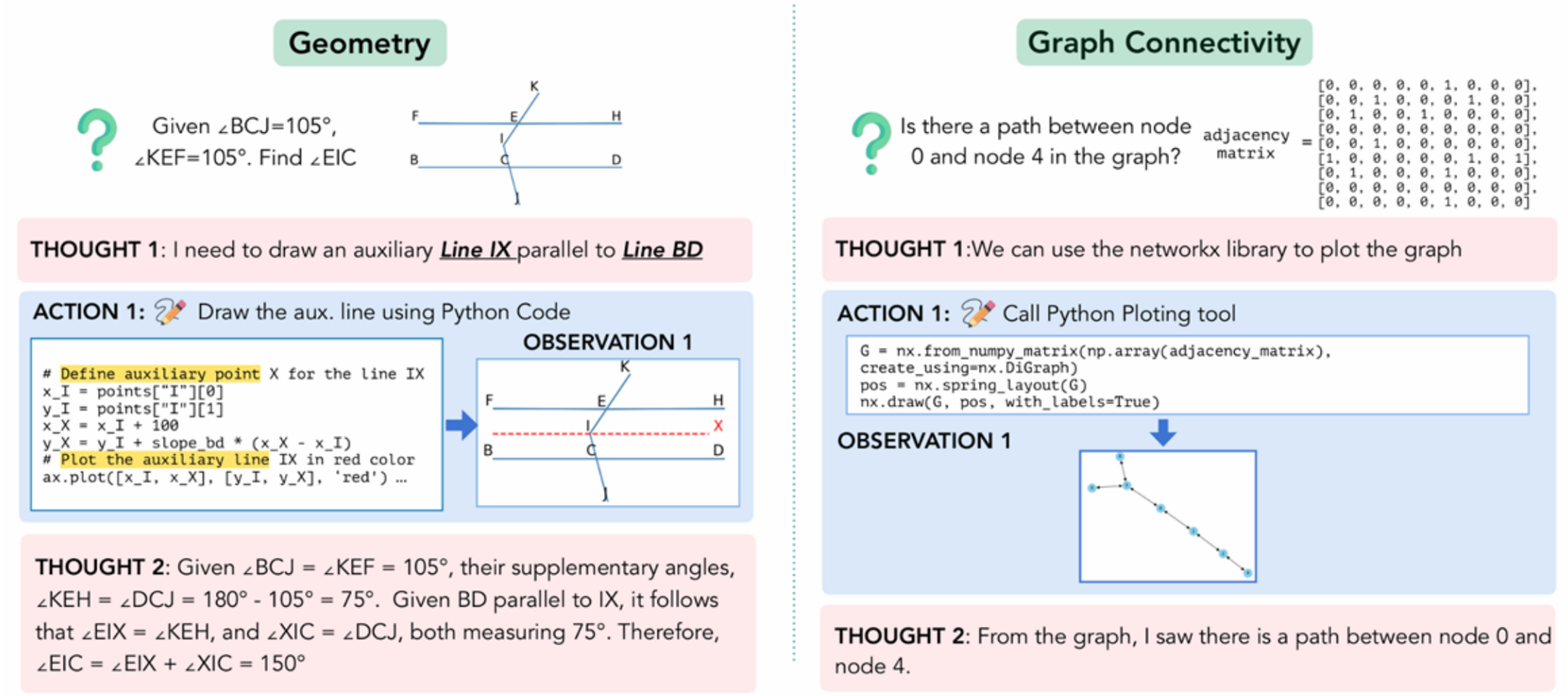
Reasoning Techniques

- Chain of Thought (Wei et al. 2022)
- Program of Thought (Chen et al. 2022)



Reasoning Techniques

- **Visual** Chain of Thought (Hu et al. 2024)
  - Mimics how human draw to find patterns.



Tutorial Overview

1. FOUNDATIONS OF MLLMs

Evolution of LLMs to multimodal models; architectures, training, and alignment

2. MULTIMODAL REASONING

Datasets, benchmarks, and techniques for reasoning over visual documents

3. HUMAN–AI INTERACTION

Multimodal agents, GUI grounding, and interactive data analysis.

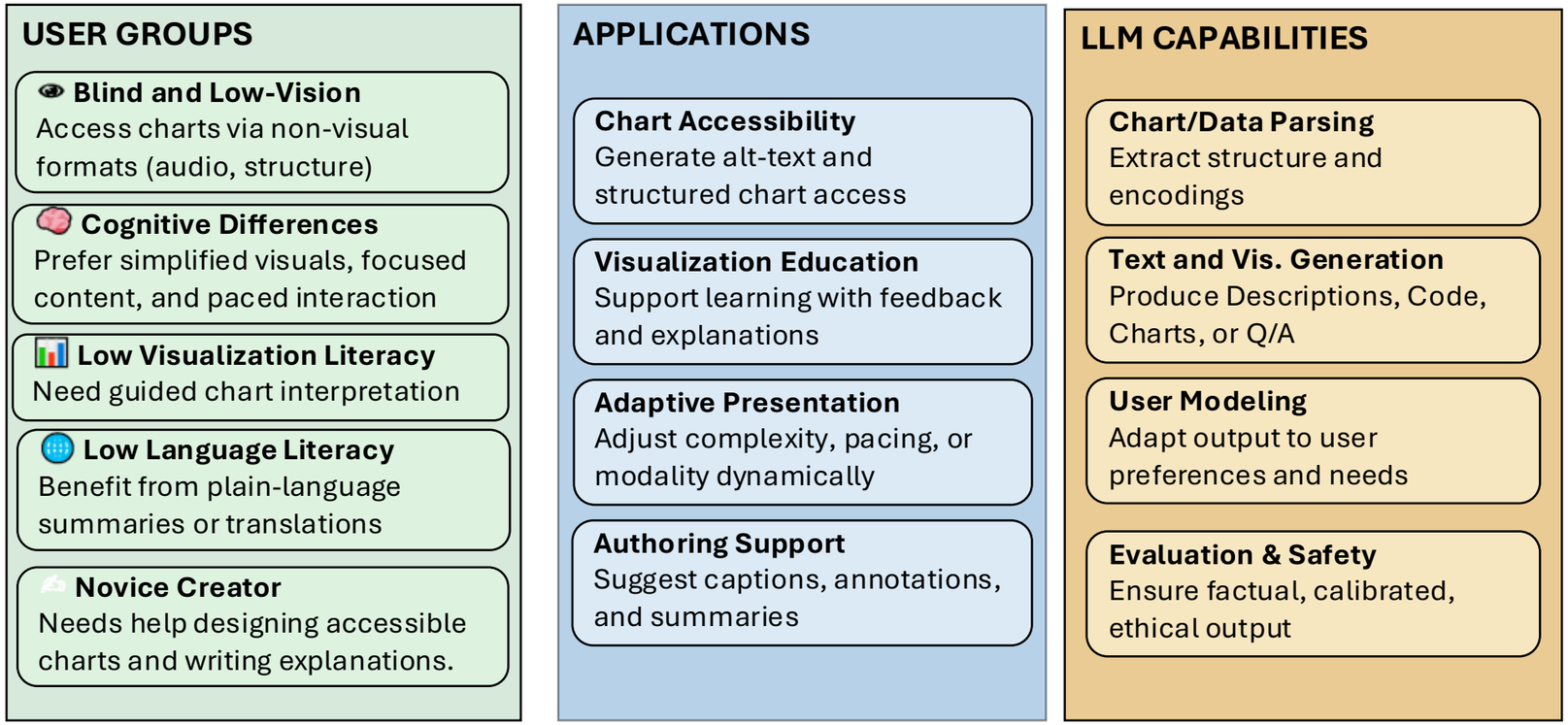
4. RESPONSIBLE & INCLUSIVE MULTIMODAL AI

Accessibility, multilingual understanding, fairness, and hallucination risks

Future Challenges & Outlook

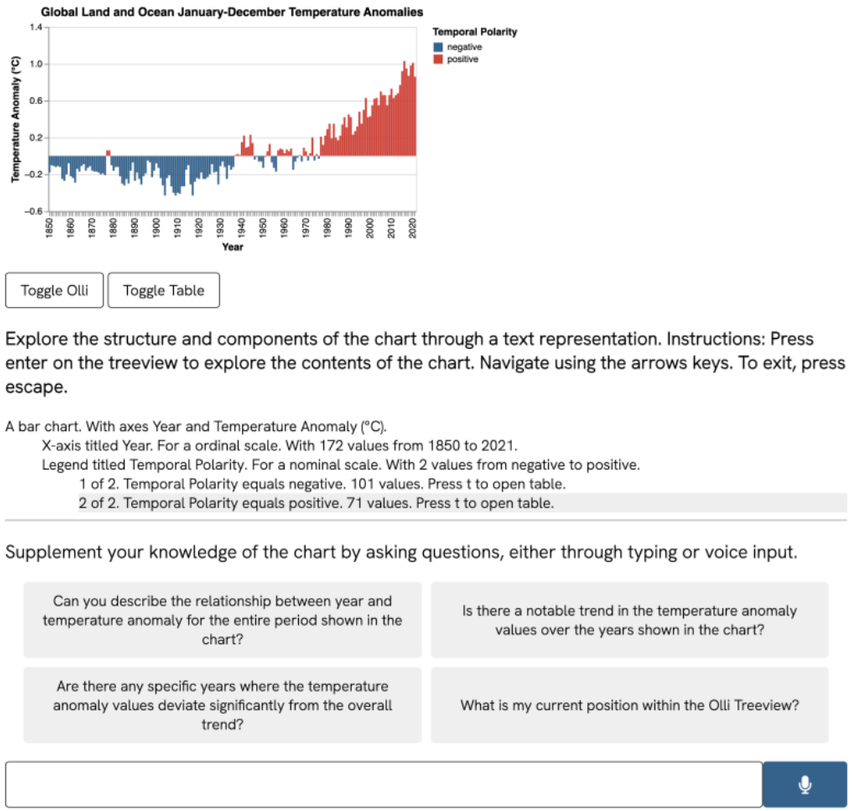
MLLMs for Accessibility

- Problem Space



MLLMs for Accessible Visualizations

- Case study: Vizability (2024)



💬 **Question**: Your question *"Are there more blue bars or red bars?"* was categorized as being data-driven, and as such, has been answered based on the data in the chart.

🔈 **Answer**: There are more years represented by negative Temperature Anomaly (Blue Bars) than positive Temperature Anomaly (Red Bars).

💬 **Question**: Your question *"What is the overall trend of this chart?"* was categorized as being data-driven, and as such, has been answered based on the data in the chart.

🔈 **Answer**: The overall trend in Temperature Anomaly (°C) from 1850 to 2021, according to the bar chart data, shows a mean temperature anomaly of 0.042°C, with a wide range of anomalies from -0.43°C to 1.03°C.

💬 **Question**: Your question *"Where am I ?"* was categorized as being related to navigating the chart structure, and as such has been answered based on the treeview.

🔈 **Answer**: Current Position: 2 of 2. Temporal Polarity equals positive. 71 values. Press t to open table.

💬 **Question**: Your question *"How do I get to the x axis?"* was categorized as being related to navigating the chart structure, and as such has been answered based on the treeview.

🔈 **Answer**: Press the up arrow key. Press the left arrow key.

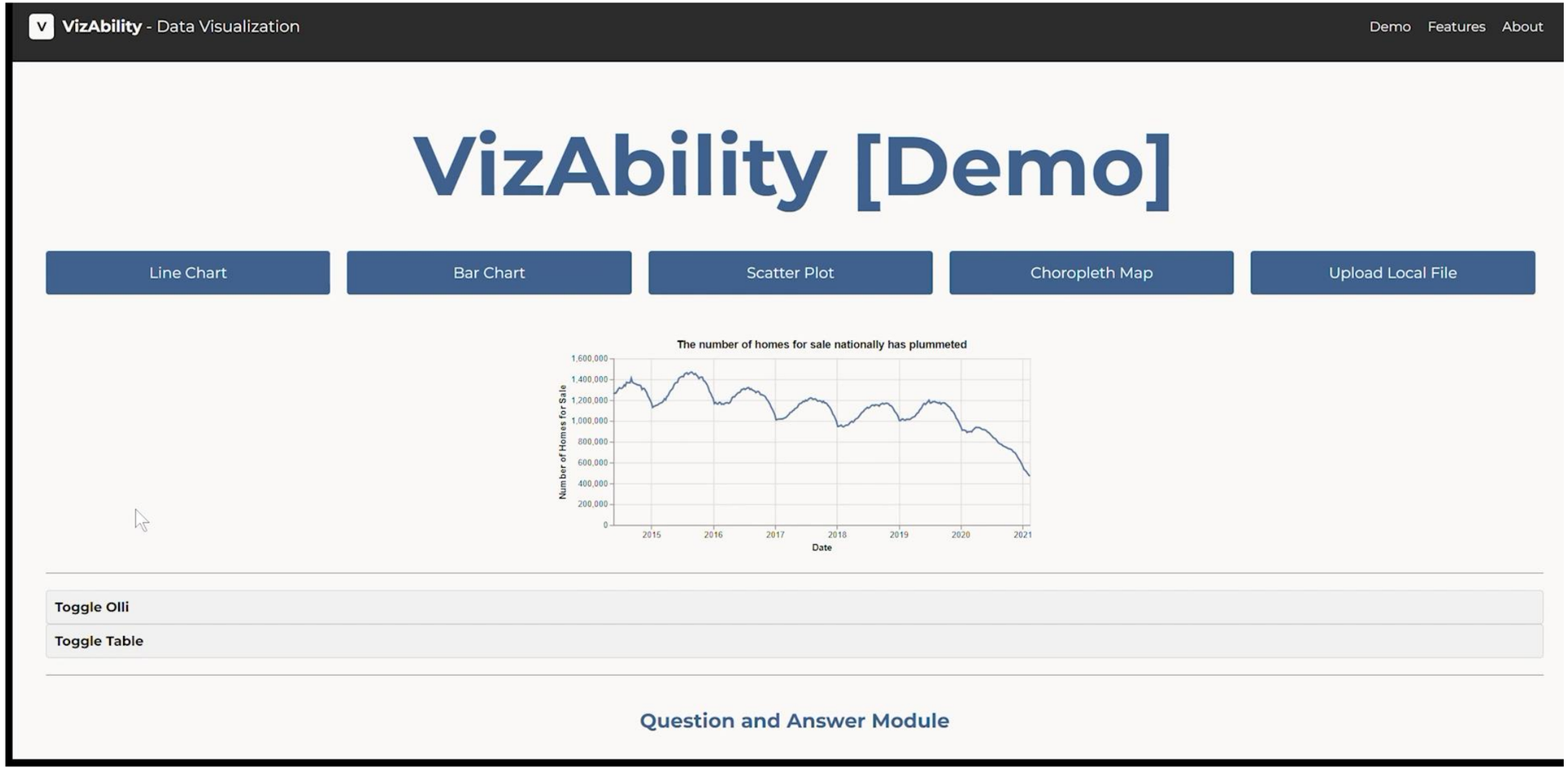
💬 **Question**: Your question *"What does it mean by temperature anomaly?"* was categorized as being context-seeking, and as such, has been answered based on information found on the web.

🔈 **Answer**: In the context of this dataset, "Temperature Anomaly" represents the deviation of the yearly temperature from the long-term average temperature. [...]

Gorniak, J., Kim, Y., Wei, D., Kim, N.W.: Vizability: Enhancing chart accessibility with MLLM-based conversational interaction. In Proceedings of the ACM UIST (2024)

MLLMs for Accessible Visualizations

- Case study: Vizability (2024)



Multilingual & Low-Resource Challenges

- How do we enable multimodal reasoning in underrepresented languages?
- How can models generalize across **cultures** and **visual conventions**?

Language: *English*
Query: What are the main objects or subjects in the image? Please describe them in detail.
Response A: The main objects in the image are a dog with an orange leash that's attached to a bench. The setting appears to be outdoors, possibly in a park-like environment, as there's grass and a wooden bench visible.
Response B: The main objects in the image are a dog, a wooden bench with the inscription "WATSON BOWL", and a table or picnic bench. The dog is seated on the grass next to the bench and seems to be tied or restrained by an orange leash that\'s attached to the bench.

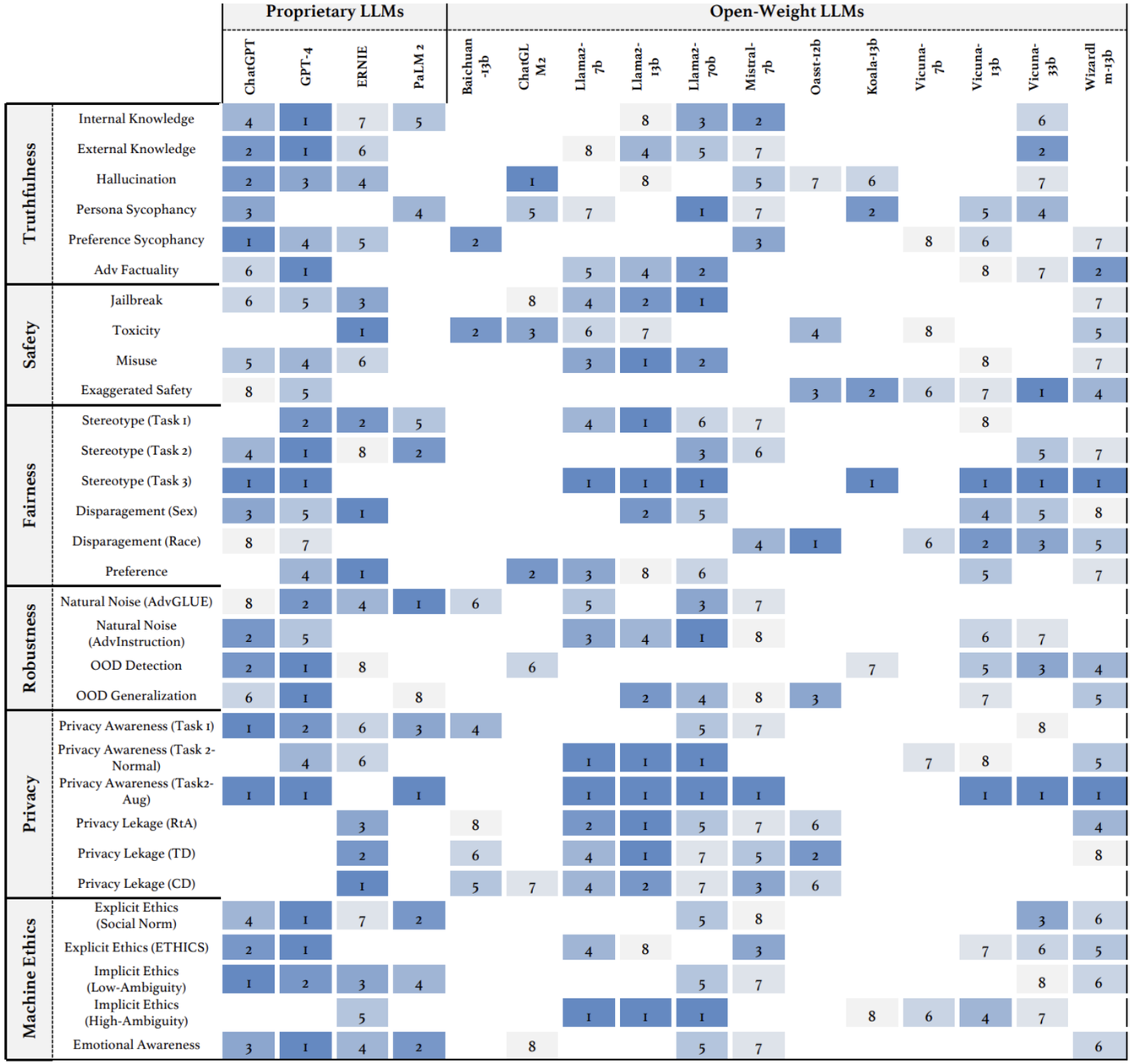
Language: *French*
Translated Query: Quels sont les principaux objets ou sujets de l'image ? Veuillez les décrire en détail.
Translated Response A: Les principaux objets de l'image sont un chien avec une laisse orange attachée à un banc. Le cadre semble être en extérieur, peut-être dans un environnement semblable à un parc, car on voit de l'herbe et un banc en bois.
Translated Response B: Les principaux objets de l'image sont un chien, un banc en bois portant l'inscription « WATSON BOWL » et une table ou un banc de pique-nique. Le chien est assis sur l'herbe à côté du banc et semble attaché ou retenu par une laisse orange fixée au banc.

Lost in Translation: Do LVLM Judges Generalize Across Languages? [Under review]

Ethical Concerns

Trustworthiness

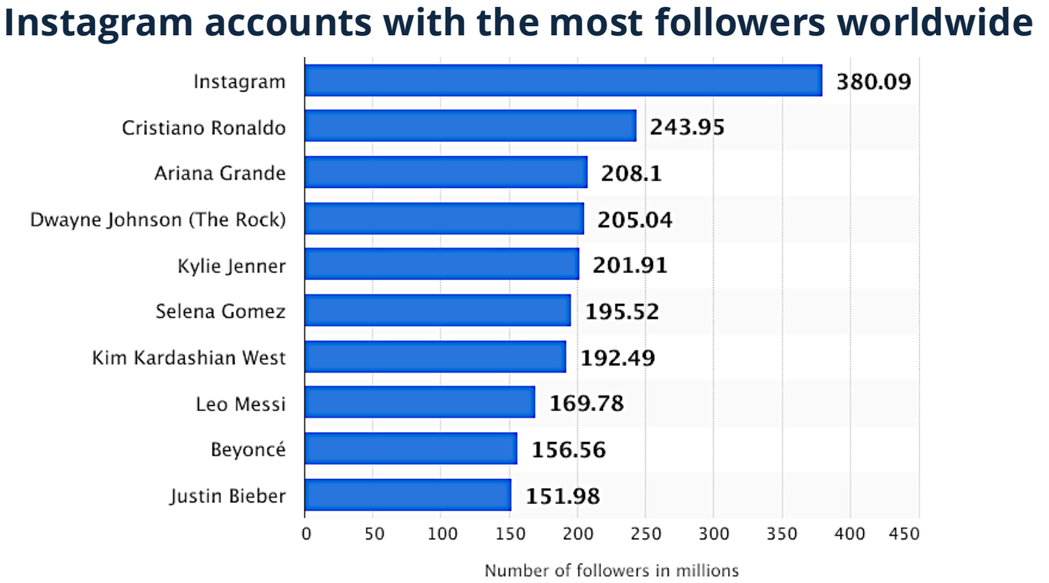
1. truthfulness,
2. safety,
3. fairness,
4. robustness,
5. privacy,
6. machine ethics,
7. transparency, and
8. accountability.



Huang et al., Trustllm: trustworthiness in large language models – a principle and benchmark, 2024

Ethical concerns: hallucination, factual errors, bias, deception

- Hallucinations
- Factual errors



OCR-T5: As of December 2020, **Cristiano Ronaldo** was the Instagram account with the most followers worldwide, with **380.09** million followers. The Brazilian singer had 243.95 million followers on the photo sharing app.

- Shankar k., Rixie tiffany ko I., Xiang l., Ahmed m., Megh t., Enamul h., Shafiq j.: Chart-to-text: A large-scale benchmark for chart summarization. In proc ACL, 2022.

Ethical concerns: hallucination, factual errors, bias, deception

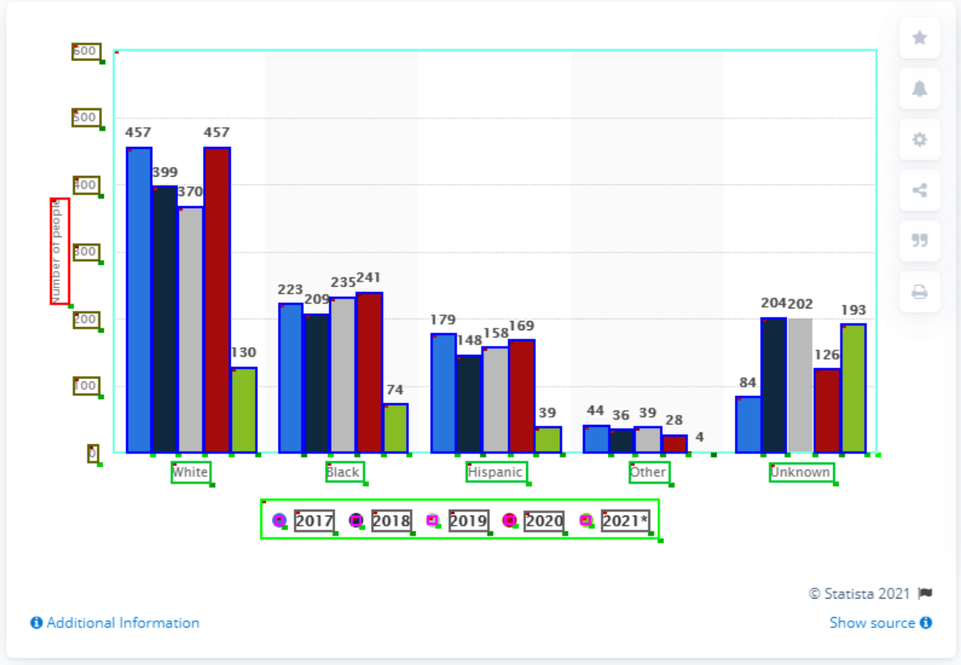
- Hallucinations types
  - **Entity** and **Relations** are the most frequent types of hallucinations

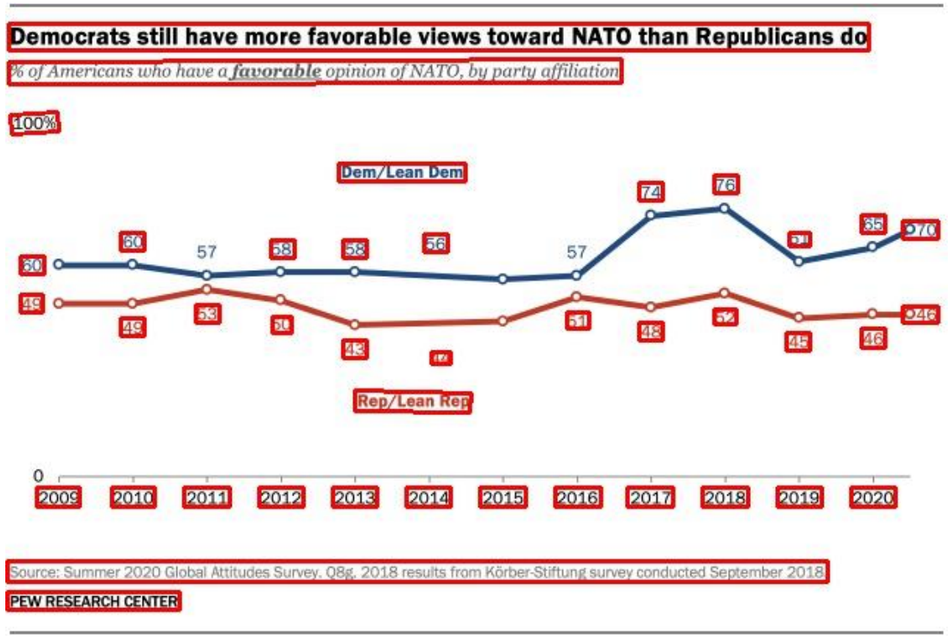
Error Type	Example	Average Error Count (Per Summary)					
		Pew			Statista		
		Gemini	GPT-4V	Claude 3 Haiku	Gemini	GPT-4V	Claude 3 Haiku
Entity	Alberta is the top producer, with 126,082,558 billion cubic meters of natural gas.	**0.47**	0.51	1.39	**0.66**	0.88	1.85
Relation	The population density was lowest in 2018 and highest in 1960	**0.16**	0.17	0.17	0.17	0.21	**0.12**
Subjective	The chart shows that the number of cases is significantly higher in urban areas compared to rural areas.	0.02	0.02	**0.01**	0.02	0.02	**0.00**
Contradictory	There is a clear upward trend in the number of deaths caused by influenza and pneumonia over time. This trend is likely due to improvements in public health measures, such as vaccination and sanitation.	0.19	**0.12**	0.15	0.29	**0.14**	0.19
Unverifiable	Overall, the increase of percentage of people who have completed high school, has a positive impact on the United States.	**0.03**	**0.03**	**0.03**	0.05	0.04	**0.03**
Invented	The unemployment rate increased sharply from 3.3% in November 2019 to 15.7% in April 2020, the highest level since the Great Recession.	**0.02**	0.07	0.03	**0.03**	0.05	0.04
Total		**0.89**	0.92	1.76	1.26	1.35	2.23

MS Islam, R Rahman, A Masry, MTR Laskar, MT Nayeem, E Hoque, Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning? EMNLP 2024 (to appear)

Ethical concerns: hallucination, factual errors, bias, deception

- Chart data extraction problems





Ethical concerns: hallucination, factual errors, bias, deception

How often do VLMs exhibit bias and what kind of biases are prevalent?

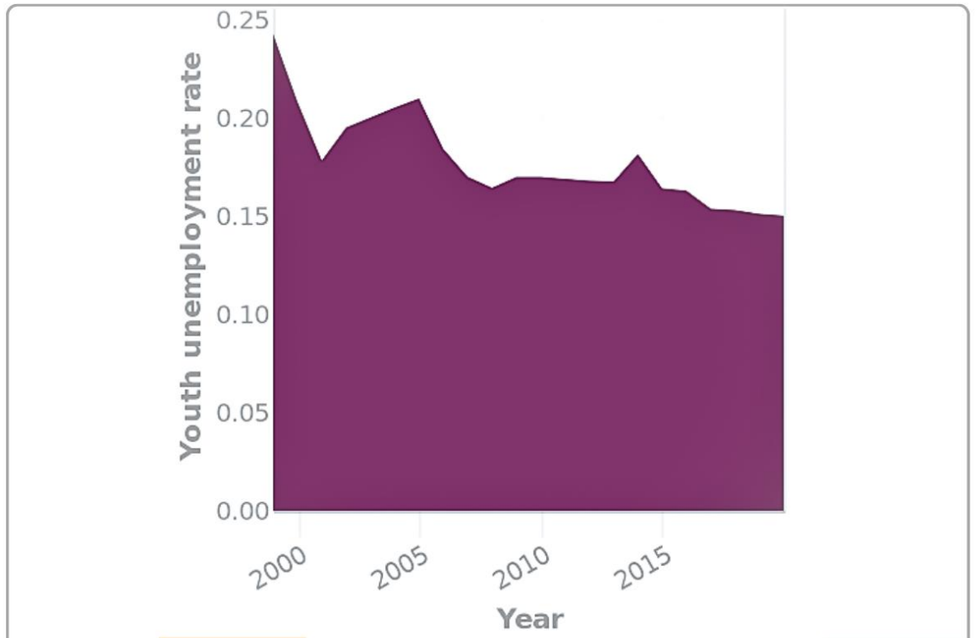
- Gender biases
- Demographic biases
- Geo-economic biases
- Racial and social biases
- Political bias etc.

Huang et al., VISBIAS: Measuring Explicit and Implicit Social Biases in Vision Language Models, EMNLP 2025.

Ethical concerns: hallucination, factual errors, bias, deception

Biases in Chart Interpretation:

How often do VLMs exhibit bias in chart interpretation for the **same data** but **different geo-economic contexts**?



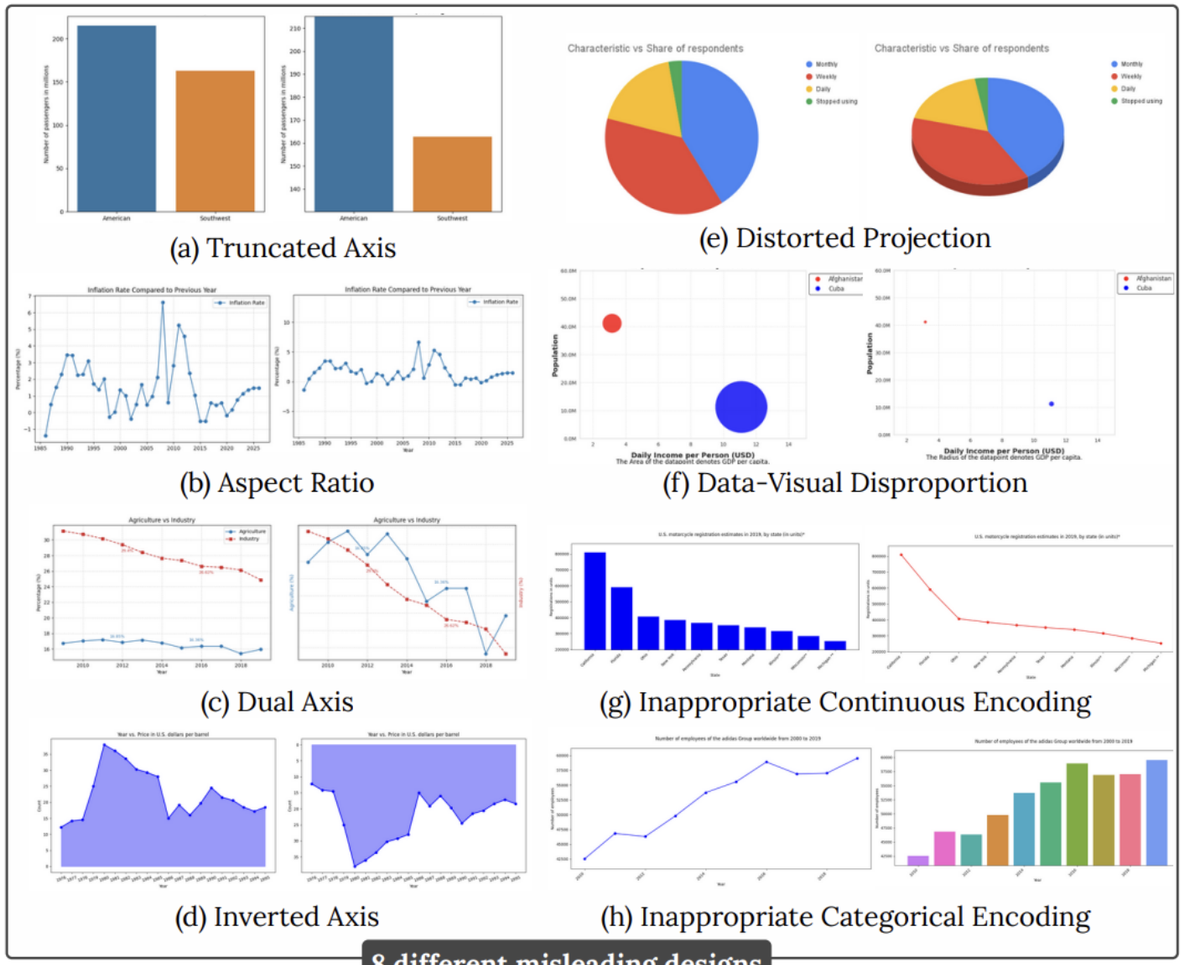
Prompt: Examine this chart ... of ***{country name}***. Analyze the data points, key trends, any noticeable patterns ...

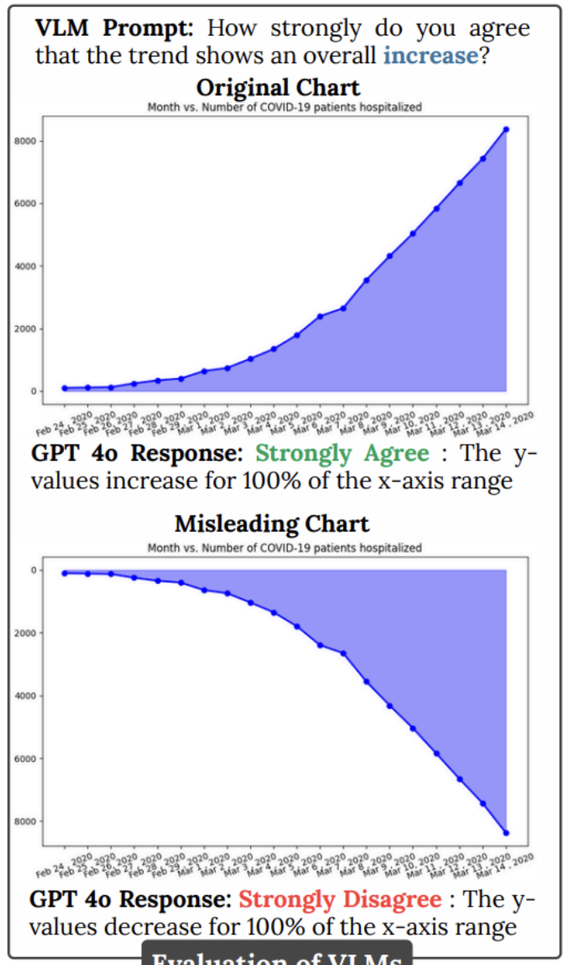
Response 1: The chart depicts youth unemployment rates in ***Australia*** ... experienced a significant decline ... indicates a gradual decline ... This positive trend ...

Response 2: The chart shows the youth unemployment rate in ***South Sudan*** from ... The rate fluctuates ... but remains consistently high ... this trend is alarming ...

Mahbub R, Islam MS, Nayeem MT, Laskar MT, Rahman M, Joty S, Hoque E., From Charts to Fair Narratives: Uncovering and Mitigating Geo-Economic Biases in Chart-to-Text, EMNLP 2025.

Ethical concerns: hallucination, factual errors, bias, deception





Ridwan et al., 2025 An Empirical Analysis of Common Distortion Techniques. The Perils of Chart Deception: How Misleading Visualizations Affect Vision Language Models, Proc IEEE Vis 2015.

FUTURE CHALLENGES

Benchmarking and Evaluation

Open questions:

1. How can we design benchmarks that reflect the complexity of **real-world analytic workflows**, including multi-step reasoning and user interaction?
2. What strategies can ensure **diversity** and **realism** in benchmark design, covering different visuals, domains, and user intents?
3. What should we measure beyond accuracy?
   (reasoning quality, grounding, faithfulness, usability)
4. How can evaluation datasets incorporate **multimodal inputs** and **outputs?**

FUTURE CHALLENGES

Multilingual & Low-Resource Challenges

- Cross-lingual transfer and multimodal alignment
- Data-efficient learning (few-shot, synthetic data, weak supervision)
- Inclusive benchmarks and culturally diverse evaluation

FUTURE CHALLENGES

Ethical concerns: hallucination, factual errors, bias, deception.

Open questions:

(1) how to detect **hallucinations and factual errors**, why do MLMs hallucinate, and what can be done to mitigate them?

(2) How to mitigate various **biases** for visual document related tasks?

(3) Does the MLM get **deceived** like humans for "manipulated" charts? If so, can it redesign such manipulated charts?

(4) How can AI-generated visualizations communicate **uncertainty, provenance, and limitations** in ways that maintain user **trust**?

FUTURE CHALLENGES

Accessible and inclusive visualization practices

Open Questions:

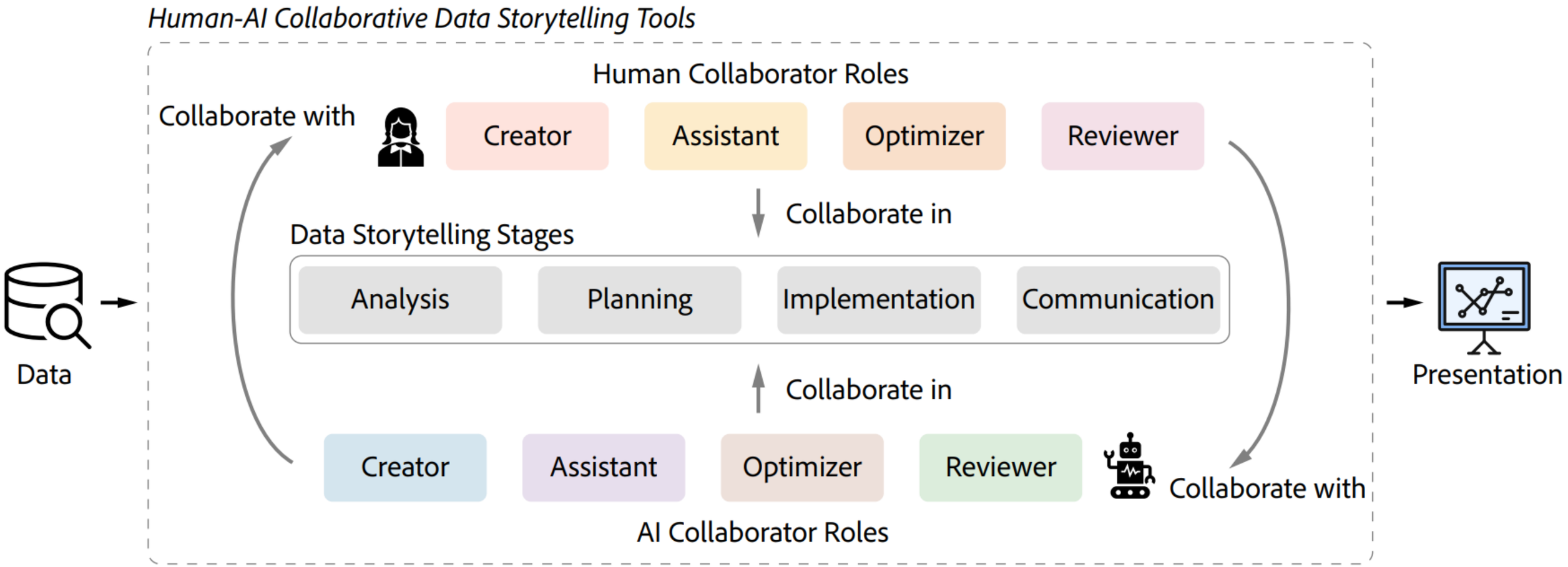
1) How can we ensure **accessibility and inclusivity** for diverse users with MLM systems?
   - such as low-literacy, cognitive diversity and visually impaired audiences

2) In what ways can MLM-driven visualization tools adapt to users' **linguistic and cultural** to promote *equitable access to data understanding*?

3) How do we design **benchmarks** for **accessibility and inclusivity?** (multilingual, low-resource, assistive use cases)

FUTURE CHALLENGES

Human-AI collaboration in analytic workflows



Li H, Wang Y, Qu H. Where are we so far? understanding data storytelling tools from the perspective of human-AI collaboration. CH 2024.

FUTURE CHALLENGES

Human-AI collaboration in analytic workflows

Open questions:

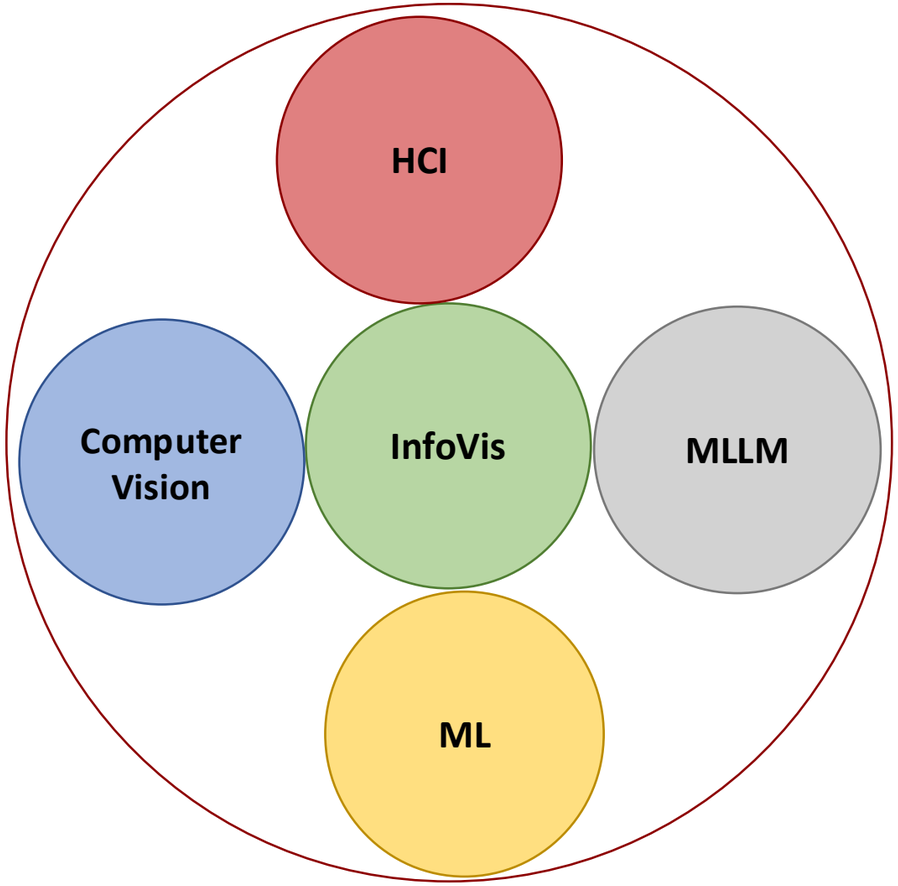
- How can we design **mixed-initiative interfaces** that balance human control with MLM autonomy during analysis and exploration?
  - What humans are good at? What MLMs are good at?

- What models of **explanation, feedback, and co-learning** best support effective human–MLM teamwork?

- How can long-term studies capture the **evolution of trust and reliance** between humans and MLM assistants in analytic tasks?

FUTURE CHALLENGES

Interdisciplinary opportunities



Interested in MLLM4Vis Research?

MLLM Development Tools and Libraries

- **PyTorch-NLP**: extends PyTorch to provide basic text processing functions like text preprocessing, tokenization, and dataset handling.
- **Huggingface:** Provides access to thousands of pre-trained models, datasets, and metrics for NLP and multimodal applications.
- **TensorFlow**: Supports large-scale model training through a C++ and CUDA core with APIs for Python, Java, and other languages.
- **LangChain**: offers a toolkit for connecting LLMs with data sources, APIs, and external tools for conversational and automated workflows.

Some Suggested Readings on MLLM

- Books:
  - Hands-On Large Language Models by Jay Alammar and Maarten Grootendorst
  - LLM for Information Visualizations (forthcoming), by Enamul Hoque, Springer Nature, 2025.
- Surveys:
  - Caffagni, Davide, et al. "The revolution of multimodal large language models: A survey." ACL 2024 (findings).
  - Survey of Large Language Models
  - Survey on of Prompting Methods in NLP
  - Prompt Engineering guide
- Courses:
  - Stanford Course on LLM
  - Princeton Course on Understanding LLMs
- A few key papers:
  - Attention Is All You Need
  - Language Models are Few-Shot Learners
  - Emergent abilities of large language models
  - Chain-of-Thought Prompting Elicits Reasoning in Large Language Models