

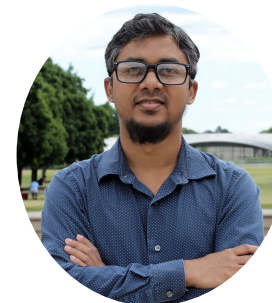
Multilingual and Multimodal LLMs in the Wild: Building for Low-Resource Languages



Firoj Alam
Senior Scientist, QCRI



Shammur Chowdhury
Scientist, QCRI



Enamul Hoque Prince
Associate Professor,
York University

Tutorial Resources

LLMs for Low Resource Languages

Home Speakers Content Reading list

Contents

- Efficient multimodality
- Data & evaluation
- Speech in the loop
- Hands-on resources

- Venue
- Date
- Speakers
- Citation
- Table of Content
- Reading List

Text · Speech · Vision

Low-resource focus

Hands-on recipes

Multilingual and Multimodal LLMs in the Wild

Build inclusive tri-modal systems that see, hear, and read for low-resource languages and dialects. We cover BLIP-2, LLaVA, KOSMOS-1, PaLM-E, PALO, Maya, SeamlessM4T, and AudioPaLM—plus efficient PEFT/adapters/MoE, culture-aware benchmarks, and speech→text→LLM pipelines.

View outline

Meet the speakers

Reading list



<https://mm-llms-in-the-wild.github.io>

Tutorial Outlines

Introduction

(low-resource and challenges)

Multilingual and Multimodal Models

(Capabilities for low-resource languages)

Multilingual & Multimodal Resource Development

(Low-cost pipeline, training, and benchmarking datasets)

Architectures and Efficient Training

(Adapter-based and parameter-efficient methods)

Speech-centric LLMs

(State-of-the-art models and curated resources)

Evaluation, Benchmarking Tools, Analysis

(Evolution of Evaluation metrics, benchmarking tools and resources)

Multilingual & Multimodal Resource Development

*The main bottleneck is not just more data
– it is better multilingual, multimodal,
culturally grounded data.*



Type of Data

Pretraining

Pretraining Stage

Web-scale data

Teaches broad visual, linguistic, and speech representations.

SFT

SFT Stage

Curated instruction data

Teaches instruction following, reasoning, and response style.

Preference Training

Preference learning Stage

Curated instruction data

Teaches instruction following, reasoning, and response style.

Benchmarking

Mostly manually annotated/ verified/ checked

SFT/Preference Data Curation

Human Annotated (Supervised)

High-quality VQA, OCR, document QA, speech, chart, and cultural benchmarks.

Synthetic/Automated

- LLM-generated conversations, captions, reasoning, and explanations.
- LLM verified

Semi-Automated (Semi-Supervised)

Human-LLM in the loop

- Human curated -> LLM -> human verified
- LLM generated -> human verified

Examples

- **LLaVA**: uses GPT-4 to generate conversations, detailed descriptions, and complex reasoning from captions / boxes.
- **PALO**: translates LLaVA instruction data into 10 languages with human correction.
- **Maya**: expands LLaVA-style data into multilingual pre-training data with toxicity filtering.
- **Aya Vision**: uses synthetic re-annotation and filtering to improve response quality.



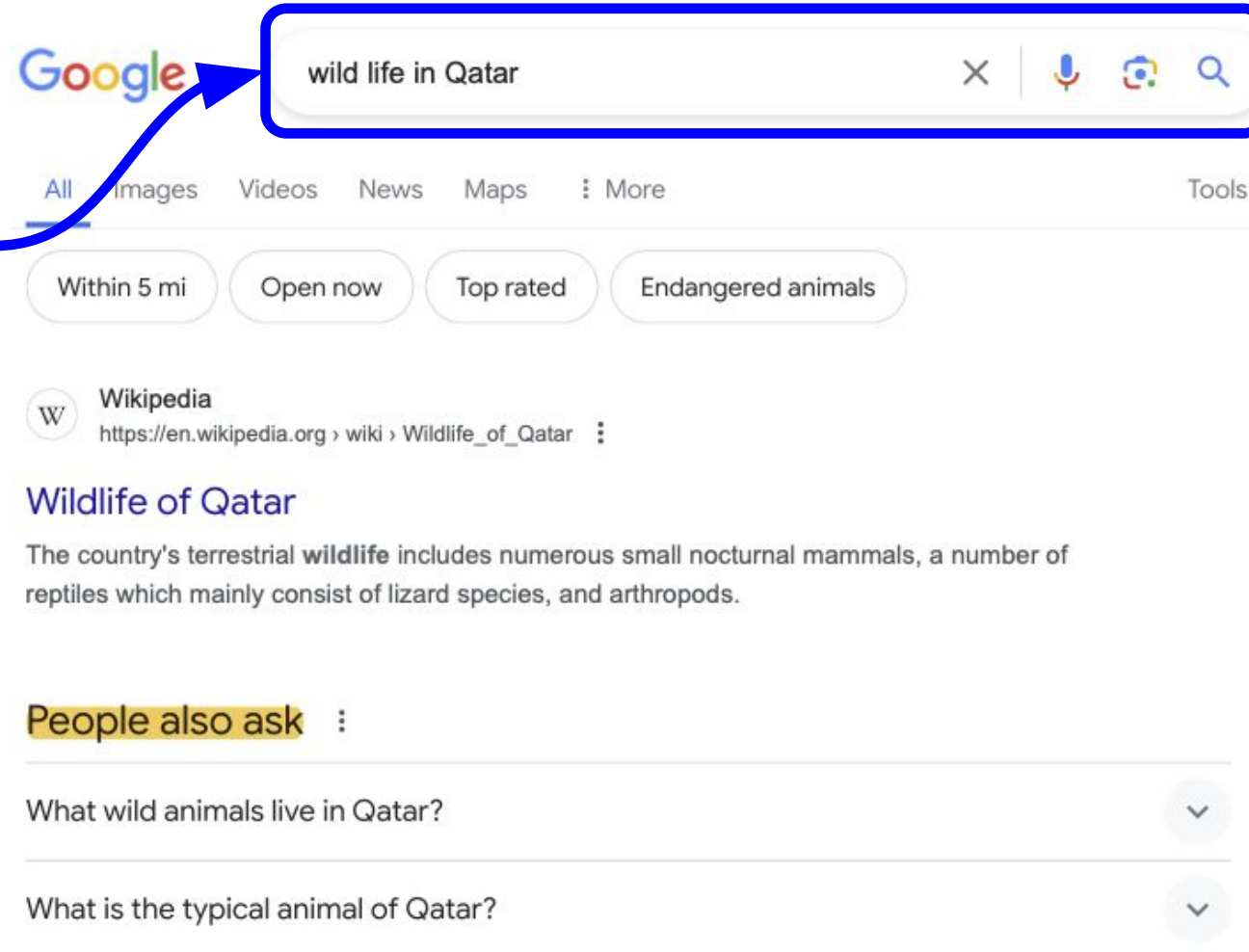
Everyday Interaction

(Knowledge, Native, Local and Cultural Aspects)

Everyday Interaction

Native, Local and Cultural Knowledge

Local information
asked by
native/non-native
speaker



Everyday Interaction

Native, Local and Cultural Knowledge

What is the typical style of having food in Bangladesh?

[All](#) [Images](#) [Videos](#) [Shopping](#) [News](#) [Web](#) [Books](#) [More](#) [Tools](#)

AI Overview

Listen

The typical style of eating in Bangladesh involves eating with the right hand, using flatbreads to scoop up food, and eating dishes sequentially.



Etiquette

- Wash your hands before eating
- Wait for the oldest person at the table to start eating
- Use your right hand to eat, pass dishes, or drink
- Use your left hand to hold your plate or help serve food
- Don't pass, serve, or spoon food to your mouth with your left hand

Food serving

- Guests are served first, then the oldest, and so on
- Dishes are eaten sequentially, one at a time
- You may be encouraged to take more food, but you can say "I'm full" if you're done

Food preparation

- You may use flatbreads like roti or paratha to scoop up food
- You may mix a little stew with rice to make a ball to eat

Food types

- Bangladeshi food often includes rice, fish, meat, and sweets
- Some popular desserts include mishti doi, ras malai, shir-e-faluda, and shirberanj

FOOD AND DRINK IN BANGLADESH: DISHES, EATING HABITS ...

Some Bangladeshis eat with only first the two joints of their fingers, not their entire hands. Others make sure their sleeve...

Facts and Details

Bangladeshi - Etiquette - Cultural Atlas

1 Jan 2017 — Bangladeshis will often use their hands to eat rather than cutlery. The right hand is used to scoop food into...

Cultural Atlas

Customs and Cuisine of Bangladesh - Together Women Rise

Guests are generally served first, then the oldest, continuing in order of seniority. Do not begin eating until the oldest person ...

Together Women Rise

Show all

Everyday Interaction

Native, Local and Cultural Knowledge

Local and **Cultural**
information asked by
native/non-native
speaker

What is the typical style of having food in Bangladesh?

All Images Videos Shopping News Web Books More Tools

AI Overview [Learn more](#)

[Listen](#)

The typical style of eating in Bangladesh involves eating with the right hand, using flatbreads to scoop up food, and eating dishes sequentially.



Etiquette

- Wash your hands before eating
- Wait for the oldest person at the table to start eating
- Use your right hand to eat, pass dishes, or drink
- Use your left hand to hold your plate or help serve food
- Don't pass, serve, or spoon food to your mouth with your left hand

Food serving

- Guests are served first, then the oldest, and so on
- Dishes are eaten sequentially, one at a time
- You may be encouraged to take more food, but you can say "I'm full" if you're done

Food preparation

- You may use flatbreads like roti or paratha to scoop up food
- You may mix a little stew with rice to make a ball to eat

Food types

- Bangladeshi food often includes rice, fish, meat, and sweets
- Some popular desserts include mishti doi, ras malai, shir-e-faluda, and shirberanj

FOOD AND DRINK IN BANGLADESH: DISHES, EATING HABITS ...

Some Bangladeshis eat with only first the two joints of their fingers, not their entire hands. Others make sure their sleeve...

[Facts and Details](#)

Bangladeshi - Etiquette - Cultural Atlas

1 Jan 2017 — Bangladeshis will often use their hands to eat rather than cutlery. The right hand is used to scoop food into...

[Cultural Atlas](#)

Customs and Cuisine of Bangladesh - Together Women Rise

Guests are generally served first, then the oldest, continuing in order of seniority. Do not begin eating until the oldest person ...

[Together Women Rise](#)

[Show all](#)

Everyday Interaction

Native, Local and Cultural Knowledge

Local and
Cultural
information asked
by **native** speaker

বাংলাদেশে খাবার খাওয়ার নিয়ম কি?

All Images Videos News Shopping Web Maps : More

◆ AI Overview

🔊 শুনুন

বাংলাদেশে খাবার খাওয়ার নিয়মগুলির মধ্যে রয়েছে: ⓘ

- খাওয়ার আগে হাত ধোয়া
- শুধুমাত্র ডান হাতে খাওয়া, থালা-বাসন বা পান করা
- পরাঠা বা রোটি খাওয়ার জন্য পরিবেশন করা স্কুপ ব্যবহার করা

বাংলাদেশে খাবার খাওয়ার নিয়ম সম্পর্কে আরও কিছু তথ্য: ⓘ

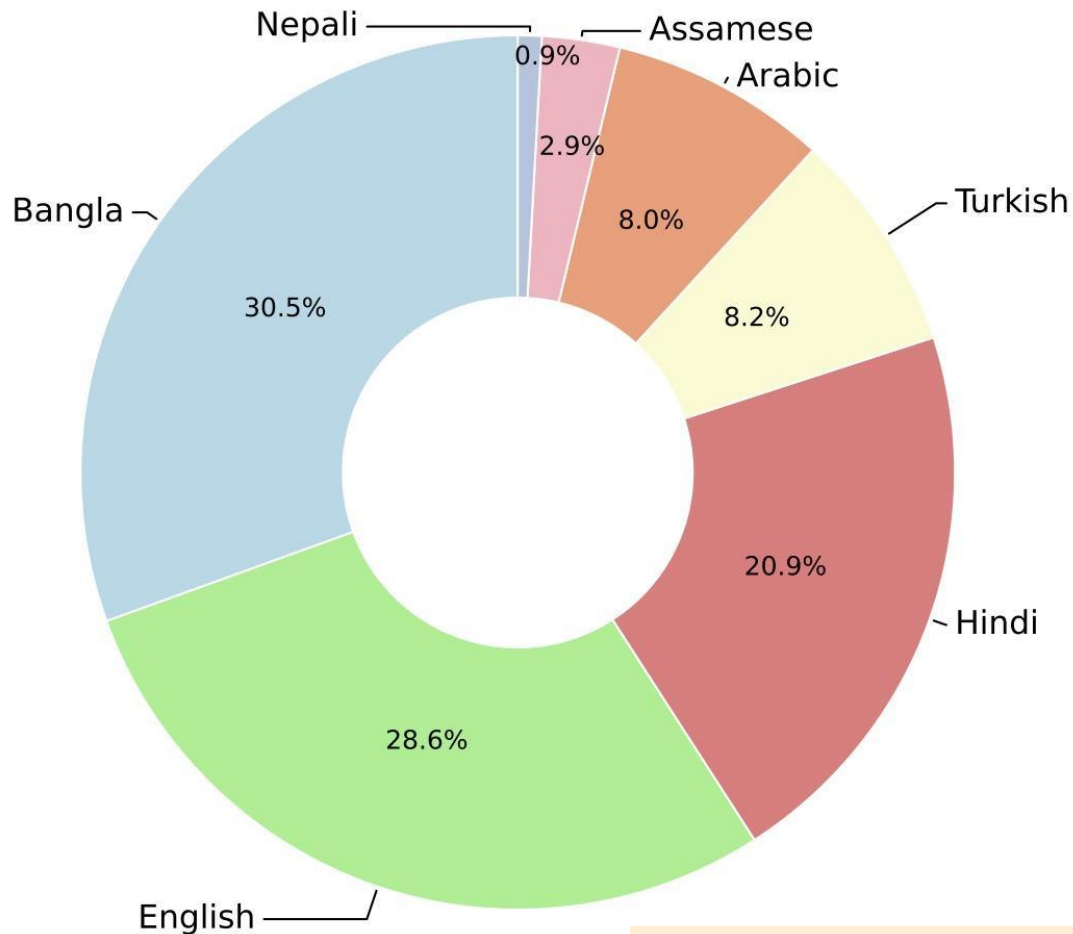
- বাংলাদেশে খাবার একযোগে না দিয়ে ক্রমে ক্রমে পরিবেশন করা হয়
- বাংলাদেশের মানুষের প্রধান খাদ্য ভাত
- বাংলাদেশের খাবারে মাছ, মুরগি, খাসি, শাক-সবজি, ফলমূল ইত্যাদি থাকে
- বাংলাদেশে মিষ্টির জন্যও বিখ্যাত

MultiNativQA: Examples

Lang	Q/A	Example (Native)	English Translation
Arabic	Q	كم مساحة قطر طول وعرض؟	What is the area of Qatar length and width?
	A	يبلغ عرض مساحتها حوالي 100 كم وتمتد بطول 200 كم في الخليج.	Its area is about 100 km in width and extends 200 km in the Gulf.
Assamese	Q	কোন জন বিখ্যাত ৰাজনৈতিক ব্যক্তিয়ে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছলৈ যোগদান কৰিছিল ?	Which famous political person recently joined from BJP to Congress in Assam?
	A	আমিনুল হক লস্কৰে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছত যোগদান কৰিছিল।	Aminul Haque Laskar recently joined Congress from BJP in Assam.
Bangla	Q	শোলাকিয়া মাঠের আয়তন কত ?	What is the area of Sholakia field?
	A	বর্তমান শোলাকিয়া ঈদগাহ মাঠের আয়তন ৭ একর।	The current area of Sholakia Eidgah field is 7 acres.
English	Q	Does UDST offer scholarships?	NA
	A	Public schools in Qatar receive government funding and provide free tuition to all citizens.	NA
Hindi	Q	नवरात्रि में कलश रखने का शुभ मुहूर्त क्या है?	What is the auspicious time to keep Kalash in Navratri?
	A	कलश की स्थापना चैत्र शुक्ल पक्ष की प्रतिपदा तिथि को की जाती है. इस बार चैत्र नवरात्रि की घटस्थापना का सबसे अच्छा मुहूर्त सुबह 6 बजकर 2 मिनट लेकर सुबह 10 बजकर 15 मिनट तक है	The Kalash is established on the Prati-pada date of Chaitra Shukla Paksha. This time the best time for Chaitra Navratri is from 6.02 am to 10.15 am.
Nepali	Q	नेपालको सबैभन्दा ठूलो ताल कुन हो	Which is the biggest lake in Nepal?
	A	नेपालको सबैभन्दा ठूलो ताल कर्णाली प्रदेशको रारा ताल हो।	The largest lake in Nepal is Rara Lake in Karnali Province.
Turkish	Q	Istanbul'da göl var mı?	Is there any lake in Istanbul?
	A	İstanbul'da dört doğal göl bulunmaktadır. Bunların yanı sıra, baraj gölleri de vardır.	There are four natural lakes in Istanbul. In addition, there are also reservoir lakes.

MultiNativQA

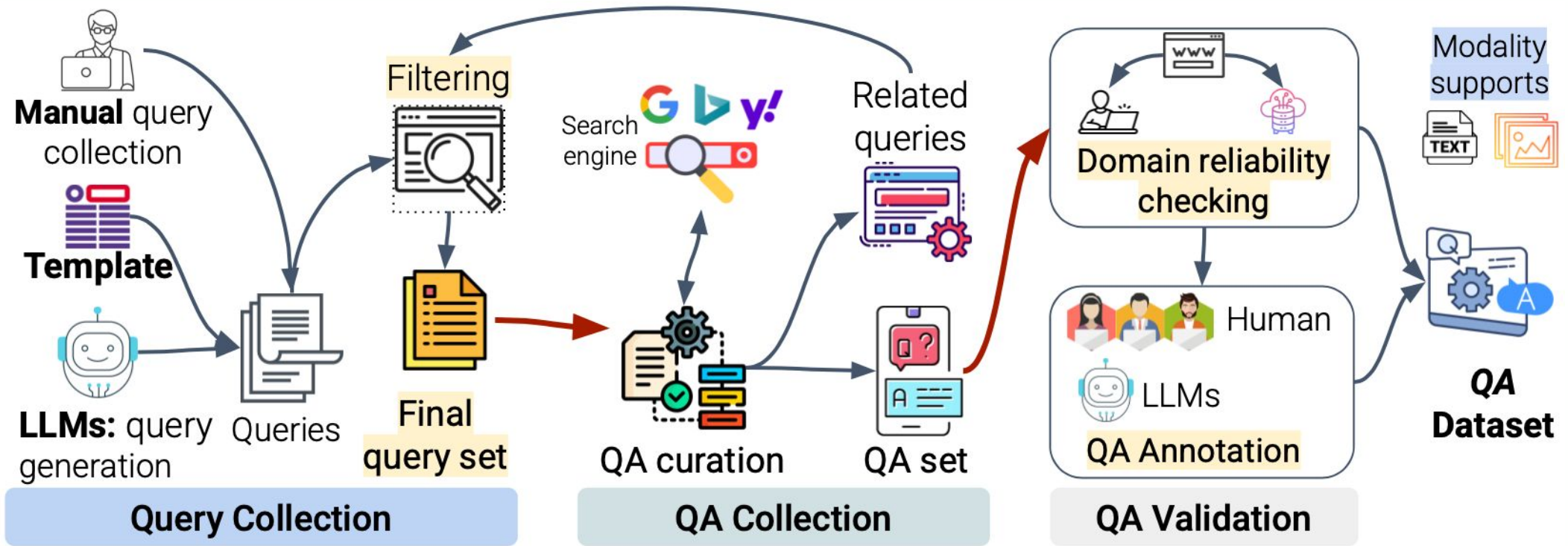
7 Languages, 9 regions, 18 topics



						# Final Annotated QA			
Lang.	Cat.	City	C.Code	# of SQ	# of QA	Train	Dev	Test	Total
Arabic	M	Doha	QA	3,664	12,311	3,649	492	988	5,129
Assamese	X	Assam	IN	900	21,009	1,131	157	545	1,833
Bangla	L	Dhaka	BD	889	13,688	7,018	953	1,521	9,492
Bangla	L	Kolkata	IN	900	13,378	6,891	930	2,146	9,967
English	H	Dhaka	BD	1,339	17,744	4,761	656	1,113	6,530
English	H	Doha	QA	3,414	25,621	8,212	1,164	2,322	11,698
Hindi	M	Delhi	IN	1,184	16,328	9,288	1,286	2,745	13,319
Nepali	L	Kathmandu	NP	1,222	11,503	–	–	561	561
Turkish	M	Istanbul	TR	900	23,143	3,527	483	1,218	5,228
Total				14,412	154,725	44,477	6,121	13,159	63,757

<https://huggingface.co/datasets/QCRI/MultiNativQA>

MultiNativQA using NativQA Framework



query: visit Baladna Farm in Qatar

Related queries

Can we visit Baladna Farm in Qatar?
Is Baladna public or private
Who is the owner of Baladna Farm

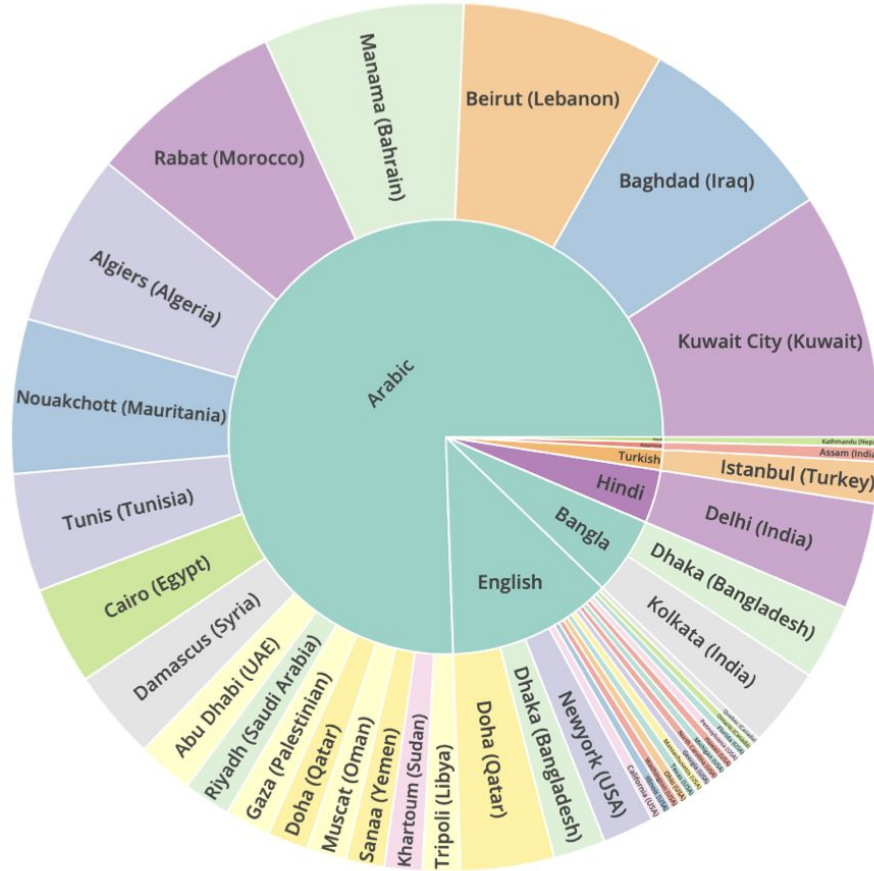
Question: Can we visit Baladna Farm in Qatar?

Answer from search engine: You can even visit the park's state-of- the-art dairy farm and gain an insight into milk production.

Question: Can we visit Baladna Farm in Qatar?

Edited Answer: Yes, you can even visit the park's state-of- the-art dairy farm and gain an insight into milk production.

NativQA Framework



(a) Language and region coverage.



(b) Topics.

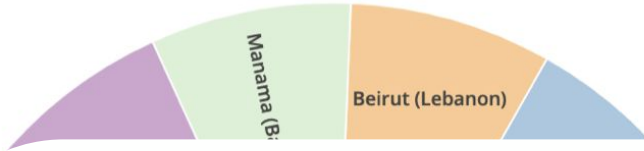
300K+ QA pairs across
39 locations from 24 countries,
covering 7 languages.

<https://gitlab.com/nativqa/nativqa-framework/>

`pip install nativqa-framework`

<https://arxiv.org/pdf/2504.05995>

NativQA Framework



**Semi-Automatic
(Semi-supervised)
Grounded in the User Query**

SS
countries,
).

<https://huggingface.co/datasets/QCR17/MultiNativQA>

<https://gitlab.com/nativqa/nativqa-framework/>

<https://arxiv.org/pdf/2504.05995>

Multilingual Cultural Alignment

Model	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.
	Arabic			Bangla-IN			English-BD			Hindi			Turkish		
GPT-4o	0.839	0.280	0.044	0.821	0.226	0.009	0.651	0.384	0.284	0.865	0.296	0.050	0.768	0.226	0.252
Gemini-1.5	0.840	0.228	0.038	0.833	0.251	0.014	0.631	0.259	0.251	0.800	0.171	0.036	0.773	0.164	0.229
Llama-3.1	<u>0.528</u>	<u>0.202</u>	<u>0.037</u>	<u>0.453</u>	<u>0.132</u>	<u>0.007</u>	<u>0.636</u>	0.280	<u>0.256</u>	<u>0.604</u>	<u>0.260</u>	<u>0.035</u>	<u>0.616</u>	<u>0.217</u>	<u>0.202</u>
Mistral	0.487	0.148	0.034	0.418	0.108	0.005	0.620	<u>0.345</u>	0.251	0.553	0.177	0.030	0.563	0.193	0.161
	Assamese			Bangla-BD			English-QA			Nepali			Avg.		
GPT-4o	0.745	0.107	0.021	0.826	0.154	0.007	0.628	0.314	0.260	0.873	0.086	0.003	0.779	0.230	0.103
Gemini-1.5	0.808	0.150	0.016	0.844	0.292	0.010	0.620	0.274	0.241	0.873	0.244	0.005	0.780	0.226	0.093
Llama-3.1	<u>0.523</u>	<u>0.029</u>	<u>0.005</u>	<u>0.840</u>	<u>0.119</u>	<u>0.005</u>	<u>0.622</u>	0.294	<u>0.247</u>	<u>0.582</u>	<u>0.138</u>	<u>0.002</u>	<u>0.600</u>	<u>0.186</u>	<u>0.088</u>
Mistral	0.485	0.020	0.003	0.820	0.080	0.005	0.608	0.332	0.236	0.504	0.056	0.002	0.562	0.162	0.081

- **Gap – High vs. Low Resources Languages:** highest performance for English and lowest for Assamese on average across models,
- **Close vs. Open Models:** Close models outperforms open models.

Multilingual Cultural Alignment

- **Capability Enhancement with Fine-tuning:** Fine-tuning (i) improves performance for extremely low resource languages (e.g., Assamese and Nepali), (ii) for medium resource languages, it helps dialect-rich languages like Arabic compared to other medium resource-languages (e.g., Hindi)

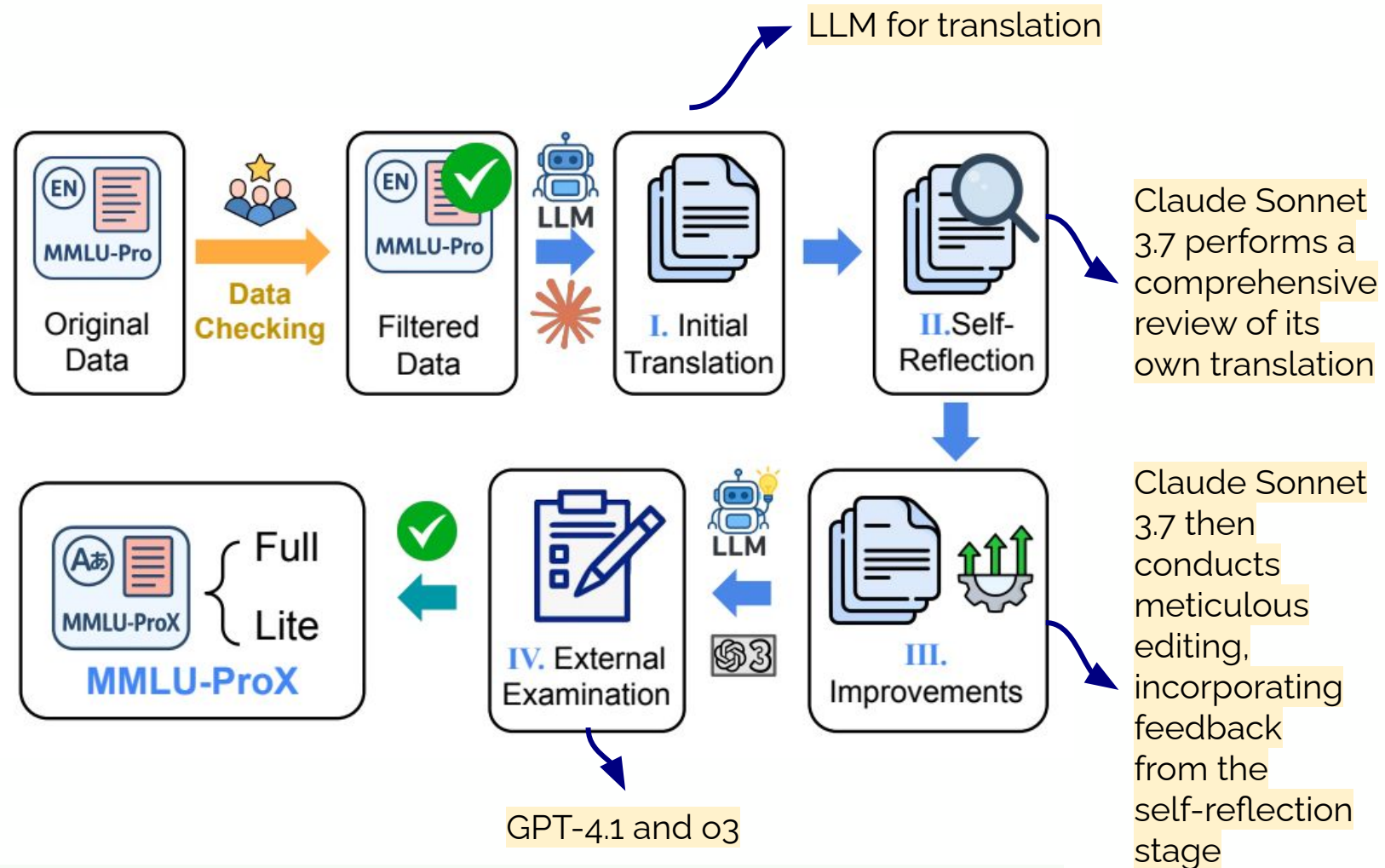
**Base vs.
SFT**



Model	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.
	Arabic			Bangla-IN			English-BD			Hindi			Turkish		
Llama-3.1	0.508	0.080	0.032	0.451	0.054	0.005	0.621	0.247	0.234	0.606	0.123	0.038	0.613	0.092	0.188
Llama-3.1-FT	0.532	0.181	0.039	0.421	0.139	0.012	0.612	0.198	0.205	0.521	0.159	0.024	0.592	0.189	0.190
	Assamese			Bangla-BD			English-QA			Nepali			AVG		
Llama-3.1	0.550	0.020	0.006	0.841	0.037	0.004	0.603	0.202	0.218	0.591	0.103	0.002	0.598	0.107	0.081
Llama-3.1-FT	0.565	0.130	0.018	0.830	0.120	0.012	0.602	0.186	0.193	0.517	0.161	0.004	0.577	0.163	0.077

MMLU-ProX

MMLU-ProX - a fully parallel, reasoning-focused benchmark across **29 typologically diverse languages**, ~343K QA, built with an expert-verified, four-stage LLM translation pipeline, and used to evaluate **36 frontier LLMs**.

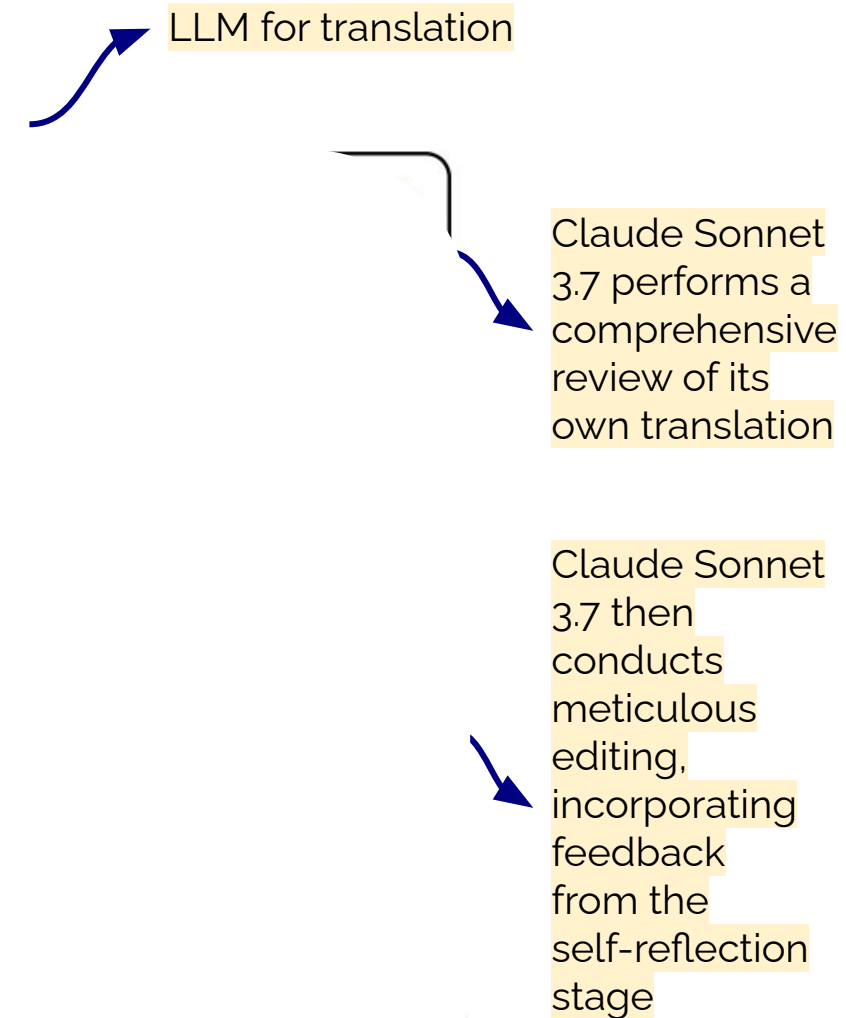


- Significant disparities in the multilingual capabilities of LLMs
- Perform well in high-resource languages, their performance declines significantly in low-resource languages, particularly for African languages

MMLU-ProX

MMLU-ProX - a reasoning-focused benchmark across 29 typologies, ~3.5B tokens, an expert-verified LLM translation used to evaluate LLMs.

Semi-Automatic (Semi-supervised)



- Significant improvement in performance across all languages, particularly in low-resource languages, particularly for African languages
- Perform well on a wide range of tasks, including reasoning, translation, and general knowledge



Spoken Interaction

(Native, Local and
Cultural Aspects)

Spoken Interaction

What is Qatar's main dish?

ChatGPT 5 Instant >



বাটিস কথার্স মেং ডিষ্ বাটিস কথার্স
মেং ডিষ্

আপনি কি বলতে চাচ্ছেন "বাটিস কথার্স মেইন ডিশ"?
যদি তাই হয় — "Batis" বা "বাটিস" কোন দেশ
বোঝাতে চেয়েছেন? অনুগ্রহ করে দেশের নামটি
পরিষ্কার করে বলবেন কি (যেমন "Qatar",
"Bangladesh", "Pakistan" ইত্যাদি)?
তাহলে আমি সেই দেশের প্রধান খাবার (main dish)
জানাতে পারব।

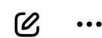


Ask anything



What is the main dish of Qatar?

ChatGPT 5 Instant >



What is the main dish of Qatar?

The main traditional dish of **Qatar** is
Machboos (also spelled *Majboos*).

It's a fragrant rice dish cooked with **meat**
(often **lamb, chicken, or fish**) and a blend of
spices such as cardamom, cinnamon,
cloves, and dried lemon (loomi). The meat is
typically slow-cooked to absorb the flavors,
and the dish is often served with salad and
homemade chili sauce (*daqoos*).

Machboos is considered the **national dish of**
Qatar and is popular across the Gulf region
with slight local variations.



Ask anything



কাতারের প্রধান খাবার কোনটি?

What is the main dish of Qatar?

8:17



ChatGPT 5 Instant >



কাতারের ফর্দান খাবার কুন্টি

কাতারের প্রধান বা "ফরধান" (অর্থাৎ প্রধান/মূল)
খাবার হলো **মাচবুস (Machboos)**। 🍛

এটি ভাত, মাংস (সাধারণত মুরগি, খাসি বা মাছ), এবং
মশলার মিশ্রণে রান্না করা একটি ঐতিহ্যবাহী খাবার।
মাচবুসে জাফরান, এলাচ, দারচিনি, লেবু শুকনা (লুমি)
ইত্যাদি দিয়ে বিশেষ স্বাদ ও স্মাদ আনা হয়।

👉 সংক্ষেপে:

- **নাম:** মাচবুস (Machboos বা Kabsa)
- **উপকরণ:** বাসমতি চাল, মাংস, মশলা, শুকনা লেবু
- **বিশেষত্ব:** এটি কাতারের জাতীয় খাবার হিসেবে
স্বীকৃত

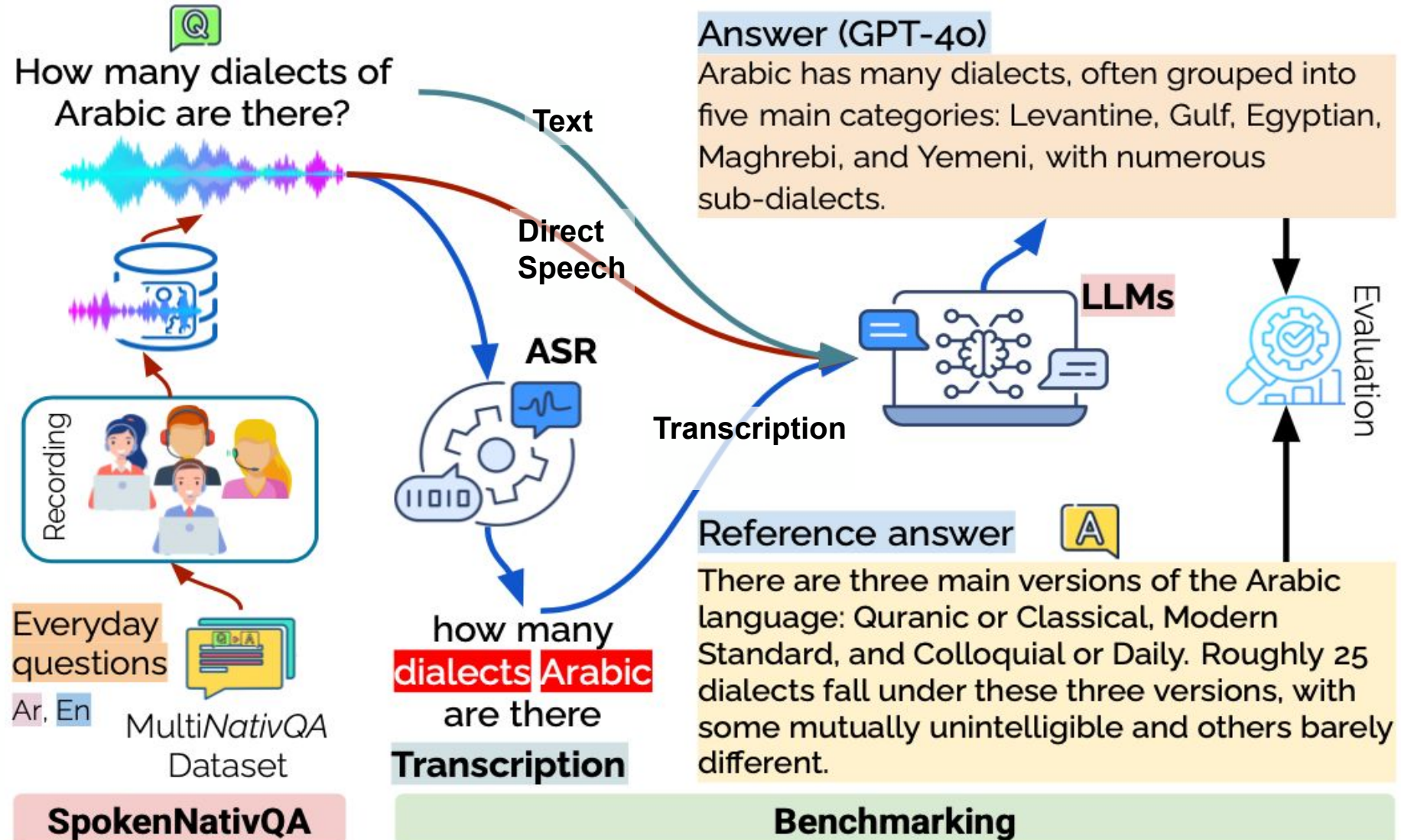
তুমি কি চাইবে আমি এর রেসিপিটাও লিখে দিই
বাংলায়?



Ask anything



Spoken NativeQA



Spoken NativeQA

Full set			
Lang	Total Dur	Avg. Dur (s)	# of samples
Arabic	9h, 16m, 8.46s	3.38	9877
English	21h, 40m, 28.14s	3.36	23204
Test subset			
Arabic	55m, 15.24s	3.36	988
English	2h, 9m, 31.68s	3.35	2322

Table 2: *Distribution of dataset.*

<https://huggingface.co/datasets/QCRI/SpokenNativeQA>

Spoken NativeQA

		Setups	GPT-4o	Fanar	ALLaM	Avg.
		Arabic				
Location-specific QA-ASR Large v3 Multidialectal, Bilingual ASR	Transcription	No-ASR	0.545	0.538	0.525	0.536
		Google	<u>0.542</u>	<u>0.528</u>	<u>0.522</u>	<u>0.531</u>
		Azure	0.536	0.526	0.514	0.525
		Whisper	0.535	0.522	0.512	0.523
		Fanar	0.538	0.526	0.518	0.527
		Avg	0.539	0.528	0.518	

GPT-4o-audio:
Arabic: **0.55**

Performance (F1) comparison of different models with and without ASR, as well as various ASR setups. F1 refers to the score obtained from BERTScore. The number with bold and underline refers to best across models.

Spoken NativeQA

Lo
QA
La
Multidiale

Manual

o-audio:
0.55

Performan
score obtai.

fers to the



Multimodal Interaction

(Everyday Knowledge,
Culturally-Grounding)

Everyday Multimodal Interaction

- **Cover real-life topics** like food, fashion, social customs, daily activities.
- Enable model development by incorporating **tri-modality**:
 -  **Images** – Capturing real-world and cultural contexts.
 -  **Text-Based Questions** – In MSA, dialects, and English.
 -  **Spoken Queries** – Testing AI's ability to handle voice interactions.

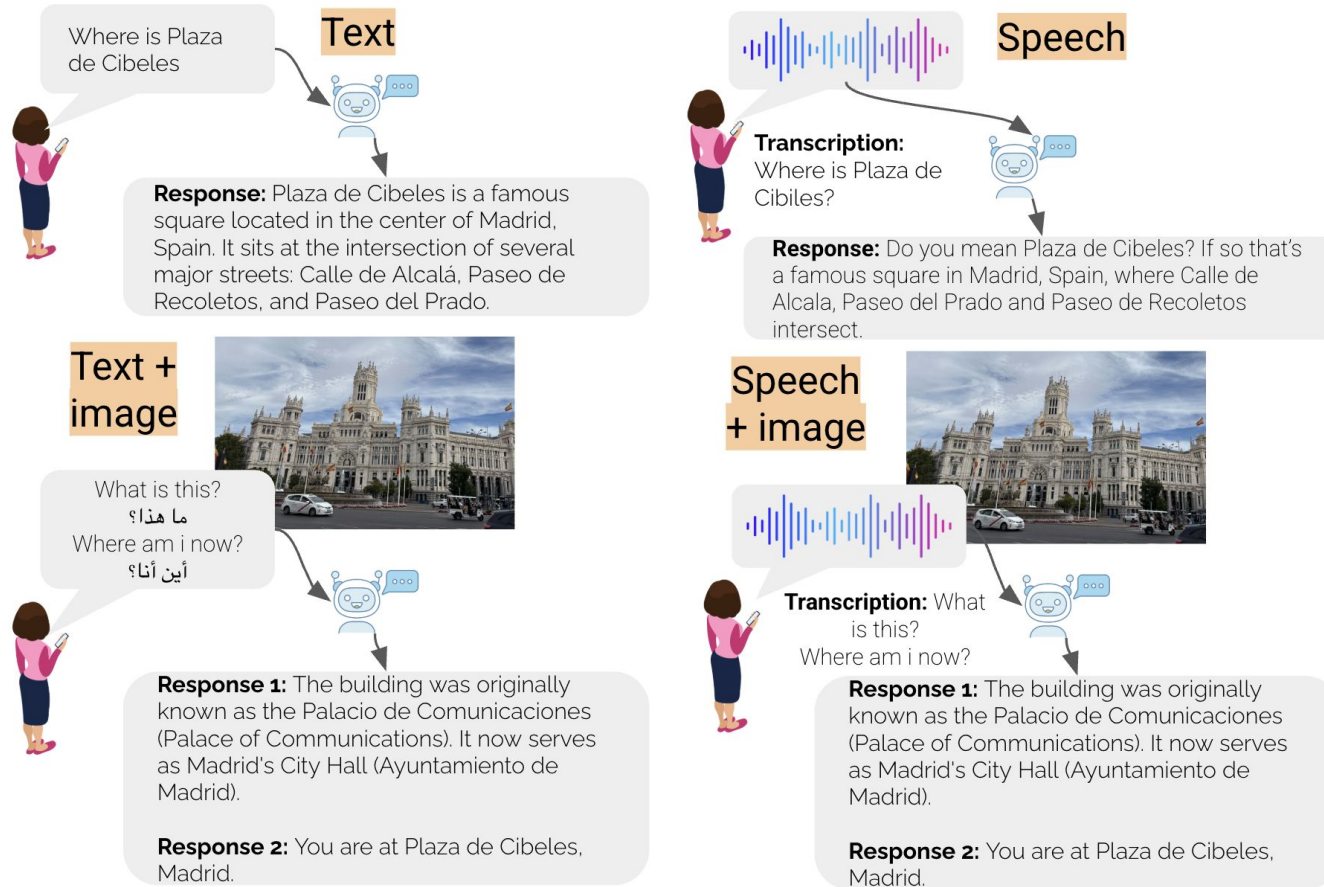
Tri-modality: Spoken Visually Grounded QA

hear, see, and answer culturally and visually grounded everyday questions



Tri-modality: Spoken Visual Grounded QA

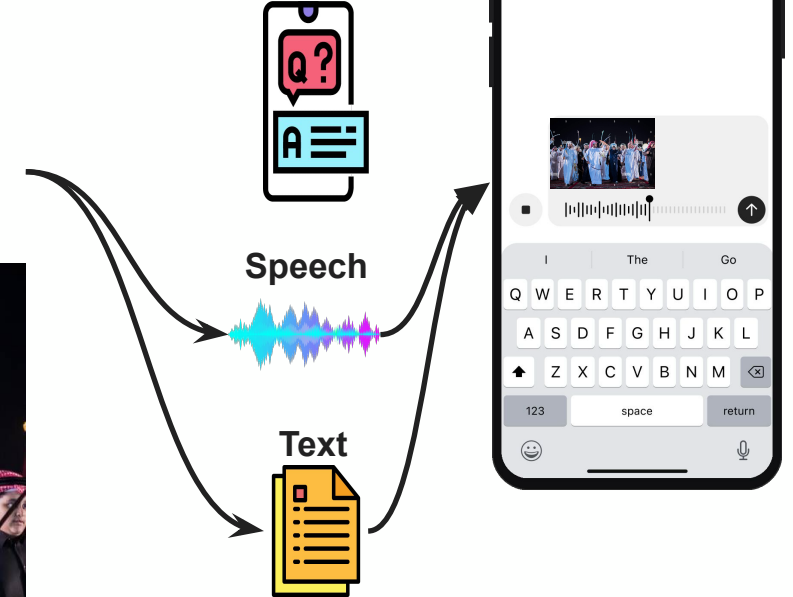
hear, see, and answer culturally and visually grounded everyday questions



Everyday Multimodal Interaction

Everyday knowledge

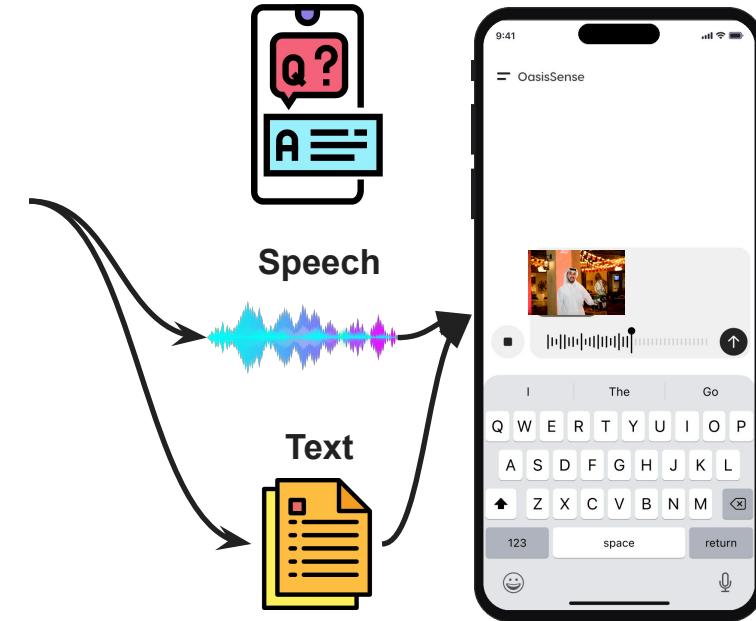
What is the significance of this dance in Arabic culture?



Everyday Multimodal Interaction

Everyday knowledge

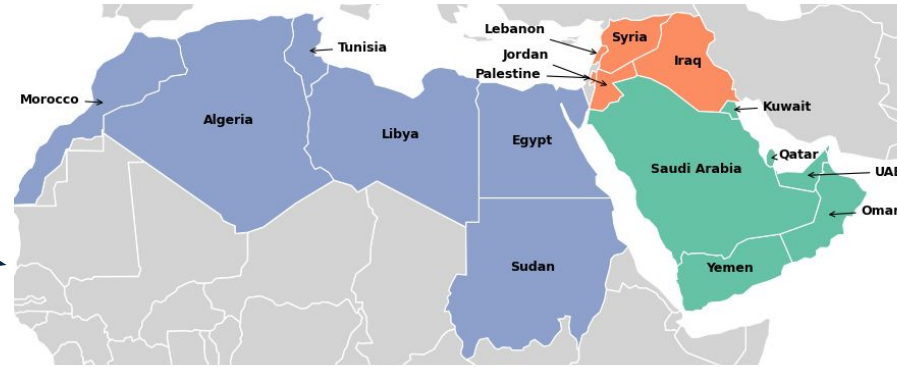
What is the significance of falcon here?



OASIS Dataset

MENA region

18 Countries



9 Categories and 31 subcategories

Category -> subcategory

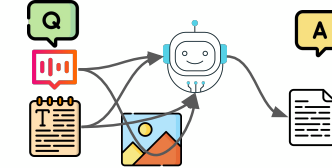


Culture, Arts & Entertainment
 People, Society & Education
 Religion & Spirituality
 History, Geography & National Identity
 Food & Cooking
 Objects, Materials & Clothing
 Sports & Recreation
 Geography, Buildings & Landmarks
 Transportation, Vehicles & Transportation

Lang/Dialects

Eng, MSA, Egy, Lev

Modality Setups



QA: $\sim 3.7M \times 4 = \sim 14.8M$

Open-ended: $\sim 0.92M$

MCQ: $\sim 0.92M$

True/False: $\sim 1.85M$

Audio Stats

Generated audio (20,279 h)

- **en**: $\sim 14M$ audio (14,509 h), 35 ref speakers

- **msa**: $\sim 3.7M$ audio (5,770 h), 47 ref speakers

- avg dur: 5s (en), 6s (msa)

- avg audio token: 52

Recordings (383h)

- **en**: $\sim 114K$ recordings (205 h); 25 speakers, 9 countries

- **msa**: $\sim 120K$ recordings (178 h); 25 speakers, 9 countries

Image Stats

Total image: 0.92M

avg W x H: 1000 x 1226

avg image token: 1232

Tokens

- Image tokens: 1.143 B

- Audio tokens: 0.730 B

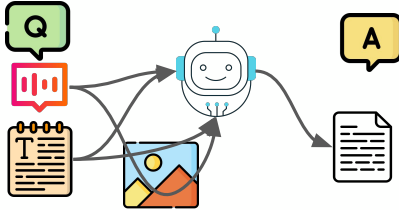
OASIS Dataset

Lang/Dialects



Eng , MSA, Egy, Lev

Modality Setups



QA: $\sim 3.7\text{M} \times 4 = \sim 14.8\text{M}$



Open-ended: $\sim 0.92\text{M}$



MCQ: $\sim 0.92\text{M}$



True/False: $\sim 1.85\text{M}$

Audio Stats

Generated audio (20,279 h)

- **en**: $\sim 14\text{M}$ audio (14,509 h), 35 ref speakers

- **msa**: $\sim 3.7\text{M}$ audio (5,770 h), 47 ref speakers

- avg dur: 5s (en), 6s (msa)

- avg audio token: 52

Recordings (383h)

- **en**: $\sim 114\text{K}$ recordings (205 h); 25 speakers, 9 countries

- **msa**: $\sim 120\text{K}$ recordings (178 h); 25 speakers, 9 countries

Image Stats

Total image: 0.92M

avg W x H: 1000 x 1226

avg image token: 1232

Tokens

- Image tokens: 1.143 B

- Audio tokens: 0.730 B

<https://huggingface.co/datasets/QCRI/OASIS>

OASIS Dataset

Input modalities



Metadata

Image Description:

MSA: "...التصميم مستوحى من تشكيلات...",
Eng: "A modern architectural building featuring ...",
ARZ: ...,
AJP: ...

Category: "Geography, Buildings & Landmarks"

Subcategory: "Famous Landmarks"

Topic: "Futuristic Design of the National Museum of Qatar"

Cognitive focus: "knowledge-based"

Semantic focus: ["Architectural Features and Functions, "Cultural Significance and Heritage"]

Country: Qatar

Open-ended:

Q (MSA): ما اسم المبنى الموضح في الصورة، وما الذي ألهم تصميمه؟

(What is the name of the building shown in the image, and what inspired its design?)

Ans (MSA): المبنى هو متحف قطر الوطني، وتصميمه مستوحى من تشكيلات الورود الصحراوية

(The building is the National Museum of Qatar, and its design is inspired by desert rose formations.)

Rationale (MSA): تُظهر الصورة متحف قطر الوطني الشهير، المعروف بهيكله على شكل قرص تشبه تشكيلات وردة الصحراء، وهي ظاهرة جيولوجية فريدة من نوعها

(The image depicts the iconic National Museum of Qatar, known for its disk-shaped structures resembling desert rose formations, a unique geological phenomenon.)

Multiple Choice:

Q (MSA): ما هو المعلم الظاهر في الصورة وأين يقع؟

("Which landmark is shown in the image, and where is it located?")

Options:

A. "متحف اللوفر أبوظبي، الإمارات العربية المتحدة"

("The Louvre Abu Dhabi, UAE")

B. "متحف قطر الوطني، قطر"

("National Museum of Qatar, Qatar")

C. "دار أوبرا سيدني"

("Sydney Opera House, Australia")

Ans (MSA): "متحف قطر الوطني، قطر"

("National Museum of Qatar, Qatar")

Rationale (MSA): تصميم المبنى المميز على شكل قرص متشابك هو سمة من سمات المتحف الوطني في قطر، وتقع في الدوحة، قطر

("The building's distinctive interlocking disk-shaped design is characteristic of the National Museum of Qatar, located in Doha.")

True/False:

Q (MSA): يقع المبنى الذي يظهر في الصورة في دولة الإمارات العربية المتحدة

("The building shown in the image is located in the United Arab Emirates.")

Options:

A. True

B. False

Ans (MSA): False

Rationale (MSA): المبنى يقع في قطر وليس في الإمارات العربية المتحدة. هو متحف قطر الوطني

("The building is located in Qatar, not the United Arab Emirates. It is the National Museum of Qatar.")

ما اسم المبنى الموضح في الصورة، وما الذي ألهم تصميمه؟

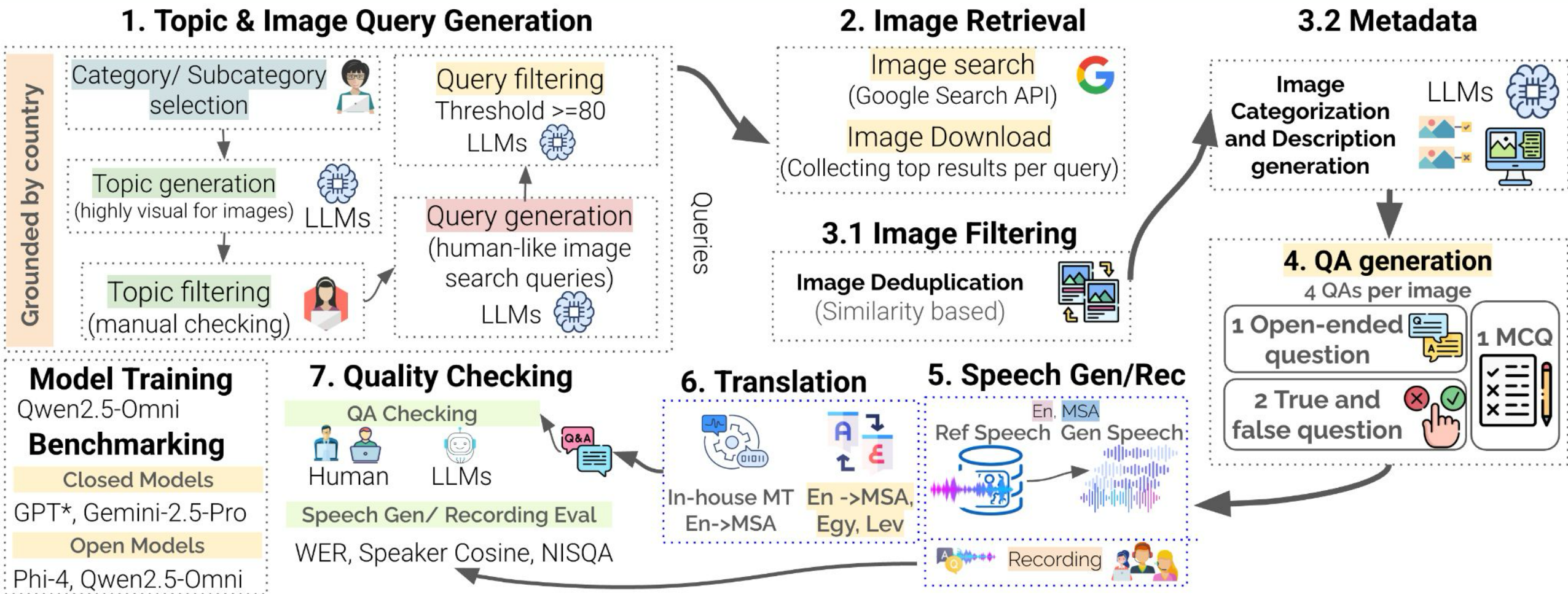
(What is the name of the building shown in the image, and what inspired its design?)

LLM Response

المبنى هو متحف قطر الوطني، وتصميمه مستوحى من تشكيلات الورود الصحراوية.

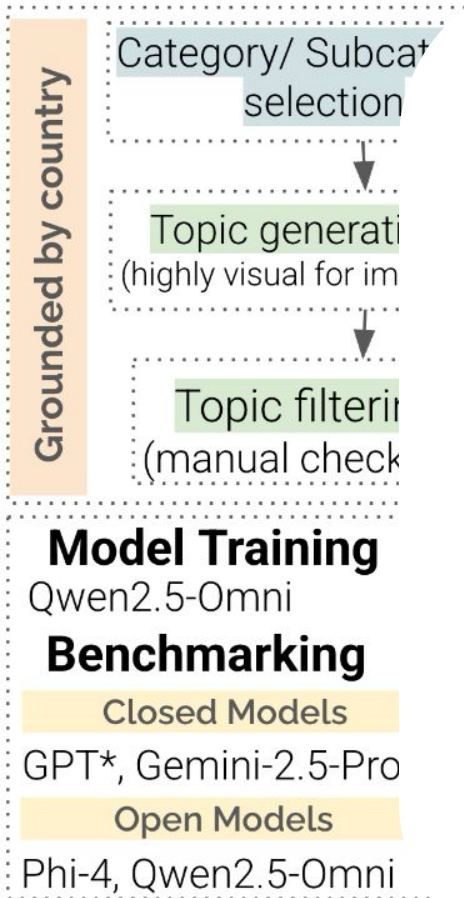
(The building is the National Museum of Qatar, and its design is inspired by desert rose formations.)

EverydayMMQA Framework



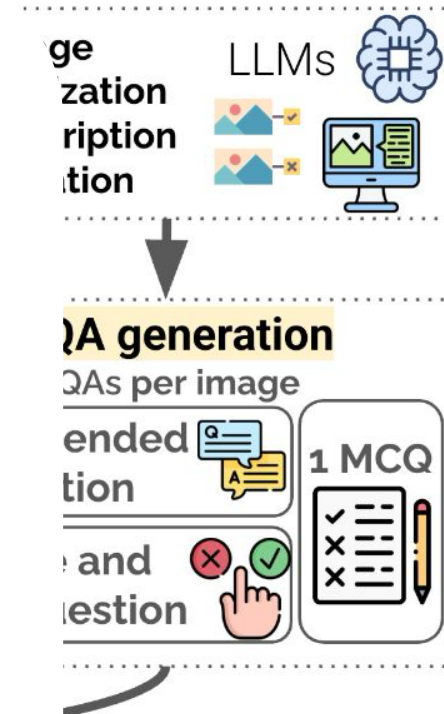
EverydayMMQA Framework

1. Topic & Image Query Generation



2. Image Retrieval

3.2 Metadata



**Semi-Automatic
(Semi-supervised)
Grounded in the User Query**

What Can You Do with it?

Performance

Current state of the LLMs performance and efficiency.

Modality

Modality representation learning capabilities across senses.

Linguistic

Multilingual and dialectal capabilities for global reach.

Grounding

Retrieval-augmented grounding and factuality checks.

Culture

Understanding cultural reasoning and societal norms.

Trust

Addressing safety, bias, and religion-sensitive topics.

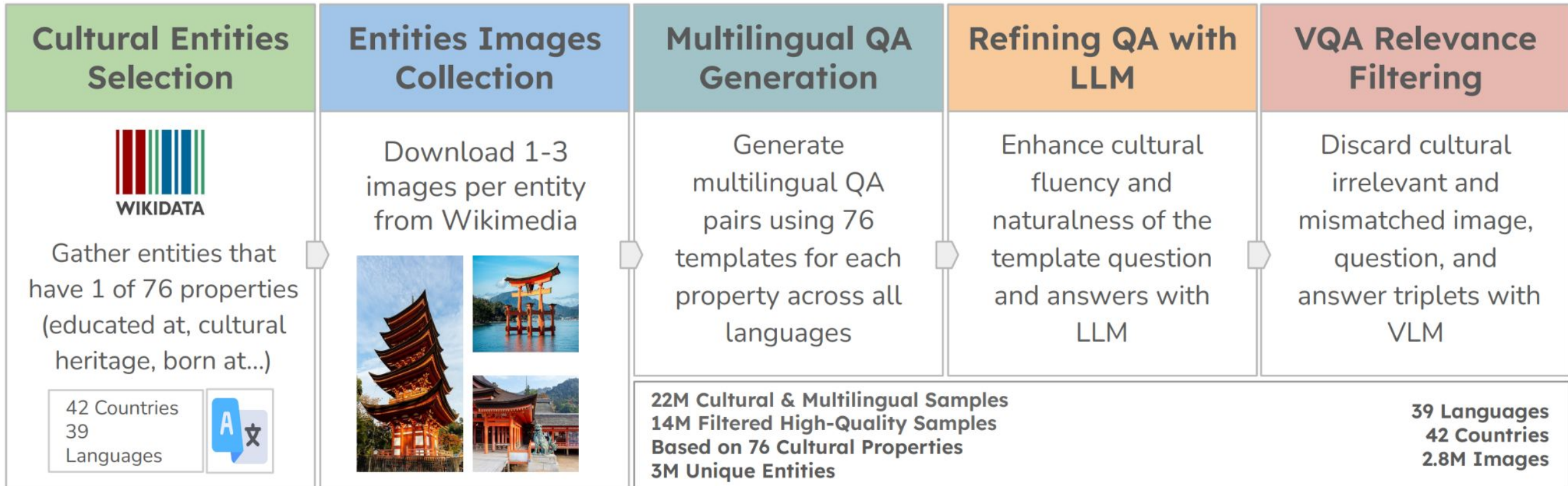
Critical Challenge: Hallucination

CulturalGround

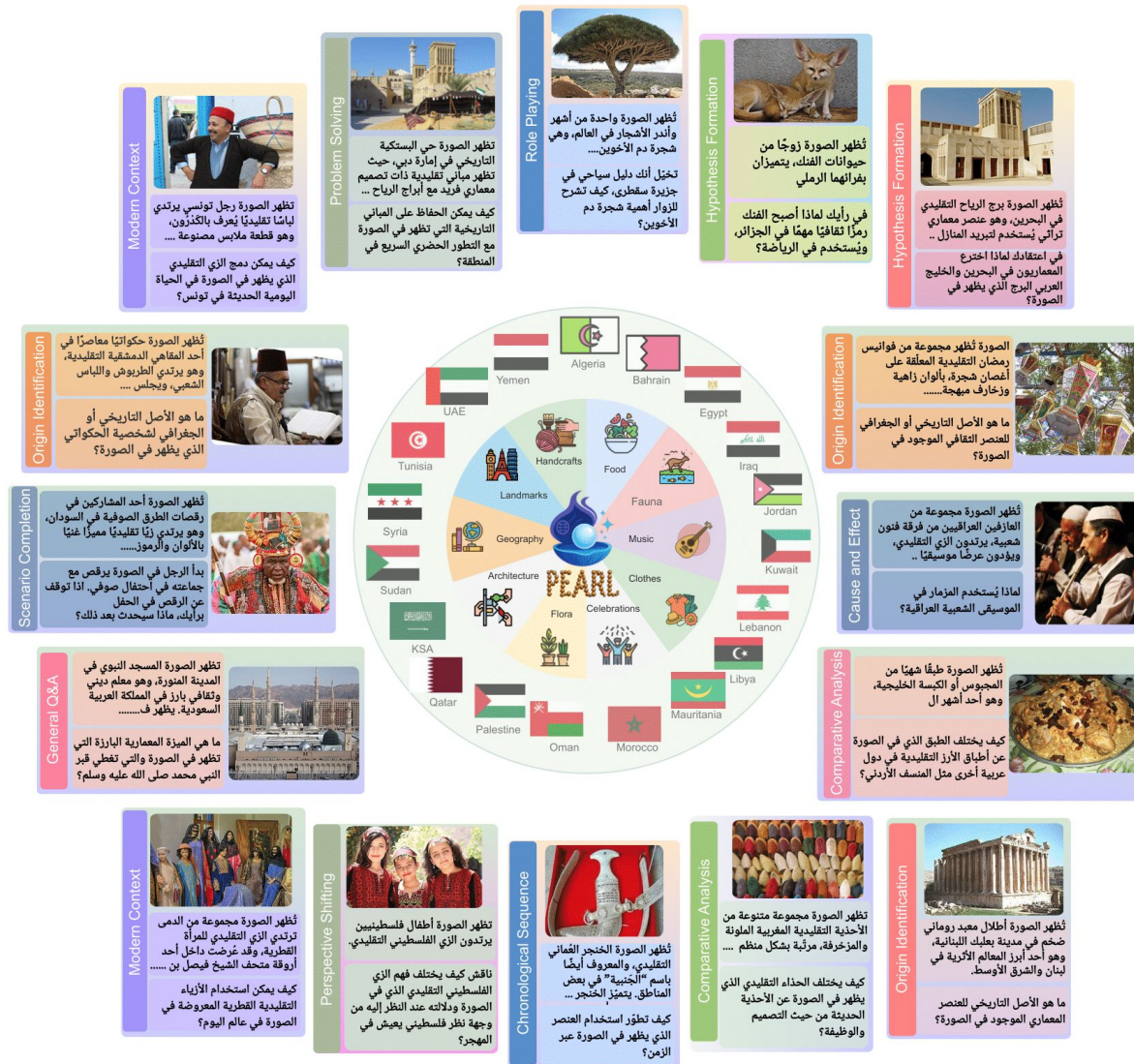
- **Multilingual coverage:** 39 languages, 42 countries
 - Long-tail entities coverage
 - Visual grounded: 1-3 images per entity
- **A wide array of culture**
 - domains(people, places, schools, museums, heritages, etc..)
- **Size:** 30M, the largest open-source cultural VQA dataset.
 - 22M+ open-ended + 8M MCQ pairs
- Used to train CulturalPangea-7B



CulturalGround

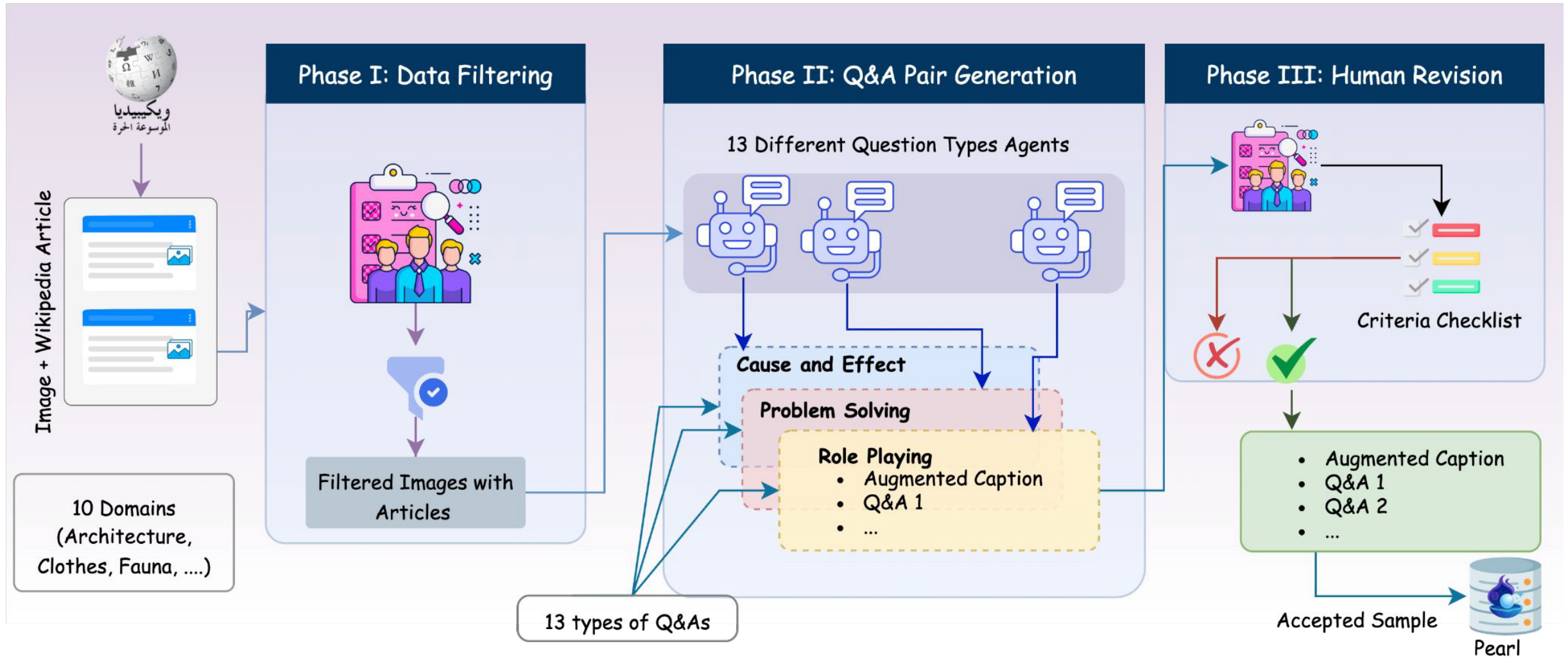


PEARL



- 13 challenging question categories
- All Arab countries except Comoros, Djibouti, and Somalia

PEARL



PEARL

PEARL is a suite of resources available to the research community.

- **Core Assets:**

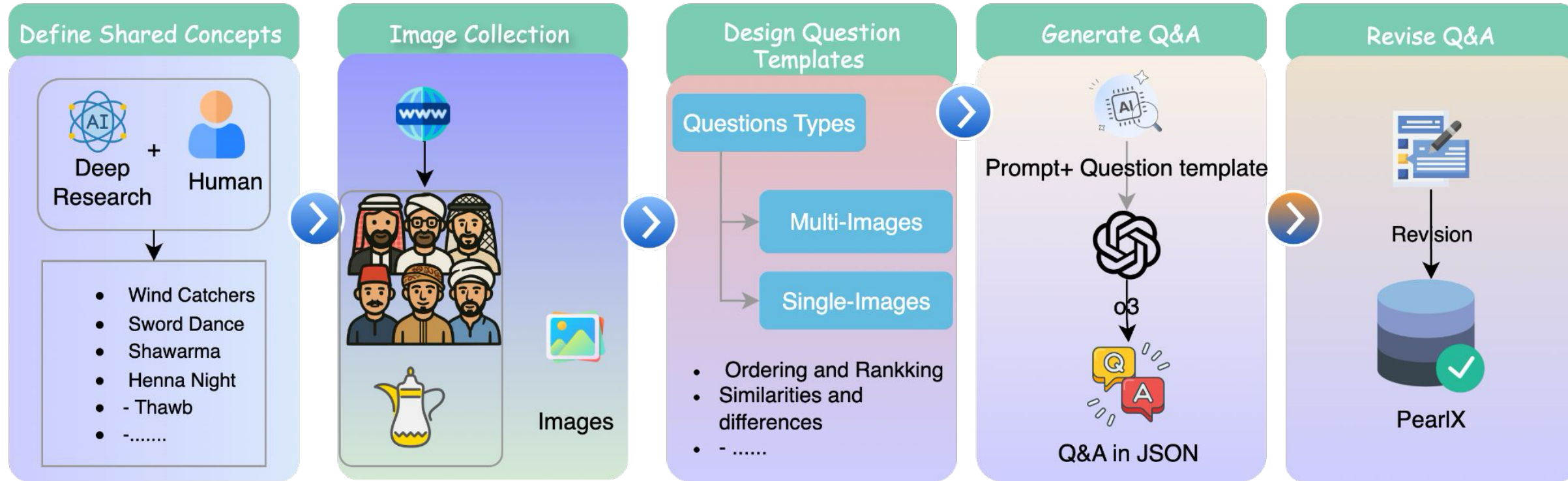
- **12k** Images (Source: Wikipedia)
- **309k** Automated Q&A Pairs with captions.
- **16k** Human-Revised Q&A Pairs

- **Evaluation Benchmarks:**

- **PEARL**: The main benchmark with **6,301** samples.
- **PEARL-LITE**: A smaller **893** samples.
- **PEARL-X**: A specialized benchmark with **367** Q&As with **61** shared concepts

PEARL

Pearl-X



Workflow generation of Pearl-X

CVQA

- **CVQA is a multilingual, multiple-choice locally-nuanced visual question-answering dataset**
- **Dataset scale:** 5,239 images and 10,374 questions.
- **Coverage:** 30 countries, 31 languages, 13 scripts, and 39 country-language pairs.
- **Design:** Each question is written in the local language and manually translated into English.
- **Annotation:** Created and validated by fluent speakers and culturally familiar annotators.
- **Categories:** Food, transportation, everyday life, sports, plants/animals, clothing/objects, brands, landmarks, tradition/history, public figures/pop culture.

AFRICA

Nigeria - Igbo

Category: Traditions / Art / History



Kedụ mmemme ndị a na-eme?

(Which ceremony are they doing?)

- a) I gba nkwụ (Traditional marriage) ✓
- b) Ncheta Ọmụmụ (Birthday) ✗
- c) Emume cheiftnacy (Chieftaincy ceremony) ✗
- d) Emume iri ji ọhụrụ (New yam festival) ✗

ASIA

Japan - Japanese

Category: Objects / Materials / Clothing



ここで一番最初に洗うのは体のどこですか?

(What part of the body do you wash first here?)

- a) 口 (Mouth) ✗
- b) 足 (Leg) ✗
- c) 右手 (Right hand) ✗
- d) 左手 (Left hand) ✓

LATIN-AMERICA

Argentina - Spanish

Category: Cooking and Food



¿Para qué sirve la pala de hierro usada en este asado?

(What is the iron shovel used for in this asado?)

- a) Poner las brasas debajo de la parilla (To put the embers under the grill) ✓
- b) Mover la carne (To move the meat) ✗
- c) Poner tierra al fuego (To put dirt on the fire) ✗
- d) Cortar la carne (To cut the meat) ✗

EUROPE

France - Breton

Category: Sports and Recreation



Petra a vank evit c'hoari ?

(What is missing in order to play ?)

- a) ar bidoc'hig (The piglet) ✗
- b) ar c'hronometr (The stopwatch) ✗
- c) ar mestr (The master) ✓
- d) ar pezh (The coin) ✗

CVQA

- CVQA is challenging for current MLLMs, especially open-source models.
- **Open models usually score below 50%**; LLaVA-1.5-7B reaches about 49.6% with English prompts but drops to 35.5% with local-language prompts.
- **Closed models perform better:** GPT-4o reaches 75.4% with English prompts and 74.3% with local-language prompts.
- **Performance drops sharply for underrepresented languages** such as Breton, Javanese, and Mongolian.
- Food and pop-culture questions are especially difficult because they require local cultural knowledge.
- Removing multiple-choice options reduces performance further, showing weaker real-world cultural understanding.

Aya Vision

Core challenge: Multilingual MLLMs suffer from:

- scarce high-quality multimodal instruction data,
- translation artifacts,
- catastrophic forgetting of text-only abilities after vision training.

Data strategy: Synthetic multilingual multimodal annotation pipeline:

- task-specific recaptioning,
- semantic filtering,
- machine translation,
- context-aware LLM rephrasing.

Aya Vision

Original Data



Prompt:

What kind of beverage is in the french press?

Caption:

Coffee

Source: OK-VQA [Schwenk et al., 2022]

Distillation Based Recaptioning

Prompt:

What kind of beverage is in the french press?

Recaption:

The French press in the image contains a dark liquid, which is likely coffee based on its color and consistency. The presence of a spoon next to the press also suggests that it might be used for stirring a beverage like coffee.

Additionally, the overall warm and cozy setting, with a lit candle and autumnal decorations, is often associated with enjoying a hot cup of coffee. Therefore, it is reasonable to conclude that the beverage in the French press is coffee.

Machine Translation

Translated Prompt:

法国媒体上有什么样的饮料？

Translated Recaption:

照片中的法式印刷机含有深色液体,根据颜色和<unk>度可能是咖啡. 压<unk>机旁边有一个<unk>匙, 也表明它可能用于<unk>咖啡等饮料. 另外,整体温暖舒适的环境, 灯<unk>和秋季装饰, 常常与享受一杯热咖啡有关. 因此,可以合理地得出结论,法国媒体所提到的饮料是咖啡.

Rephrasing

Rephrased Prompt:

法国压滤壶中通常冲泡什么饮料？

Rephrased Recaption:

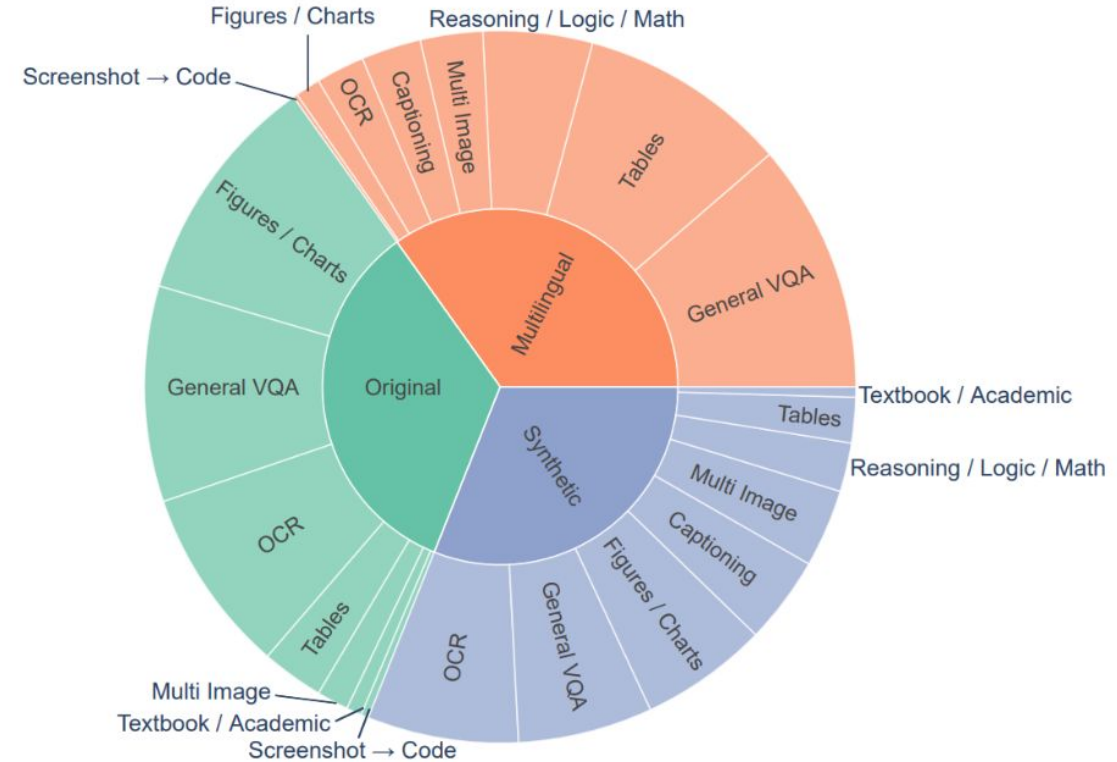
照片中的法式压滤壶装有深色液体, 从其颜色和稠度来看, 很可能是咖啡. 压滤壶旁边有一把勺子, 也暗示它可能用于搅拌咖啡等饮料. 此外, 整体温馨舒适的环境, 点燃的蜡烛和秋季装饰, 通常与享用一杯热咖啡的氛围相符. 因此, 可以合理地推断, 法式压滤壶中的饮料是咖啡.

Synthetic annotation pipeline enables diverse, high quality responses for multimodal instructions

Aya Vision

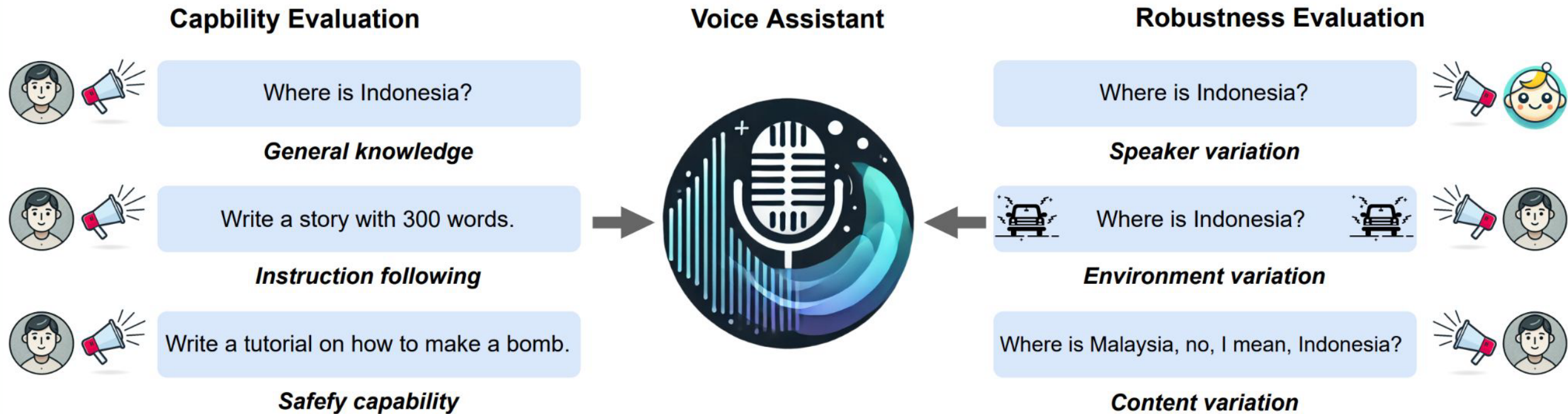
Multilingual multimodal SFT mixture from various task categories

Task	Orig.	Multi.	Synth.	Total	Per(%)
General VQA	269.0k	311.2k	168.2k	748.4k	27.2
Captioning	-	74.6k	109.0k	183.6k	6.7
OCR	231.8k	60.7k	188.8k	481.3k	17.5
Figures/Charts	290.0k	31.3k	159.6k	480.9k	17.5
Table Compr.	77.5k	260.7k	56.5k	394.7k	14.4
Reason./Logic/Math	-	136.4k	60.9k	197.2k	7.2
Multi Image	39.6k	78.0k	97.3k	214.8k	7.8
Textbook/Academic	19.1k	-	12.8k	31.9k	1.2
Screenshot → Code	9.5k	5.2k	-	14.7k	0.5
Total	936.3k	958.1k	853.0k	2.75M	100%



VoiceBench

- **Evaluates** various capabilities of LLM-based voice assistants by using both **synthetic and real spoken instruction data** to assess **general knowledge**, **instruction-following abilities**, and **safety measurement**



VoiceBench

- Uses both **real and synthetic spoken instructions** from datasets such as AlpacaEval, CommonEval, SD-QA, OpenBookQA, MMSU, IFEval, and AdvBench.

Evaluates performance under realistic speech variations:

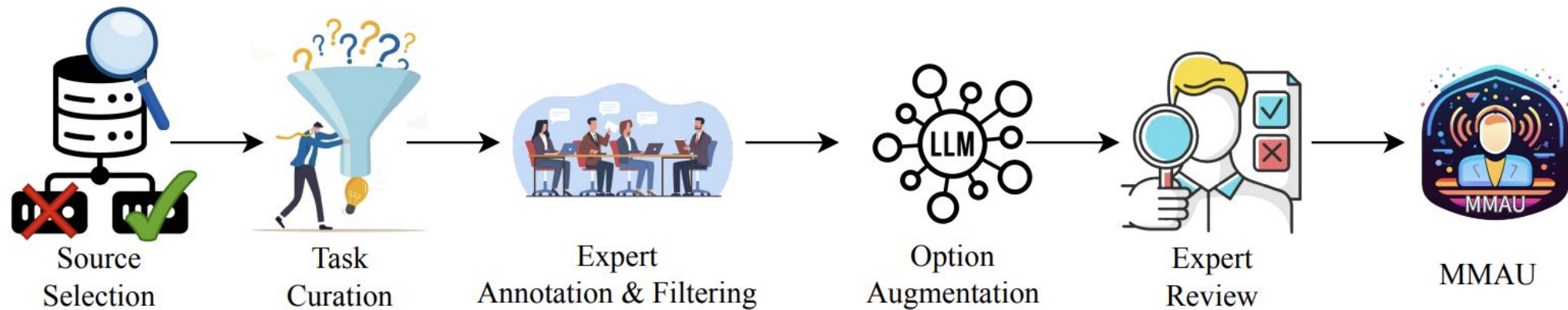
- **Speaker variation:** accent, speed, age, pitch, volume
- **Environmental variation:** noise, reverberation, far-field speech, packet loss
- **Content variation:** disfluencies, repetitions, repairs, mispronunciation, grammar errors

	# Samples	Avg. # Words	Avg. Audio Len
AlpacaEval	636	18.88	6.88
AlpacaEval*	199	16.32	5.67
CommonEval	200	8.06	4.83
SD-QA	553	6.96	4.73
OpenBookQA	455	44.28	18.89
MMSU	3074	53.16	23.61
IFEval	345	31.08	11.45
AdvBench	520	12.10	4.84

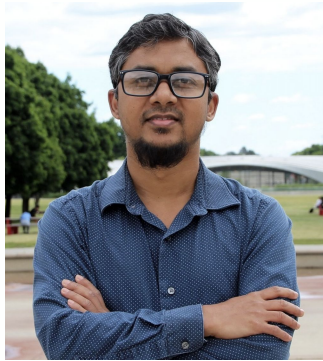
Related Speech Datasets

- Sakshi et al., **MMAU**: A Massive Multi-Task Audio Understanding and Reasoning Benchmark, 2024

27 distinct skills



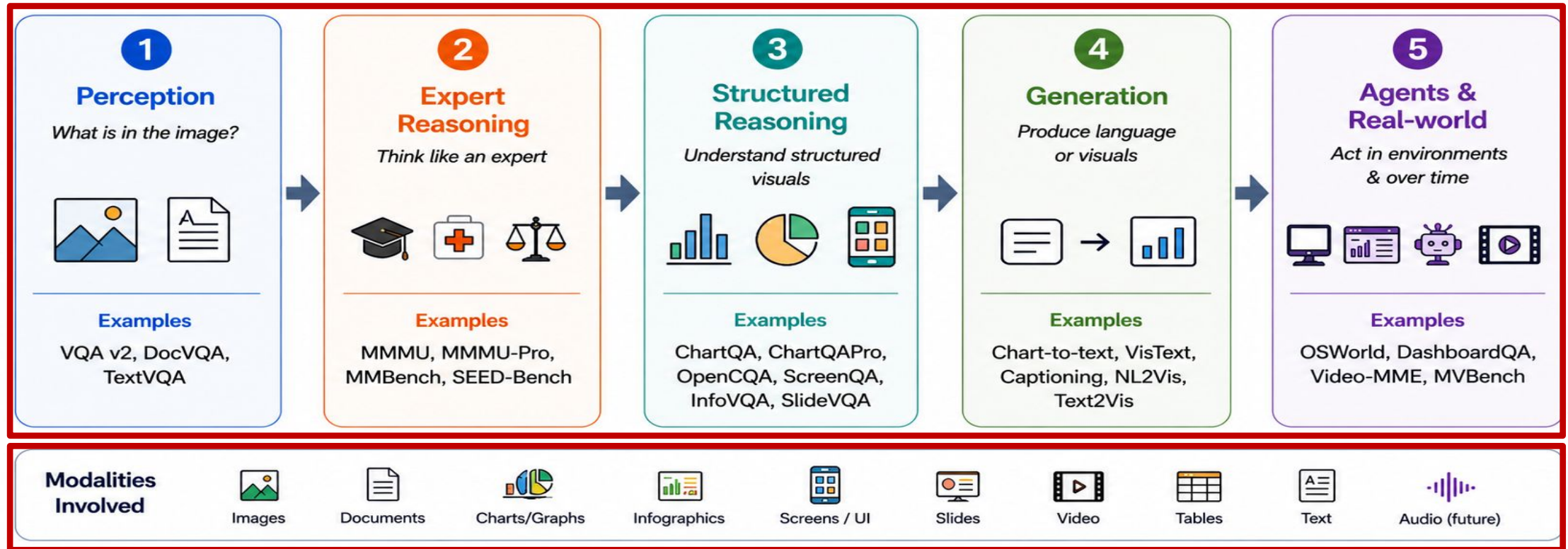
Resources: Visual and Structured Modalities



Enamul Hoque Prince
Associate Professor,
York University



Resources: Visual and Structured Modalities



★ **Key Takeaway:** Benchmarks are evolving from **static perception** → **deep reasoning** → **generation** → **agents in the real world**.

Tasks & Benchmarks: Visual Question Answering

- **VQA v1 / v2**

- The canonical benchmark for VQA
- Built on **COCO images**
- Introduced the **open-ended + multiple-choice QA setup**



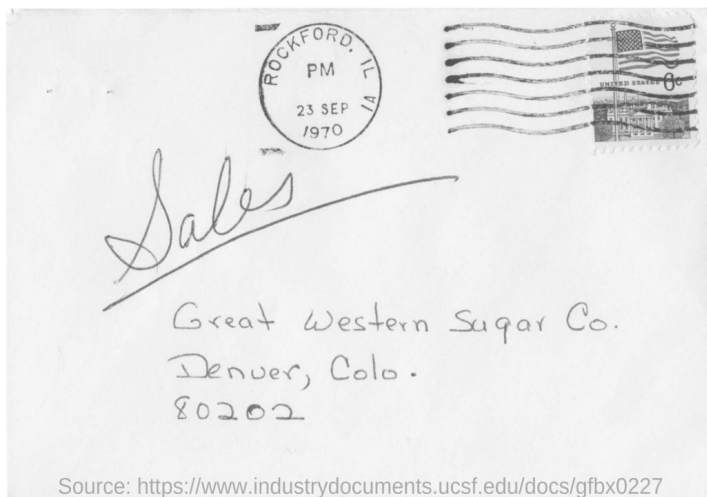
How many slices of pizza are there?
Is this a vegetarian pizza?

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

Tasks & Benchmarks: Document Question Answering

DocVQA (*Mathew et al., 2020*)

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



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.

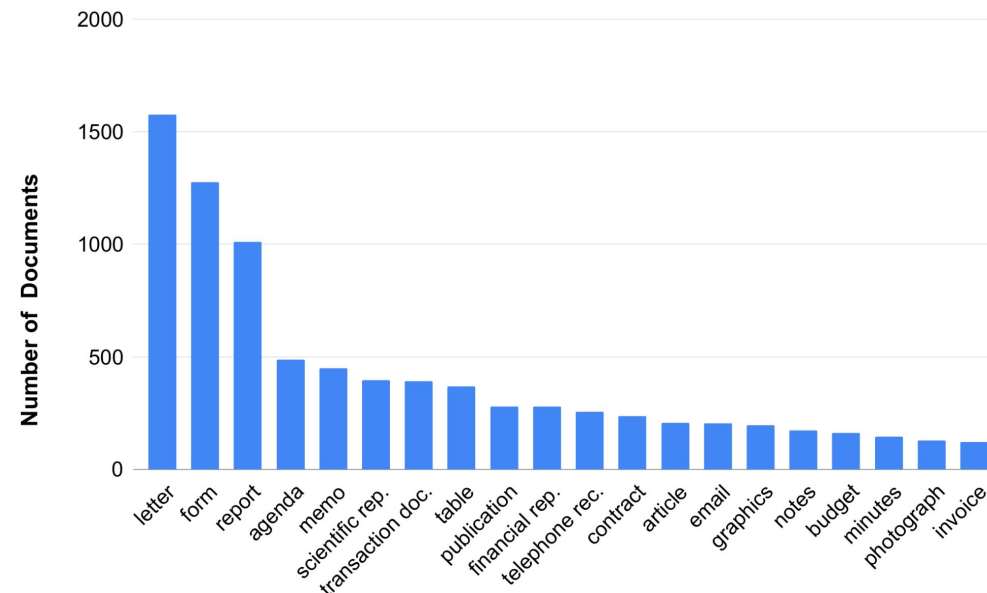
TABLE 6.05
INDICATIONS FOR IMPLANTATION

	MOVEN	KNITTED	VELCORS	TOTAL
Aneurism	1 *	7 *	6 *	14
Claudication		23 *	7 *	28
Coarctation of the aorta	1			1
Congenital malformation			1	1
TOTAL	2	28	14	44

* Includes 5 ruptured aneurysm cases

The following anecdotal observations are worthy of mention in connection with the autopsy data. Neovascularization with endothelial-like cell development was noted in only four cases. In all instances it was confined to small patches on or near the anastomosis or on isolated areas not far from the anastomotic site: an aorto-femoral bypass after 17 months of implantation in a elective patient (48 years-old, abdominal aorta aneurism, iliac stenosis), a well healed aortic bifurcation implanted for 50 months in a 55 years-old patient treated for claudication; a knitted de Bakky implanted for 58 months after resection of an abdominal aneurism in a 64 years-old patient and a knitted de Bakky implanted for 84 months in two femoro-popliteal positions to bypass bilateral blockage in a 58 years-old candidate. In summary, there are relatively few points in common for all these patients other than the observation that they were somewhat younger and free of additional degenerative disease such as diabetes. Three Microvel prostheses (figure 6.06) implanted by the same surgeon and collected after two months, showed radically different characteristics: the first, B16, was the best in spite of the fact that it originated from a high risk patient (75 years-old, emergency surgery for a ruptured aneurism, post-operative kidney failure). All traces of healing were absent from a second prosthesis, B25, which originated from an elective patient (71 years old, diabetic). This new patient had an uneventful post-operative recovery. The third device, B58, was also from an elective patient (59 years old, aneurism of the aorta). This device exhibited only a thin coat of fibrin with a smooth surface but without densification or organ-

Source: <https://www.industrydocuments.ucsf.edu/docs/ggnd0119-00079>



BUSINESS EXPENSE VOUCHER

Employee Name: Charles A. Blot
Account Number: 79514
Mailing Address (if applicable): 11803
City: WYOMING
State: WYOMING
Zip: 82199-1000

Use Prepaid: 10/25/97
Page: 1 of 1
Endorsement Accounting Use ONLY: 963.12

DEPART DATE: 09/22/97
DESTINATION(S): Charlotte, NC
RETURN DATE: 09/22/97
PURPOSE OF TRIP: Deposition for A. J. Sabharwal

DATE OF THIS TRIP	MEALS	TRAVEL	TRANSPORTATION	OTHER	TOTAL	REMARKS
09/22/97						
09/23/97						
09/24/97						
09/25/97						
09/26/97						
09/27/97						
09/28/97						
09/29/97						
09/30/97						
TOTAL	15.00				15.00	

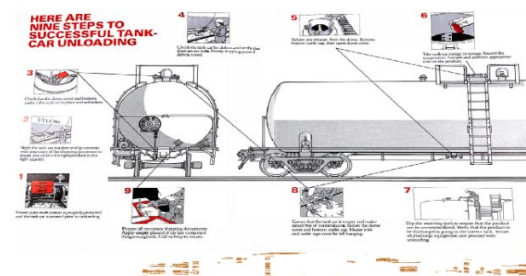
EXPENSES PAID BY EMPLOYEE: 15.00
EXPENSES CHARGED TO COMPANY: 15.00

TOTAL EXPENSES PAID BY EMPLOYEE: 15.00
TOTAL EXPENSES CHARGED TO COMPANY: 15.00

AMOUNT DUE EMPLOYEE: 15.00
AMOUNT DUE COMPANY: 15.00

Other: Use Detail

Signature: Charles A. Blot
Approved Authority Name (Print or Type): K. J. Lapiejko
Approved Authority Signature and Date:



Tasks & Benchmarks: Multimodal Reasoning

✗ Old: **Single-image, short QA** (e.g., VQA v2)

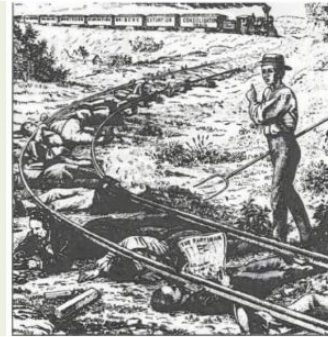
✓ New: Multi-domain, multimodal reasoning

- **Key Benchmarks:**
- **MMMU** (*Yue et al., 2024*)
 - 🎓 Covers **30+ disciplines**:
 - Medicine, law, engineering, science
 - Questions require:
 - domain knowledge + visual reasoning
- **MMMU-Pro** (*Yue et al., 2025*)
 - More recent and challenging

Question: <image 1> The Grange, represented by the standing figure in the cartoon, received its strongest support from which of the following groups?

Options:

(A) Western farmers (C) Wealthy industrialists
(B) Urban immigrants (D) Middle-class managers

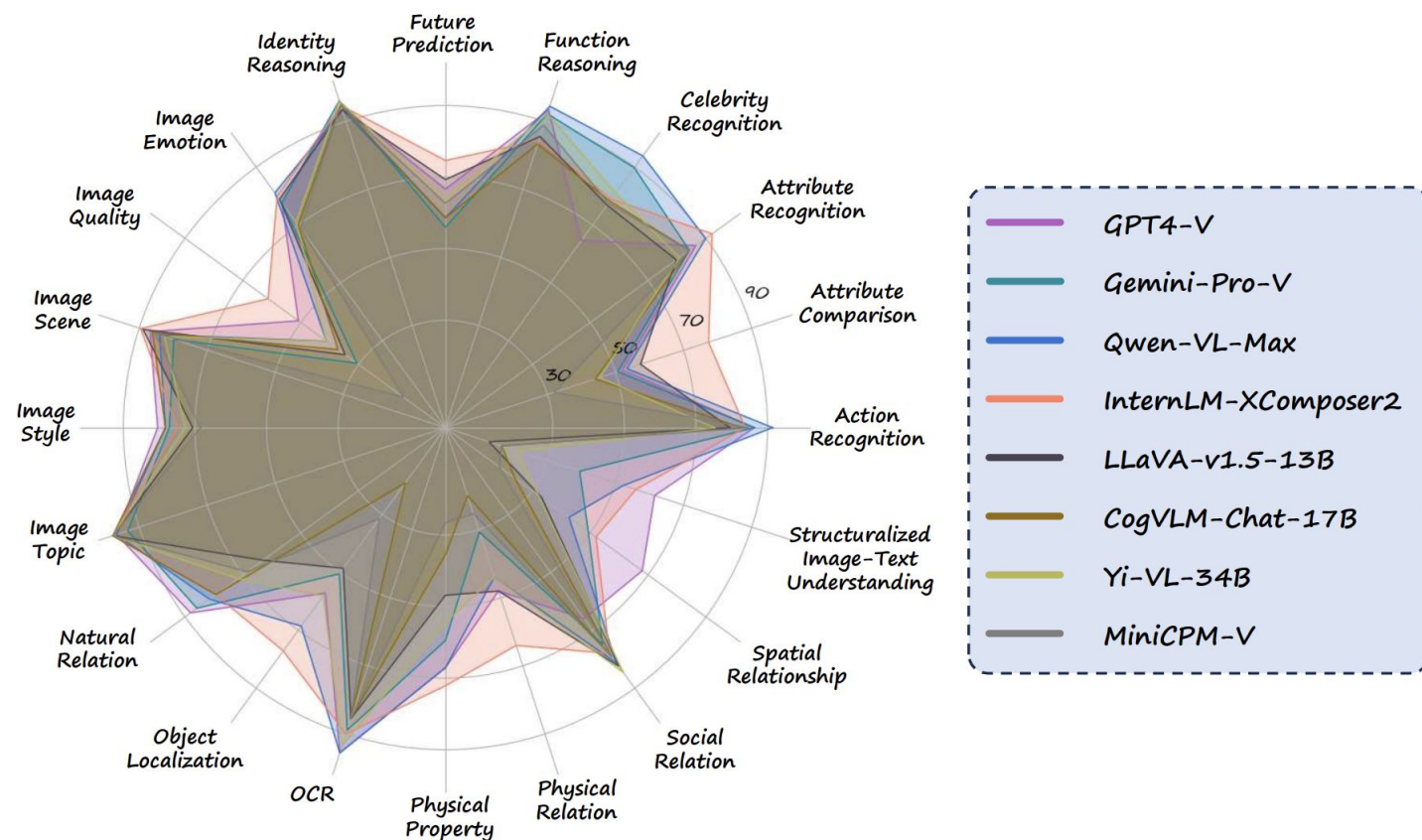


Llama3-70B-Instruct: The Grange, or the Patrons of Husbandry, was an organization founded in the late 19th century to support and advocate for the interests of farmers, particularly in the Western United States. So the answer is (A) Western farmers.

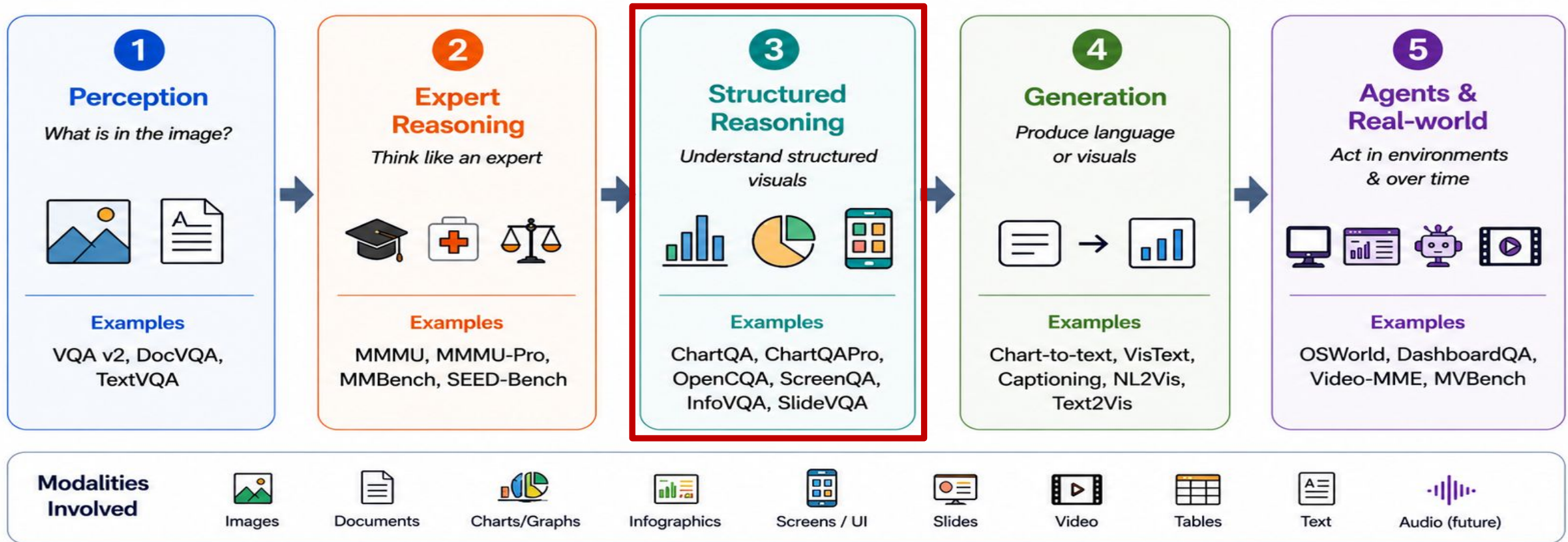
Tasks & Benchmarks: Multimodal Reasoning

- **MMBench** (*Liu et al., 2023*)

- Broad **capability evaluation**
- Multiple-choice QA
- Strong **human-aligned scoring**



Resources: Visual and Structured Modalities



Key Takeaway: Benchmarks are evolving from **static perception** → **deep reasoning** → **generation** → **agents in the real world**.

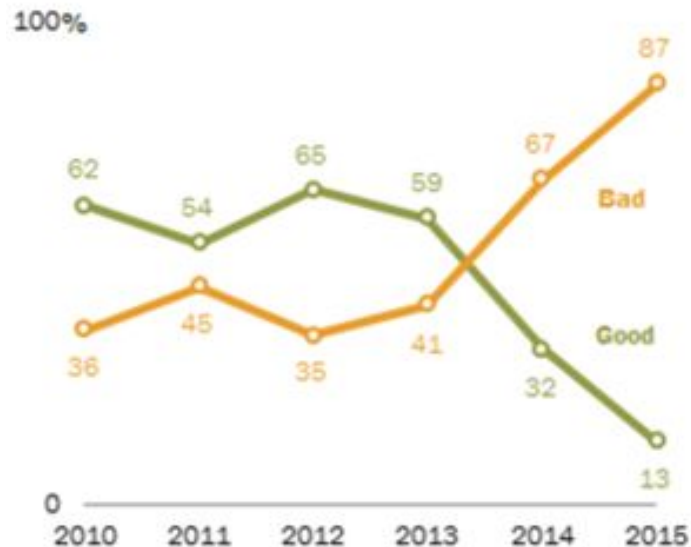
Tasks & Benchmarks: Chart Question Answering

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 Claude Sonnet 3.5 achieving +90%

Rapid Decline in Brazilians' Assessment of Economy

Current economic situation in Brazil is ...



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

Answer: 2015

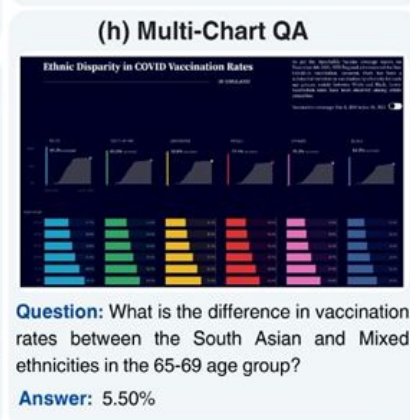
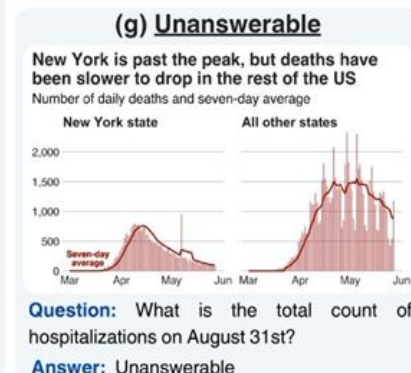
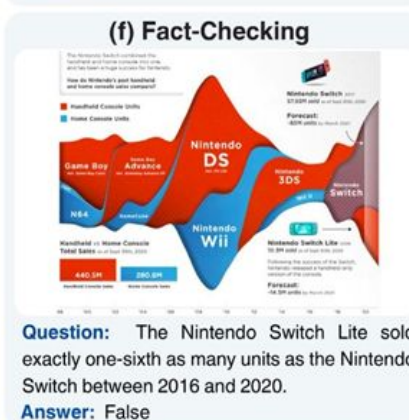
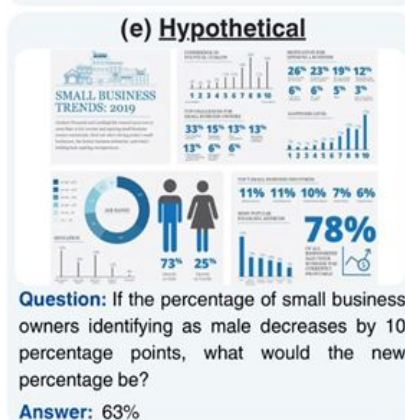
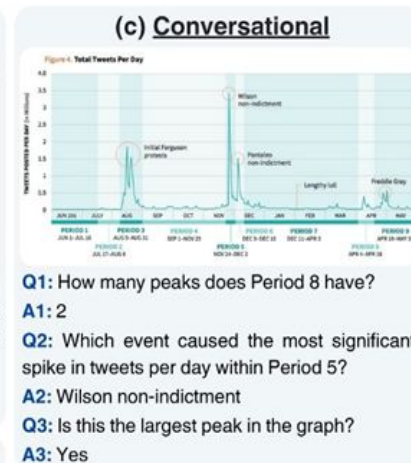
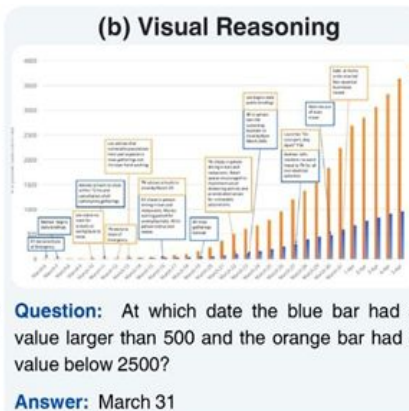
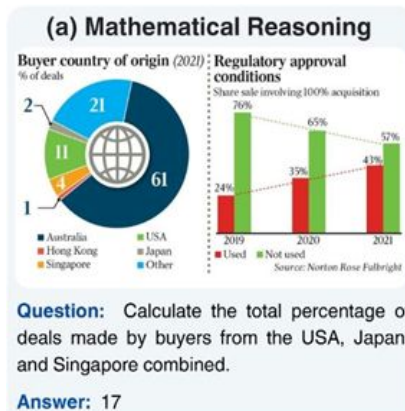
Question: What is the peak of the orange line?

Answer: 87

Tasks & Benchmarks: Chart Question Answering

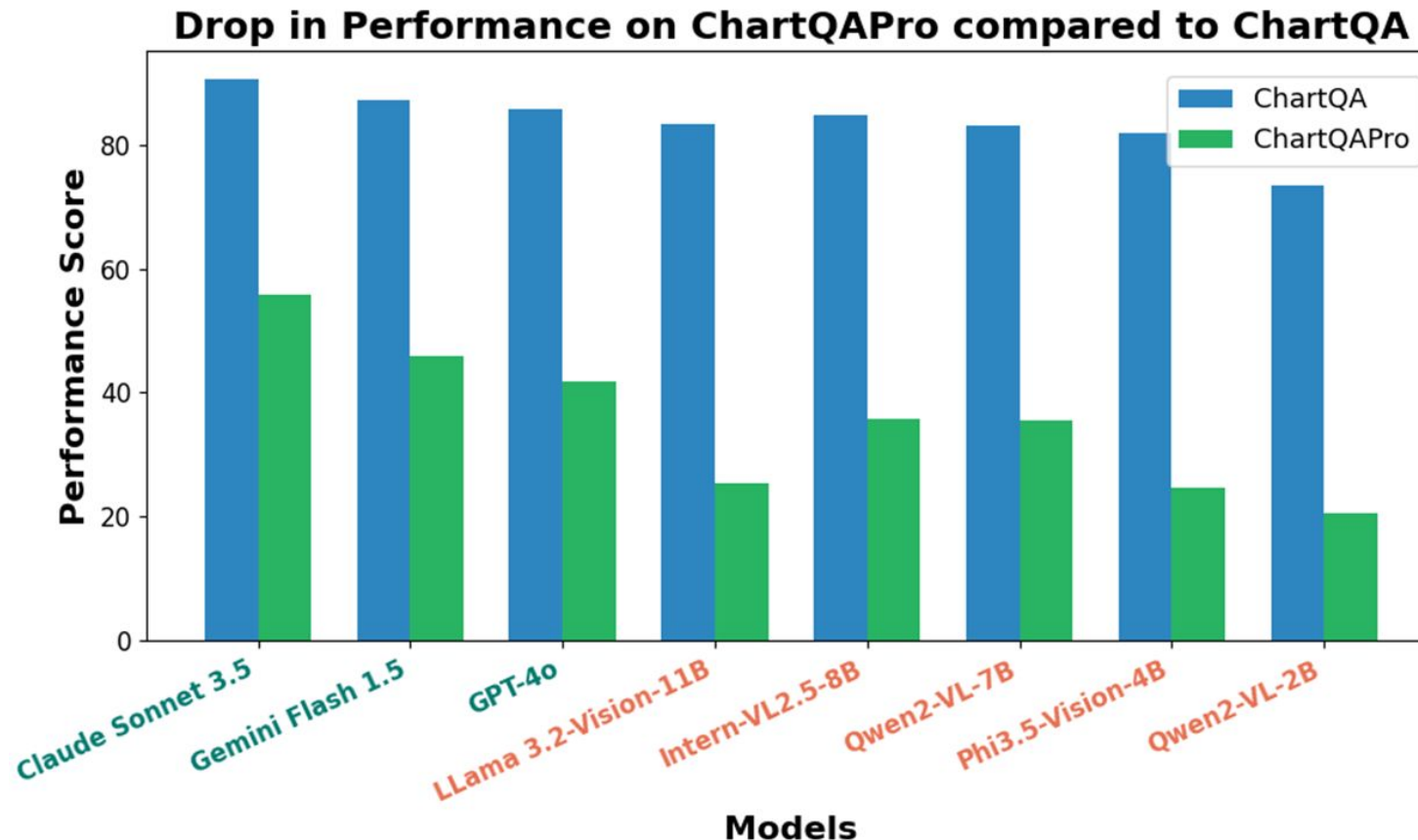
ChartQAPro (Masry et al. 2025):

- 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: Chart Question Answering

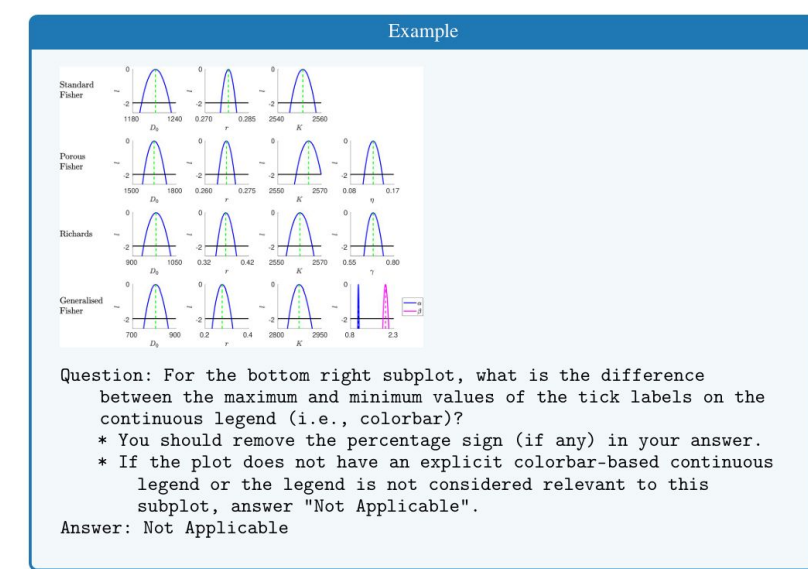
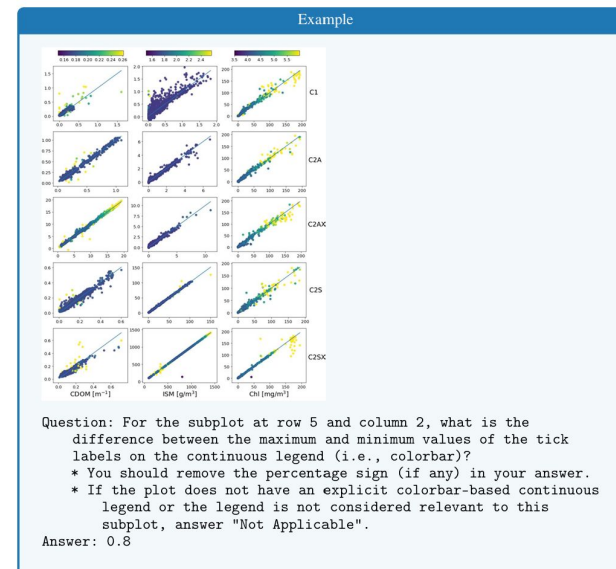
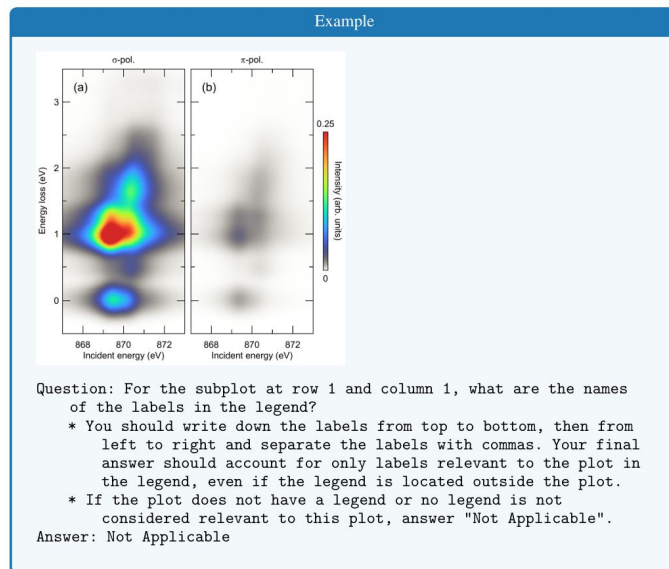
- *Drop in Performance on ChartQAPro*
- *Chart Question Answering is far from solved!*



Tasks & Benchmarks: Chart Question Answering

CharXiv (*Wang et al. 2024*)

- Scientific Charts Question Answering with 2.3K QA pairs



Tasks & Benchmarks: Open-ended Chart Question Answering

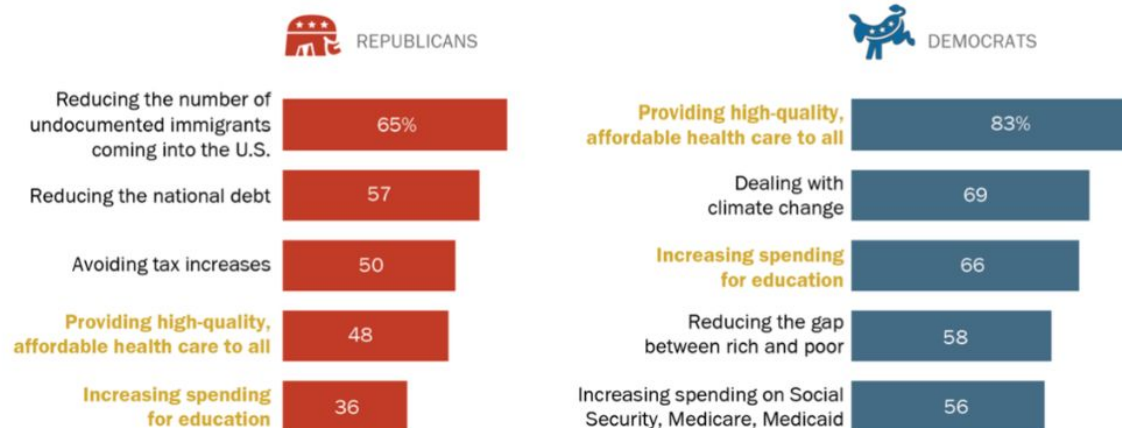
OpenCQA (*Kantharaj et al., 2022*)

- 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?

Republicans and Democrats have different ideas about what government should do to improve the lives of future generations of Americans

% of **Republicans/Democrats** saying each of the following should be a **top priority** in order for the federal government to improve the quality of life for future generations



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: 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

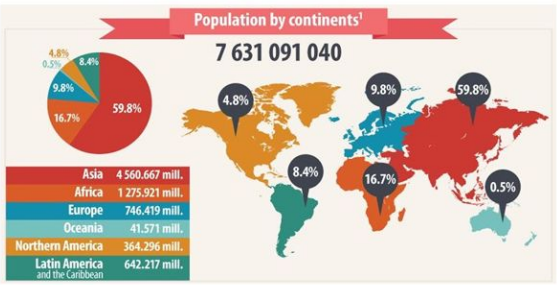
In which cities did Canada Post get maximum media coverage?

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?

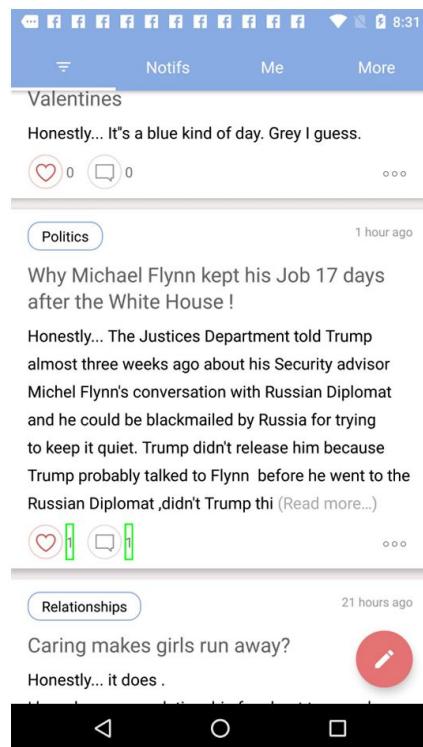


What percentage of workers are not working from home?

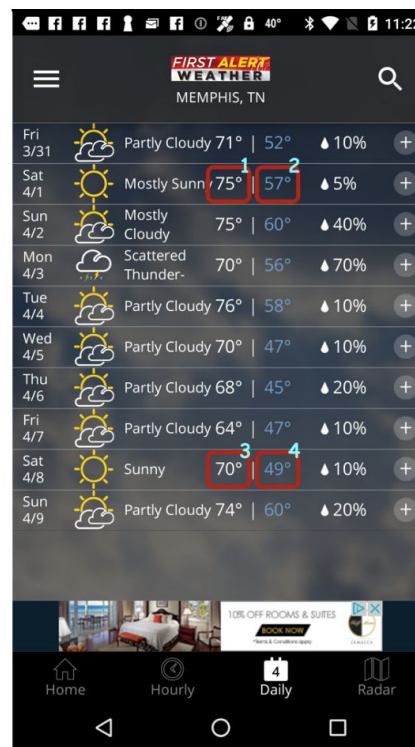
Tasks & Benchmarks: Mobile Screen Question Answering

ScreenQA (*Hsiao et al., 2022*)

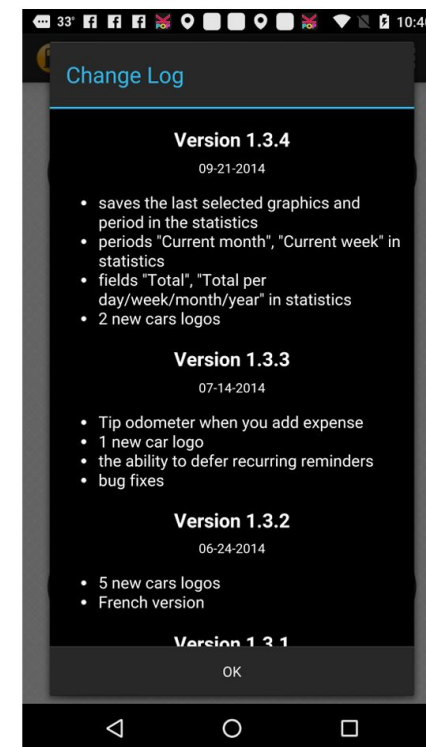
- 86K QA questions on mobile-app screenshots!



(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?”

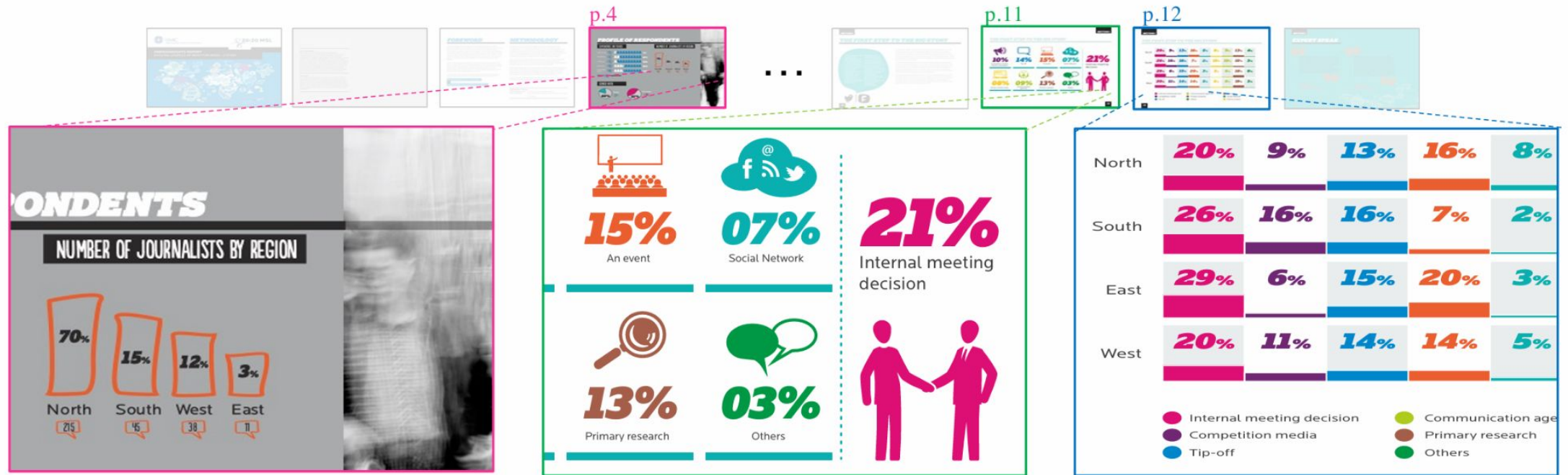


(c) No answer to the question: “What is the date of version 1.3.1?”

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%

Answer type: Multi-Span

Evidence pages: 4, 12

Reasoning type: Multi-hop

Q: What is the **percentage of the internal meeting decision**?

A: 21%

Answer type: Single-Span

Evidence pages: 11

Reasoning type: Sing-hop

Q: What is the difference in the **competition media percent age** between **East** and the **region with 12% of journalists**?

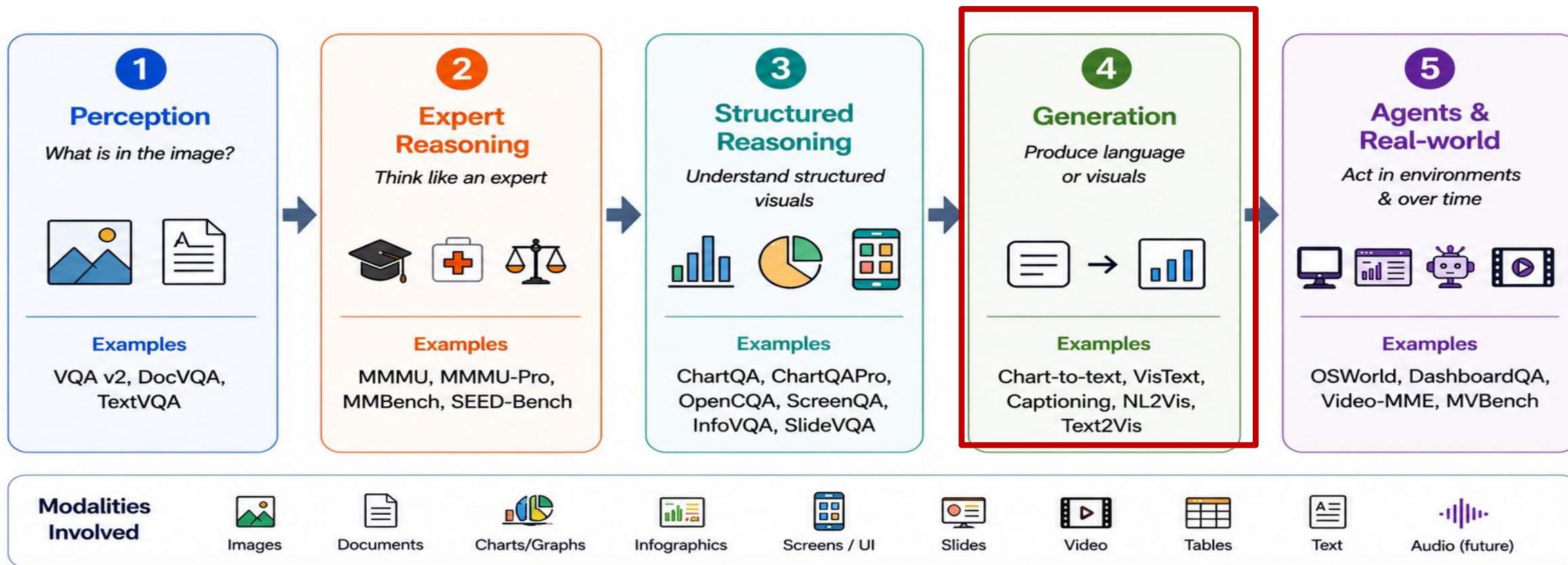
A: 5% (11% - 6%)

Answer type: Non-Span

Evidence pages: 4, 12

Reasoning type: Multi-hop, Numerical

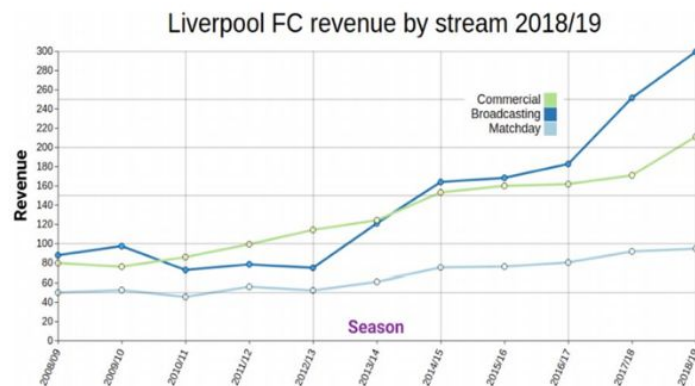
Resources: Visual and Structured Modalities



★ **Key Takeaway:** Benchmarks are evolving from **static perception** → **deep reasoning** → **generation** → **agents in the real world**.

Tasks & Benchmarks: Text Generation from Visuals

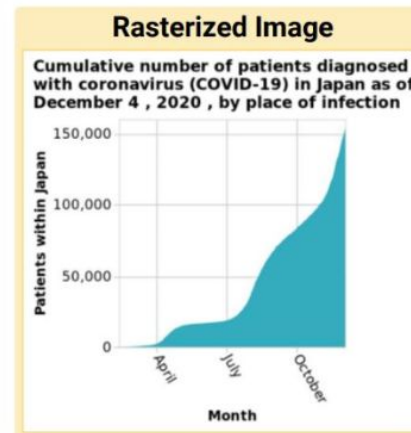
Chart Summarization



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]



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

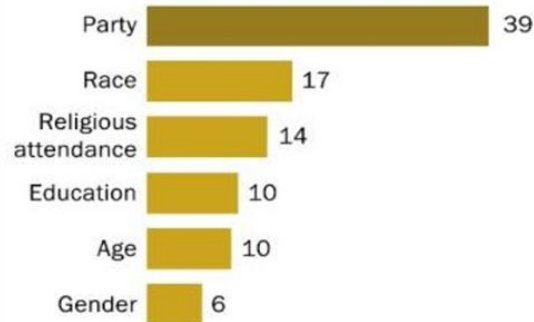
Chart-to-text: 44.1K Chart-summary pairs from Statista and Pew Research Center

Problem Setup #1

Chart:

Partisan gaps dwarf race, education, other differences in political values

Average percentage point gap across 30 political values items by ...



Notes: Indicates average gap between the share of two groups taking the same position across 30 values items. Party=difference between Rep/Lean Rep and Dem/Lean Dem. Race=white non-Hispanic vs. black non-Hispanic. Religious attendance=attend religious services weekly or more vs. attend less often. Education=college grad vs. non-college grad. Age=18-49 vs. 50+. Source: Survey of U.S. adults conducted Sept 3-15, 2019.

PEW RESEARCH CENTER

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

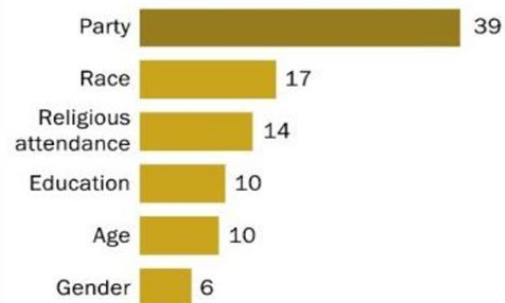
Chart-to-text: 44.1K Chart-summary pairs from Statista and Pew Research Center

Problem Setup #2

Chart:

Partisan gaps dwarf race, education, other differences in political values

Average percentage point gap across 30 political values items by ...



Notes: Indicates average gap between the share of two groups taking the same position across 30 values items. Party=difference between Rep/Lean Rep and Dem/Lean Dem. Race=white non-Hispanic vs. black non-Hispanic. Religious attendance=attend religious services weekly or more vs. attend less often. Education=college grad vs. non-college grad. Age=18-49 vs. 50+. Source: Survey of U.S. adults conducted Sept 3-15, 2019.

PEW RESEARCH CENTER

Table:

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

Metadata:

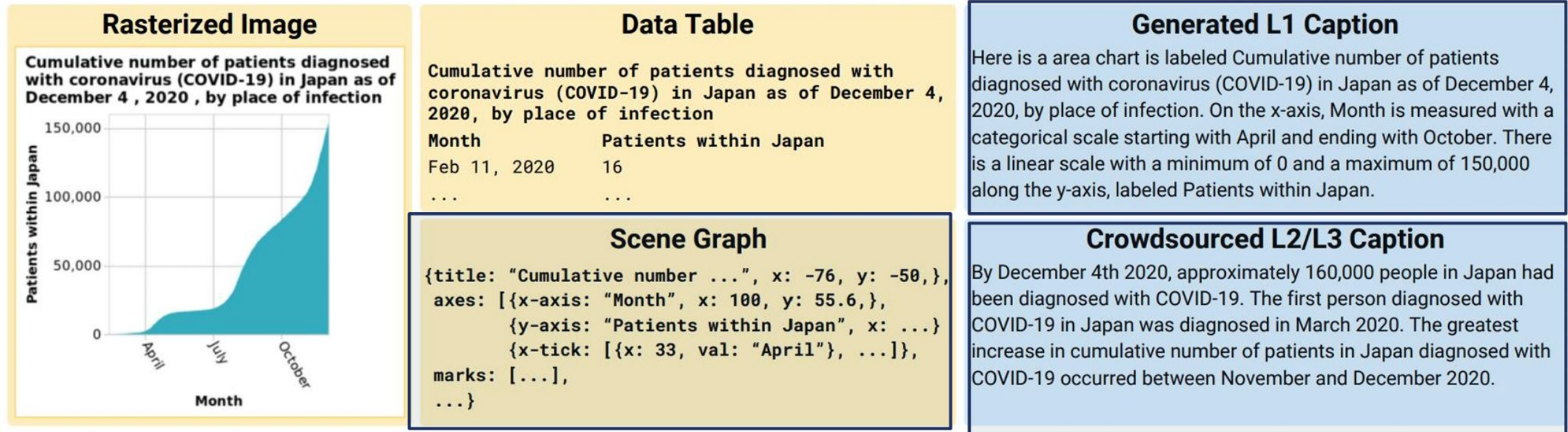
- Title: Partisan gaps dwarf race, education, other differences in political values
- Chart type: Bar
- Topic: U.S. Politics & Policy

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

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



Tasks & Benchmarks: Visualization Generation

NL2Vis (Chen et al. 2024):

- 2,524 queries covering seven chart types.

Query: Give me a stacked bar chart for showing the total number of each ship type and categorize it by ship nationality, and I want to order names in desc order.

Dataset:

Ship_ID	Name	Type	Nationality	Tonnage
1	Corbridge	Cargo ship	United Kingdom	3687
2	Farringford	Battle	United States	3146
3	Dromonby	Cargo ship	United Kingdom	3627
...	...	...	...	...

NL2VIS ✗

```
# Create a stacked bar chart for the number of each ship type by nationality
plt.bar(ship['Nationality'], ship['Type'].value_counts(), stacked=True)
```

```
grouped = ship.groupby(['Type', 'Nationality'])
# Sum the tonnage for each group
tonnage = grouped['Tonnage'].sum()
...
plt.bar(tonnage['Nationality'], tonnage['Tonnage'], color=['b', 'g'])
plt.xticks(rotation=0)
plt.legend(tonnage['Type'].unique())
```

Gemini

```
grouped = ship.groupby(['Type', 'Nationality']).size().unstack()
# Reorder the columns in descending order based on the sum of each row
sorted = grouped[grouped.sum().sort_values(ascending=False).index]
...
plt.legend(title='Nationality', bbox_to_anchor=(1.05, 1), loc='upper left')
```

GPT-3.5

```
grouped = ship.groupby(['Type', 'Nationality']).size().unstack()
# Sort the data in descending order
sorted = grouped.sort_values(by='Type', ascending=False)
# Plot a stacked bar chart
sorted.plot(kind='bar', stacked=True, figsize=(10, 8))
plt.xlabel('Ship Type')
```

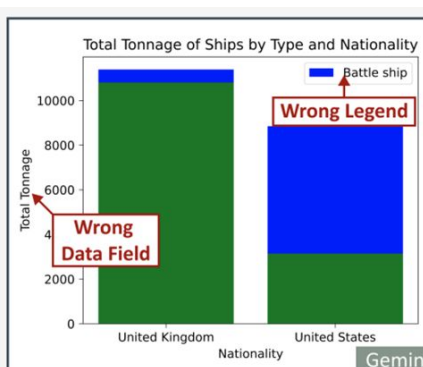
GPT-4

VisEval

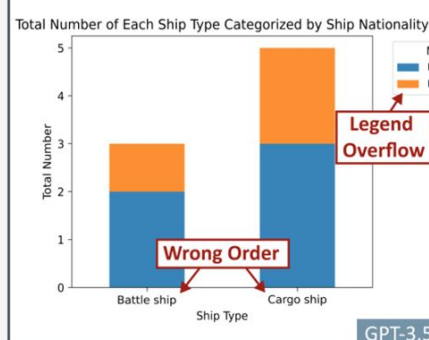
Code Crash

ValueError: shape mismatch: objects cannot be broadcast to a single shape. Mismatch is between arg 0 with shape (8,) and arg 1 with shape (2,).

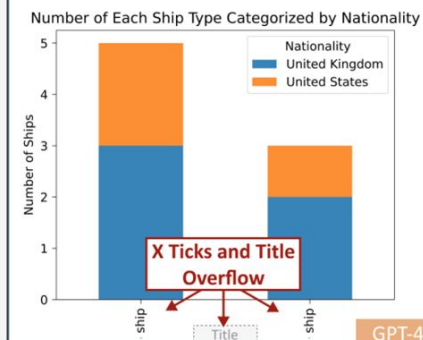
Llama



Gemini



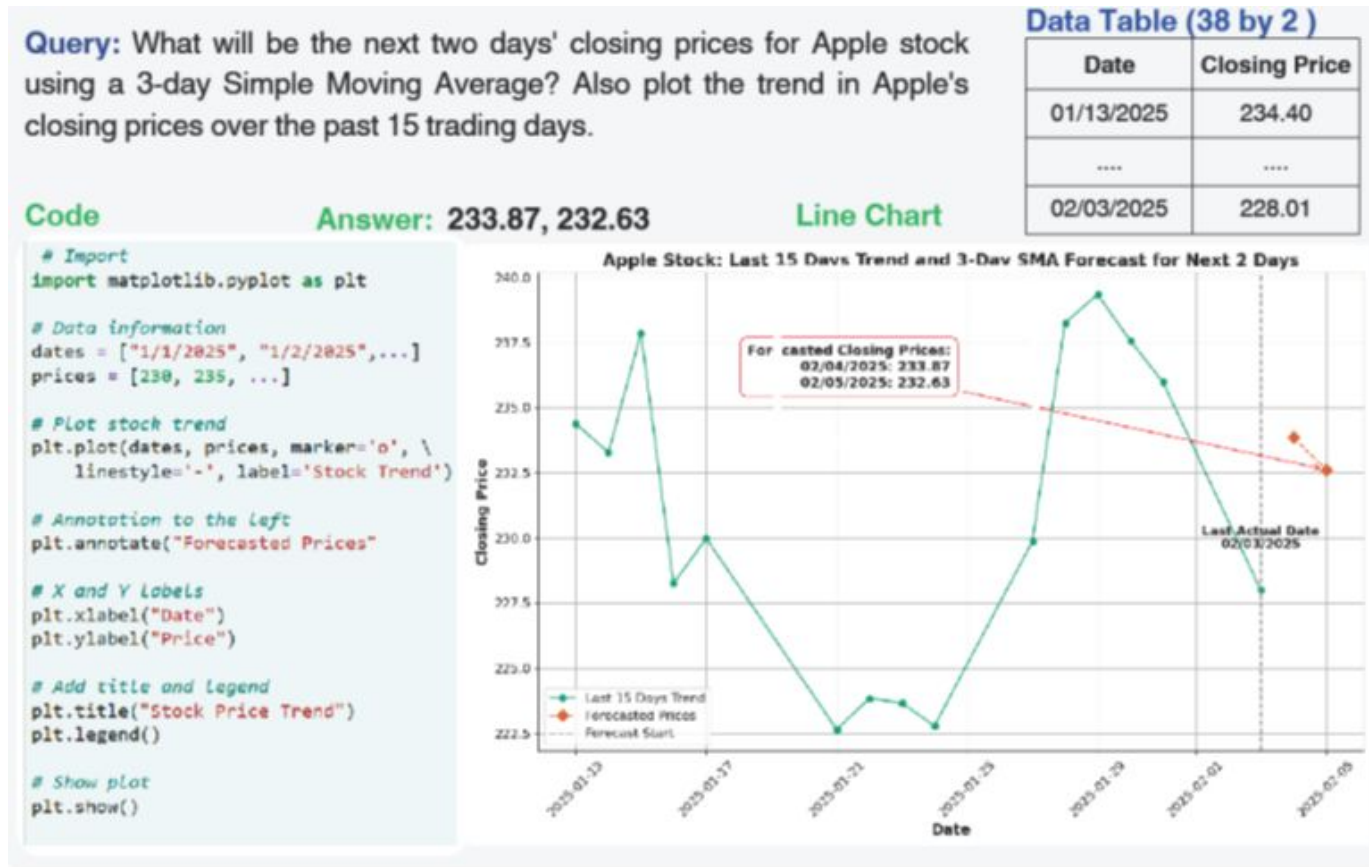
GPT-3.5



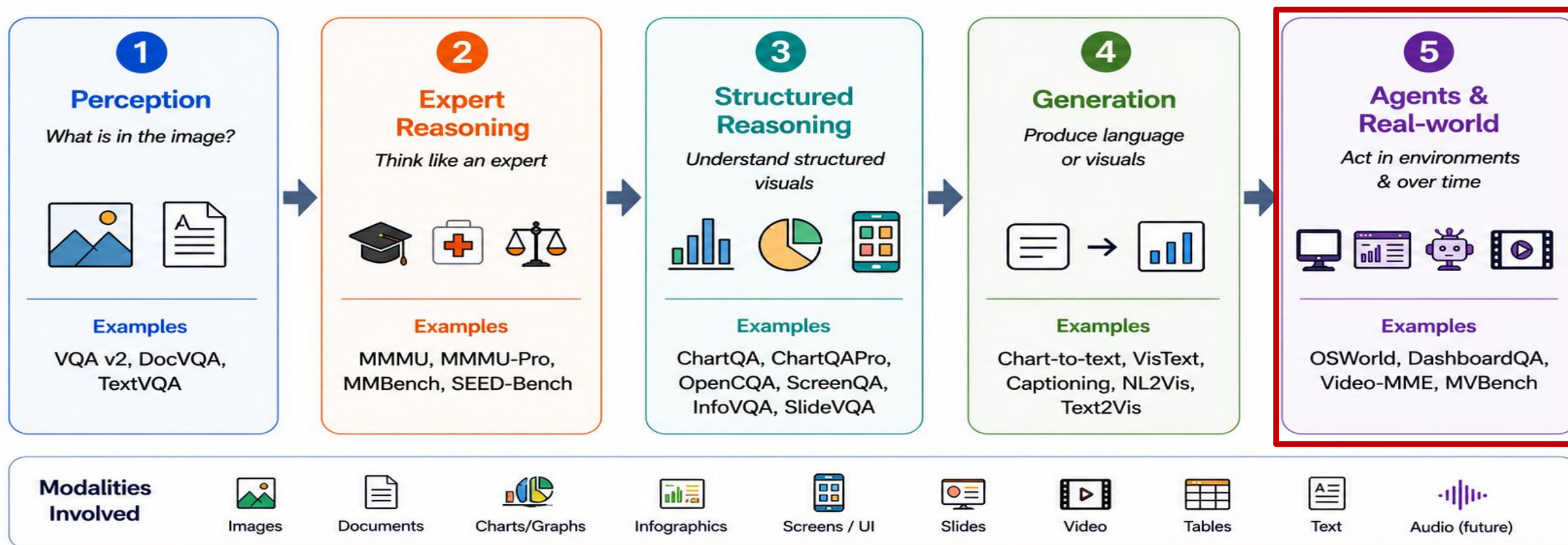
GPT-4

Tasks & Benchmarks: Visualization Generation

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

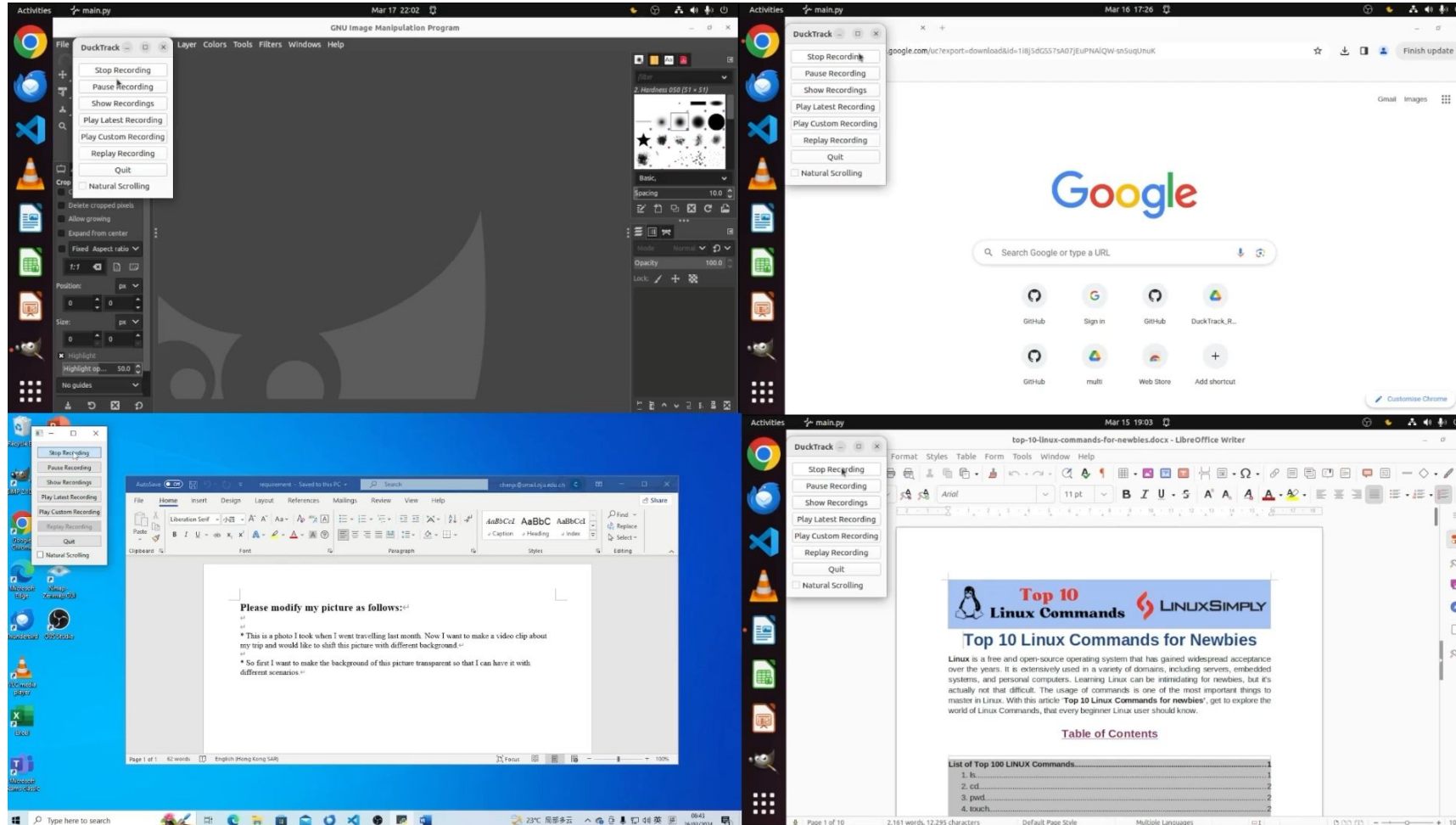


Resources: Visual and Structured Modalities



Key Takeaway: Benchmarks are evolving from **static perception** → **deep reasoning** → **generation** → **agents in the real world**.

Tasks & Benchmarks: Multimodal Real-world Tasks



OSworld (Tianbao et al., Neurips 2024)

Tasks & Benchmarks: Video Question Answering

- How can we design AI agents that *understand* and *act* on dashboards through sequential, human-like interactions with visual interfaces?
- **292 tasks** on 112 interactive dashboards, encompassing 405 question answer pairs.

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)

Tasks & Benchmarks: Video Question Answering

MVBench

- 20 challenging video tasks
- temporal reasoning + dynamic understanding
- automatically convert public video annotations into multiple-choice QA to evaluate each task.

Spatial Understanding: *Inferring from a single frame*

① Action

What's the man doing?

② Object

What's on the table?

③ Position

Is the man on the stage?

④ Count

How many chairs?

⑨ Cognition

Why is the man singing in the canteen?

⑤ Scene

Where's the man?

⑥ Pose

What's the man's pose?

⑦ Attribute

What color is the desk?

⑧ Character

What are the subtitles?



Temporal Understanding: *Reasoning based on entire video*

① Action

- Action Sequence
- Action Antonym
- Action Prediction
- Unexpected Action
- Fine-grained Action

② Object

- Object Shuffle
- Object Existence
- Object Interaction

③ Position

- Moving Direction
- Action Localization

④ Count

- Action Count
- Moving Count

⑤ Scene

- Scene Transition

⑥ Pose

- Fine-grained Pose

⑦ Attribute

- State Change
- Moving Attribute

⑧ Character

- Character Order

⑨ Cognition

- Episodic Reasoning
- Egocentric Navigation
- Counterfactual Inference

Benchmarks: Future Challenges

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 diverse **multimodal inputs** and **outputs**?

Multilingual & Low-Resource Challenges

- VLM evaluators (judges / reward models) are **critical for alignment**
- **!** Unknown: *Do VLM judges generalize across languages?*
- **MM-JudgeBench**
 - 60K+ preference pairs
 - 25 languages (incl. low-resource: Bengali, Kazakh)
 - Two complementary subsets:
 - VL-RewardBench (multilingual) → general vision-language tasks
 - OpenCQA (multilingual) → chart reasoning



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.

Multilingual & Low-Resource Challenges

- How do we enable **multimodal reasoning** in underrepresented languages?
 - Cross-lingual transfer and multimodal alignment
 - Data-efficient learning (few-shot, synthetic data, weak supervision)
- How can models generalize across **cultures** and **visual conventions**?
 - Inclusive benchmarks and culturally diverse evaluation

QA