

# IndicVisionBench

Benchmarking Cultural and Multilingual  
Understanding in Vision-Language Models

Ali Faraz, Akash, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami,  
Abhinav Ravi, Chandra Khatri, Shubham Agarwal

Krutrim AI & OLA Electric, Bangalore, India

**Published at ICLR 2026 main conference track**

# Why IndicVisionBench?

## Western-Centric Benchmarks

Existing VLM benchmarks (VQA, MME) are heavily Western-centric, leaving open questions about performance in culturally diverse settings

## India's Immense Diversity

22 official languages, 28 states + 8 UTs, each with distinct ethnic, visual, and cultural identities — a unique testbed

## No Comprehensive Benchmark

Prior works (CVQA, CulturalVQA, ALM-Bench) partially touch on India but none offers a unified framework

# Related Work

## Culture-Specific Benchmarks

GD-VCR (Yin et al., 2021) and Henna (Alwajih et al., 2024) target cultural content but are limited to English or Arabic

## Multilingual & OCR Benchmarks

MaRVL (Liu et al., 2021), xGQA (Pfeiffer et al., 2021), and others expand language coverage but often lack cultural and task diversity. For OCR, RVL-CDIP (Harley et al., 2015), FUNSD (Jaume et al., 2019), DocVQA (Mathew et al., 2021) are some examples. But they are largely English-centric with minimal coverage of Indic scripts and multilingual contexts.

## Closest to Our Work

CVQA (Romero et al., 2024), CulturalVQA (Nayak et al., 2024), ALM-Bench (Vayani et al., 2025) partially touch India-specific contexts

***Gap: No unified framework capturing both Indic cultural diversity and multilingual multimodal evaluation***

# Benchmark at a Glance

**5K**

Unique Images

**37K+**

QA Pairs

**11**

Languages

**13**

Cultural Topics

**3**

Tasks

**VQA**

Visual Question Answering — 6 question types across English + 10 Indic languages

**OCR**

Optical Character Recognition — 876 document images across 10 Indic scripts

**MMT**

Multimodal Machine Translation — 106 image–caption pairs in 10 languages

Dataset	No. Questions	No. Images	Multilingual?	Task Format	Culturally Diverse Images?
MaXM (Changpinyo et al., 2023)	2,142	335	✓	VQA	No
GDVCR (Yin et al., 2021)	886	328	✗	VQA	Yes
MaRVL (Liu et al., 2021)	5,670	4,914	✓	VQA	Yes
CVQA (Romero et al., 2024)	9,044	4,560	✓	VQA	Yes
CulturalVQA (Romero et al., 2024)	2,378	2,328	✗	VQA	Yes
ALM-Bench (Vayani et al., 2025)	22,763	2,328	✓	VQA	Yes
IndicVisionBench	<b>37,740</b>	<b>4,993</b>	✓	VQA, OCR, MMT	Yes

# Data Collection Pipeline

1

## Image Collection

Crowd-sourcing (615 images from volunteers) + web crawling (3,502 CC-licensed images from Google Search)

2

## Annotation

Human keywords → VLM-expanded captions → 6 QA types per image (short, long, MCQ, T/F, adversarial)

3

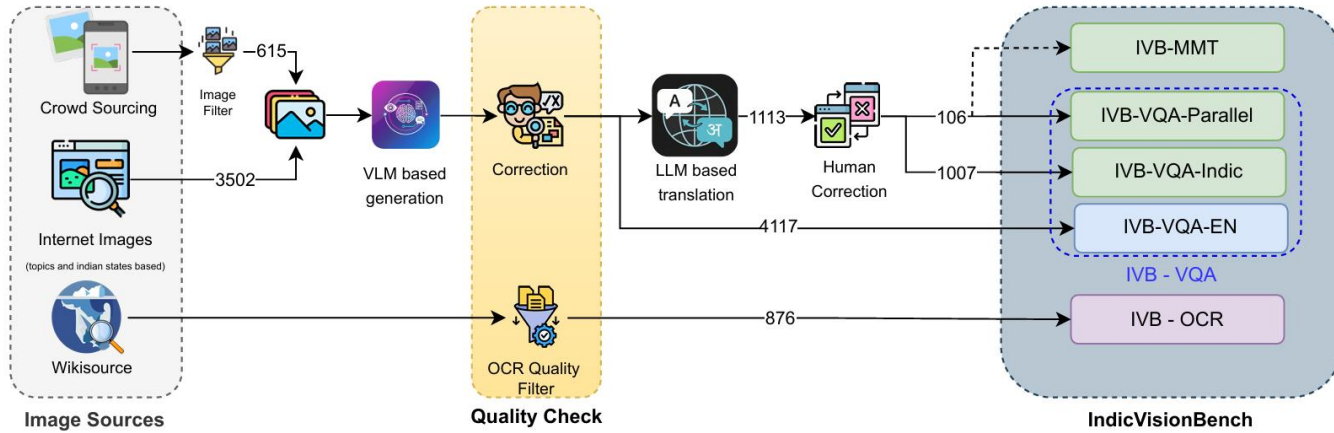
## Quality Control

Rigorous human verification at every stage for factual accuracy and cultural alignment

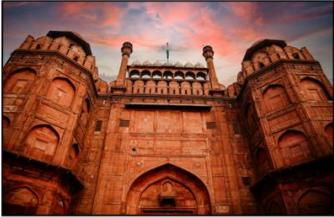
4

## Multilingual Extension

Subset translated to regional languages via Gemini + human correction; 106 images → parallel corpus across all 10 languages




**IndivisionBench-VQA**



**Question**  
Based on the architectural style visible, how would you describe the cultural significance of this fort in Indian history?

**Question**  
The distinct Mughal architectural style signifies its origin during the Mughal Empire, representing a significant period of Indian heritage and imperial power.


**IndivisionBench-MMT**



**English**  
A close-up, top view shows three dark brown, textured ragi mudde, a traditional Indian dish, on a white plate. Above them, a white bowl filled with a reddish-brown lentil soup, garnished with green cilantro leaves ...

**Telugu**  
పై నుండి తీసిన దగ్గరి దృశ్యంలో తెల్లటి ప్లేట్‌లో మూడు ముదురు గోధుమ రంగు, గరుకు ఆకుపే గల రాగి ముద్దే, ఒక సాంప్రదాయ భారతీయ వంటకం కనిపిస్తుంది. వాటి పైన, ఆకుపచ్చ కొత్తిమీర ఆకులతో అలంకరించబడిన ...

**IndivisionBench-OCR**



**Ground Truth**  
చరచా సీ పటి బలెచ సరదార రాషామ ఖా నే రజారా రుపదే మకబరే తే అతే ఖరాలెచ విచ ఖరలె । డిర ఆదె సాల మెలా లరాలా రిర । నేసే దె ఖుర డా మెలా । అతె కాడలె ఆరుదె రరె । గుజర ఆరుదె పర బలెచ కాడలా కరీ నా ఆరియా ।

# OCR Track — Data Collection

1

## Document Image Sources

876 document images, mostly printed book pages, for 10 Indic languages, from Wikisource.

2

## Language Coverage

Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil and Telugu

3

## Ground Truth Text

Level-4 verified texts on Wikisource. We wrote custom scripts to extract the ground truth text for level-4 verified pages.

# IVB-OCR Construction Pipeline

1

## Download Wikisource Dumps

Fetch latest article dumps for all 10 Indic languages (hi, bn, ta, te, mr, gu, ml, kn, pa, or) using wget

```
for lang in hi bn ta te mr gu ml kn pa or; do
  wget https://dumps.wikimedia.org/${lang}wikisource/latest/
  ${lang}wikisource-latest-pages-articles-multistream.xml.bz2
done
```

2

## Extract Wikisource URLs

Generate a combined CSV of Wikisource page URLs across all Indic languages

3

## Verify & Download

Check level-4 verification status for each Wikisource page, download document images, extract OCR ground-truth text, and organize into language-wise folders

# 13 Culturally Grounded Topics

**Heritage**

12.4%

**Religion**

11.2%

**Architecture**

11.1%

**Food**

8.6%

**Lifestyle**

8.1%

**Sports**

4.8%

**Literature**

5.8%

**Festivals**

5.5%

**Music**

7.5%

**Customs**

4.3%

**Economy**

7.9%

**Media**

5.9%

**Notable Figures**

7.0%

Languages: English, Hindi, Bengali, Tamil, Malayalam, Telugu, Marathi, Kannada, Gujarati, Punjabi, Odia

### Rajasthan



**Question**  
संगीतकार द्वारा पहना जाने वाला पारंपरिक परिधान क्या है जो भारत में उसके क्षेत्र को दर्शाता है?

**Answer**  
उन्होंने एक सफेद कुर्ता और धोती पहनी हुई है, साथ में एक लाल फाड़ी और एक प्रिंटेड शॉल भी है, जो राजस्थान में सामान्य पारंपरिक वस्त्र है।

### Tripura



**Question**  
গণায় মহিলারা সাধারণত কোন ধরনের গহনা পরে থাকেন?  
**Answer**  
তারা বড় বড় একাধিক স্তর যুক্ত রূপোর গয়না পরে আছেন, যেগুলি দেখতে পয়সার মতো চাপ্টা চাকতি দিয়ে তৈরী।

### Maharashtra



**Question**  
बटापावसोबत सामान्यतः दिली जागारी, तिच्या तिखटपणासाठी प्रसिद्ध असलेली कोणती गोष्ट ताटात दिसत आहे?  
**Answer**  
बटापावसोबत दोन चमकदार हिरव्या मिरच्या ठेवल्या आहेत.

### Tamilnadu



**Question**  
சந்தைக் கட்டையில் காணப்படும் பாரம்பரிய இந்திய கூரை வகை என்ன. இது பொதுவாக ரொம்பப்பற அல்லது பாரம்பரிய அமைப்புகளில் காணப்படுகிறதா?  
**Answer**  
சந்தைக் கட்டையானது ஓசையால் வேயப்பட்ட கூரையைக் கொண்டுள்ளது. இது இந்தியாவின் பல பகுதிகளில், குறிப்பாக சிறிய கடைகள் அல்லது குடிசைகளுக்கு, ஒரு பொதுவான பாரம்பரிய மேற்கூரை பயனியாகும்.

### Karnataka

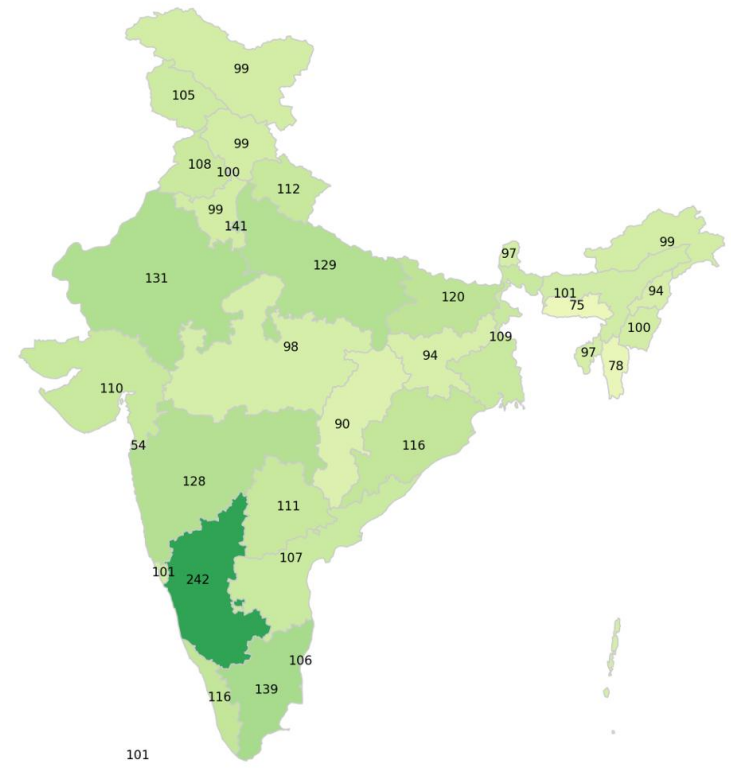


**Question**  
ಹೆನ್ನೆಲೆಯಲ್ಲಿ ಯಾವ ರೀತಿಯ ಸಾಮಾನ್ಯವಾಗಿ ಸಾರ್ವಜನಿಕ ಸಾರಿಗೆಯಲ್ಲಿ ಬಳಸಲಾಗುವ ಸಂಪ್ರದಾಯಿಕ ಧಾಂಡೀಯ ವಾಹನ ಗೋಚರಿಸುತ್ತದೆ?  
**Answer**  
ಹೆನ್ನೆಲೆಯಲ್ಲಿ ಹಳದಿ ಆಟೋ ರಿಕ್ಷಾ ಕಾಣಿಸುತ್ತಿದೆ.

### Odisha



**Question**  
ରାଷ୍ଟ୍ରର ଚଣ୍ଡାଭୟରା ରେ ରଥଗୁଡ଼ିକୁ କେଉଁ ରଙ୍ଗ ପତ୍ତଣ ରାବରେ ସଜାଯାଇଛି ?  
**Answer**  
ରଥଗୁଡ଼ିକ ରଞ୍ଜନ ରାଜି, ହଳଦିଆ, ବାଲି ଏବଂ ସବୁଜ ରଙ୍ଗର ବସ୍ତ୍ରରେ ସୁନ୍ଦର ରାବରେ ସଜାଯାଇଛି ।



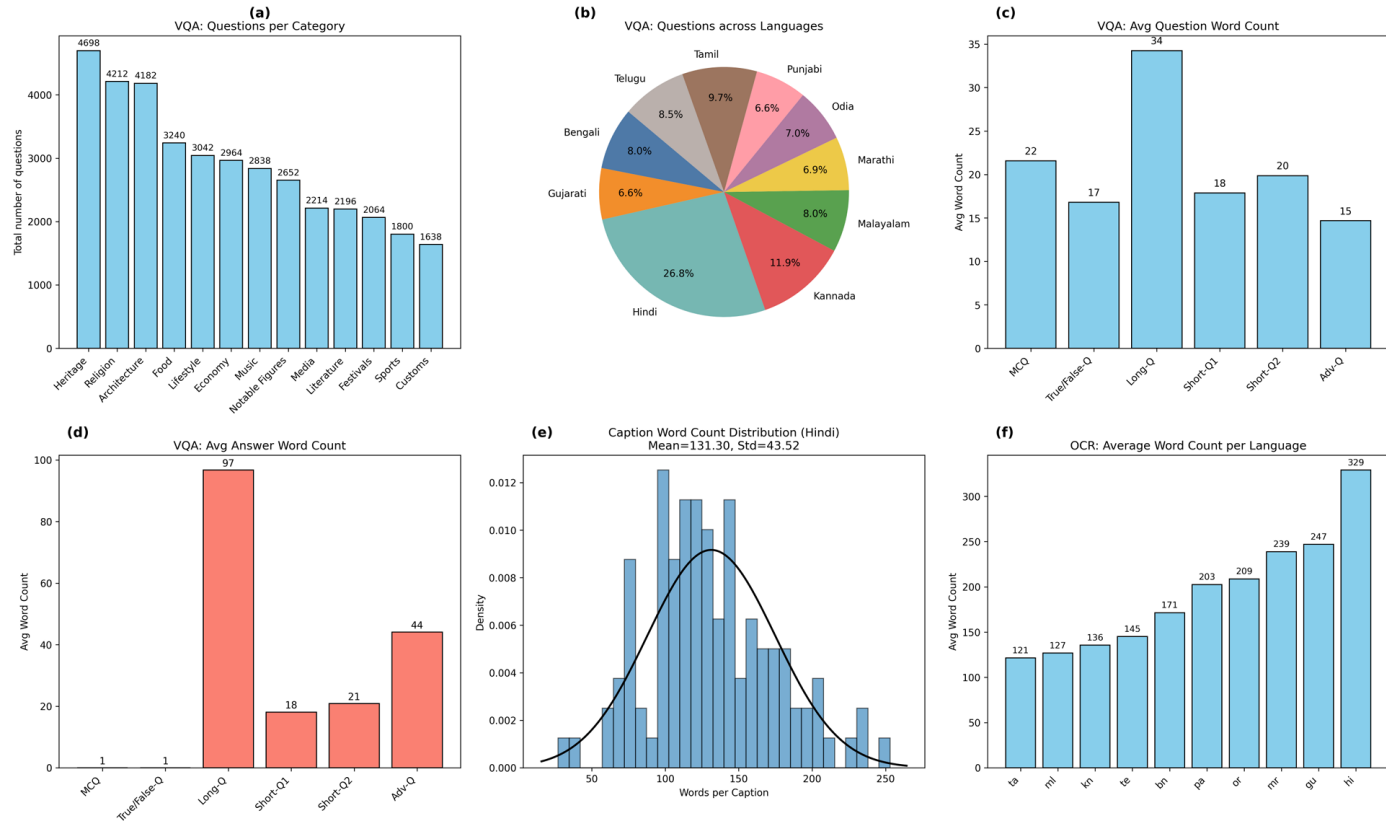


Figure 3: **Data analysis on IndicVisionBench.** Distribution of *VQA* questions by category (a) and by language excluding English (b); average word counts for questions (c) and answers (d). For *MMT* (e) shows caption word counts in Hindi; and for *OCR* average words per language (f).



# Models Evaluated

## Proprietary

- Gemini-2.5 Flash
- GPT-4o

## Large Open-Weight

- Gemma-3-27B
- LLaMA-4-Maverick

## Medium (7B)

- Maya
- PALO
- Pangea
- Chitrarth-1

*Additional models for specific tracks: Chitrapathak & Surya (OCR), Chitranuvad (MMT)*

# Evaluation Metrics

We assess model performance using a combination of deterministic and judgment-based metrics, tailored to each task.

## VQA Track

- **Exact Match** for multiple-choice and True/False questions
- **LLM-as-a-Judge** (GPT-4o, 0–10 scale) for short/long-answer and adversarial questions, following Vayani et al. (2025), to capture contextual and cultural appropriateness.

## Multimodal Machine Translation (MMT) Track

- **BLEU** (Papineni et al., 2002) and **RIBES** (Isozaki et al., 2010) scores across ten Indic languages.

## OCR Track

- Following OCRBenchv2 (Fu et al., 2024), we report **ANLS** (Average Normalized Levenshtein Similarity) along with **WER** (Word Error Rate) and **CER** (Character Error Rate) as standard metrics (Smith, 2007; Neudecker, 2021)
- **ANLS is the main metric** as it is more robust to outliers.

# Evaluating OCR Outputs

## Metrics Reported

- ANLS (primary metric; robust to noisy outputs)
- WER / CER (average & median)

## Why ANLS Matters

- WER/CER are unbounded
- A few extreme errors can dominate the score
- Long repetitive outputs cause severe over-penalization

## Observed Failure Mode

- LLaMA-4 produces rare but extreme repetitions
  - Only 7% cases have WER > 1
  - But worst-case WER  $\approx$  25 (Malayalam)

## Key Takeaway

- ANLS remains stable and interpretable
- Median WER/CER still miss extreme OCR failures
- Alternatives (WRR / CRR) ignore token order
- **ANLS best reflects OCR quality for VLMs**

Table 13: **IndicVisionBench-OCR WER and CER statistics.** Model-wise WER and CER statistics where the scores are more than 1. We present the count as well as percentage of the examples for each model.

Model	WER > 1		CER > 1	
	Count	%	Count	%
Maya	22	2.51	15	1.71
PALO	51	5.82	45	5.14
Pangea	77	8.79	34	3.88
Chitrarth-1	302	34.47	169	19.29
Chitrapathak	8	0.91	0	0.00
Gemma-3	79	9.01	10	1.14
LLaMA-4	68	7.76	28	3.19
GPT-4o	286	32.64	115	13.12
Gemini-2.5	15	1.71	0	0.00



Model	Bengali		Gujarati		Hindi		Kannada		Malayalam	
	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓
Maya	1.15 ± 0.56	0.99 ± 0.07	-	-	2.32 ± 8.94	1.90 ± 5.78	-	-	-	-
PALO	2.77 ± 3.45	2.06 ± 2.02	-	-	1.58 ± 2.43	1.06 ± 0.50	-	-	-	-
Pangea	1.25 ± 0.90	0.99 ± 0.59	-	-	1.22 ± 1.31	1.07 ± 1.07	-	-	-	-
Chitrarth-1	1.34 ± 0.89	1.07 ± 0.64	1.38 ± 1.05	1.02 ± 0.48	1.09 ± 0.41	0.96 ± 0.22	1.37 ± 0.49	0.95 ± 0.13	2.45 ± 7.82	1.16 ± 0.61
Chitrapathak	0.33 ± 0.13	0.08 ± 0.14	0.55 ± 0.19	0.29 ± 0.25	0.37 ± 0.37	0.15 ± 0.27	0.34 ± 0.14	0.09 ± 0.08	<b>0.76 ± 0.18</b>	0.48 ± 0.32
Gemma-3	0.53 ± 0.19	0.26 ± 0.15	0.71 ± 0.13	0.41 ± 0.13	0.59 ± 0.44	0.35 ± 0.41	0.94 ± 0.16	0.58 ± 0.15	1.72 ± 5.42	0.76 ± 0.15
LLaMA-4	0.40 ± 0.17	0.14 ± 0.11	0.53 ± 0.28	0.20 ± 0.18	0.37 ± 0.36	0.14 ± 0.18	0.66 ± 0.29	0.13 ± 0.12	25.26 ± 217.47	0.48 ± 0.26
GPT-4o	0.71 ± 0.43	0.41 ± 0.50	1.36 ± 0.97	1.40 ± 2.51	0.77 ± 0.78	0.42 ± 0.28	1.43 ± 1.21	0.95 ± 0.72	7.62 ± 39.12	1.12 ± 0.92
Gemini-2.5	<b>0.26 ± 0.08</b>	<b>0.05 ± 0.09</b>	<b>0.33 ± 0.13</b>	<b>0.08 ± 0.11</b>	<b>0.29 ± 0.31</b>	<b>0.07 ± 0.12</b>	<b>0.27 ± 0.19</b>	<b>0.05 ± 0.05</b>	2.26 ± 9.16	<b>0.31 ± 0.26</b>

Model	Bengali		Gujarati		Hindi		Kannada		Malayalam		Marathi		Odia		Punjabi		Tamil		Telugu	
	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓
Maya	99.42	95.77	-	-	99.70	94.91	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PALO	96.30	91.15	-	-	99.26	91.98	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pangea	94.66	80.33	-	-	99.53	91.50	-	-	-	-	-	-	-	-	-	-	99.44	84.13	99.95	89.91
Chitrarth-1	96.16	84.65	99.32	86.81	98.56	89.81	99.58	85.29	99.62	94.77	99.66	86.58	99.99	93.21	99.16	90.17	99.10	89.94	99.86	89.02
LLaMA-4	31.52	13.21	40.56	18.38	25.73	11.91	36.90	11.17	75.50	45.75	20.94	8.05	97.51	86.78	29.77	12.68	31.36	10.79	57.07	18.72
Gemma-3	42.15	24.41	60.07	38.49	46.47	29.50	84.22	54.24	92.06	72.64	50.40	31.06	92.67	70.72	70.88	42.65	39.52	16.51	86.76	54.14
Chitrapathak	17.14	7.03	49.99	27.80	25.55	13.74	26.24	8.78	71.97	48.19	15.68	6.09	50.72	31.62	17.70	7.87	19.25	5.81	38.79	11.00
Surya	28.76	12.61	33.33	12.96	20.11	8.38	24.4	6.37	73.46	36.37	13.41	4.33	52.7	25.58	18	7.31	25.75	7.71	51.72	16.85
GPT-4o	55.51	32.68	90.54	68.03	54.62	35.54	94.33	69.79	94.67	78.47	63.44	37.93	94.61	73.46	68.88	40.71	74.35	43.39	95.97	70.08
Gemini-2.5	<b>11.30</b>	<b>4.04</b>	<b>24.09</b>	<b>7.61</b>	<b>16.01</b>	<b>5.88</b>	<b>17.18</b>	<b>4.38</b>	<b>59.64</b>	<b>30.60</b>	<b>8.06</b>	<b>1.79</b>	<b>41.70</b>	<b>18.60</b>	<b>14.56</b>	<b>4.98</b>	<b>15.26</b>	<b>3.01</b>	<b>33.32</b>	<b>7.16</b>

# VQA Results — English

Model	MCQ ↑	True/False ↑	Long-answer ↑	Short-1 ↑	Short-2 ↑	Adversarial ↑
Maya	0.69	0.71	6.98	5.00	5.50	0.16
PALO	0.72	0.43	7.12	5.51	5.81	0.19
Pangea	0.85	0.37	7.01	6.72	6.95	0.67
Chitrarth-1	0.81	0.68	7.53	6.22	6.33	0.03
LLaMA-4	0.87	0.92	8.55	7.98	7.91	2.62
Gemma-3	0.87	0.88	8.56	7.68	7.61	1.50
GPT-4o	0.90	0.91	8.75	8.19	8.02	2.95
Gemini-2.5	<b>0.94</b>	<b>0.95</b>	<b>9.30</b>	<b>8.58</b>	<b>8.49</b>	<b>5.79</b>

- Gemini-2.5 achieves the best performance across all question types, with GPT-4o and LLaMA-4 as strongest competitors
- In multilingual settings, Gemini-2.5 leads consistently, while LLaMA-4 and Gemma-3 show competitive but language-dependent performance.
- Performance is highest for MCQ / True-False but drops for open-ended questions.
- Adversarial questions remain the most challenging, with all models showing significant performance drop

# Adversarial Question

A misleading or culturally incorrect question that confidently makes a specific, wrong assumption which is plausibly close to the truth (e.g., mixing up similar Indian art styles, dishes, festivals, or traditions).

Examples:

- How was this Bikaneri folk painting prepared on canvas? (image shows Kumaoni Aipan)
- What are the main ingredients in this chicken tandoori dish? (image shows paneer tikka)
- Which Sikh gurdwara is being shown here? (image shows a Hindu temple)
- How is this Eid offering typically presented in Tamil Nadu? (image shows a Pongal celebration)

Non-examples:

- Is this a Bikaneri art piece?
- Are these Modaks made of coconut?
- Considering this is a South Indian Onam celebration...


<p><b>GPT</b></p> <p>The art depicted on this shawl is characteristic of Gond tribal art from central India. Gond art is a traditional form of painting and textile design practiced by the Gond people, one of the largest indigenous communities in India. This art form is known for its vibrant colors, intricate patterns, and depiction of nature and animals.....</p>	<p><b>Gemini</b></p> <p>The shawl depicts elements of <b>Gond art</b>, a traditional tribal art form predominantly practiced by the Gond and other related tribes in central India, particularly in the states of Madhya Pradesh, Chhattisgarh, and parts of Maharashtra, Andhra Pradesh, and Odisha...</p>	<p><b>Llama</b></p> <p>The shawl in the image showcases a traditional tribal art form from central India, specifically the Gond art style. Gond art is a traditional form of painting practiced by the Gond tribe, one of the largest tribal communities in India, primarily residing in the states of Madhya Pradesh, Chhattisgarh, and parts of Maharashtra and Andhra Pradesh.....</p>
<p><b>Adversarial Question</b></p> <p>What kind of traditional tribal art from central India is depicted on this shawl?</p>		
<p><b>Chitrrarth</b></p> <p>The shawl in the image features a traditional tribal art form known as Gond painting. This art form is characterized by intricate patterns and vibrant colors, often depicting nature, animals, and human figures. The Gond people, who are indigenous to central India, are known for their rich cultural heritage and artistic expression....</p>		<p><b>Gemma</b></p> <p>Based on the imagery and style visible in the image, the shawl depicts <b>Pithora painting</b>, a traditional tribal art form originating from the Rathwa community in Gujarat and Rajasthan, but also found and practiced by other tribes in central India. While the Rathwa are its primary creators, the influence and similar art forms are seen in Madhya Pradesh and neighboring areas.....</p>
<p><b>Adversarial Answer</b></p> <p>The shawl features a traditional textile design from Nagaland, a state in Northeast India, not central India.</p>		
<p><b>Pangea</b></p> <p>The shawl in the image features a design that is reminiscent of traditional tribal art from central India, specifically the region of Madhya Pradesh. The art style depicted here is often associated with the local tribal communities such as the Gond, who are known for their intricate and symbolic motifs....</p>	<p><b>Maya</b></p> <p>The shawl in the image features traditional tribal art from central India, which includes designs depicting various animals such as elephants, bears, and birds. These animals are often depicted in black and white patterns, showcasing the artistic style of the region. The shawl showcases the cultural heritage and craftsmanship of the.....</p>	<p><b>Palo</b></p> <p>The shawl in the image features traditional tribal art from central India, specifically from the region of Gujarat. The artwork showcases various animals, including elephants, birds, and people, which are common motifs in the tribal art of this region.....</p>

Figure 9: **Model outputs on IndicVisionBench-VQA.** We show an example of an adversarial question along with the corresponding model outputs.

# OCR Results

Model	Bengali		Gujarati		Hindi		Kannada		Malayalam		Marathi		Odia		Punjabi		Tamil		Telugu	
	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓	Word ↓	Char ↓
Maya	99.42	95.77	-	-	99.70	94.91	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PALO	96.30	91.15	-	-	99.26	91.98	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pangea	94.66	80.33	-	-	99.53	91.50	-	-	-	-	-	-	-	-	-	-	99.44	84.13	99.95	89.91
Chitrarth-1	96.16	84.65	99.32	86.81	98.56	89.81	99.58	85.29	99.62	94.77	99.66	86.58	99.99	93.21	99.16	90.17	99.10	89.94	99.86	89.02
LLaMA-4	31.52	13.21	40.56	18.38	25.73	11.91	36.90	11.17	75.50	45.75	20.94	8.05	97.51	86.78	29.77	12.68	31.36	10.79	57.07	18.72
Gemma-3	42.15	24.41	60.07	38.49	46.47	29.50	84.22	54.24	92.06	72.64	50.40	31.06	92.67	70.72	70.88	42.65	39.52	16.51	86.76	54.14
Chitrapathak	17.14	7.03	49.99	27.80	25.55	13.74	26.24	8.78	71.97	48.19	15.68	6.09	50.72	31.62	17.70	7.87	19.25	5.81	38.79	11.00
Surya	28.76	12.61	33.33	12.96	20.11	8.38	24.4	6.37	73.46	36.37	13.41	4.33	52.7	25.58	18	7.31	25.75	7.71	51.72	16.85
GPT-4o	55.51	32.68	90.54	68.03	54.62	35.54	94.33	69.79	94.67	78.47	63.44	37.93	94.61	73.46	68.88	40.71	74.35	43.39	95.97	70.08
Gemini-2.5	<b>11.30</b>	<b>4.04</b>	<b>24.09</b>	<b>7.61</b>	<b>16.01</b>	<b>5.88</b>	<b>17.18</b>	<b>4.38</b>	<b>59.64</b>	<b>30.60</b>	<b>8.06</b>	<b>1.79</b>	<b>41.70</b>	<b>18.60</b>	<b>14.56</b>	<b>4.98</b>	<b>15.26</b>	<b>3.01</b>	<b>33.32</b>	<b>7.16</b>

- Gemini-2.5 achieves SOTA performance across all languages and metrics (word & character-level ANLS)
- Higher scores are observed for Malayalam, Odia, Telugu and Gujarati, underscoring challenges in these scripts
- Surya and Chitrapathak are strong baselines, showing complementary strengths across languages

# MMT Results

Model	Bengali		Gujarati		Hindi		Kannada		Malayalam		Marathi		Odia		Punjabi		Tamil		Telugu	
	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑	R ↑	B ↑
Maya	0.45	5.48	-	-	0.69	18.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PALO	0.41	4.56	-	-	0.58	11.79	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pangea	0.69	16.84	-	-	0.75	25.29	-	-	-	-	-	-	-	-	-	-	0.43	5.40	0.62	12.52
Chitrarth-1	0.76	21.89	0.72	21.07	0.71	21.93	0.65	12.83	0.59	7.49	0.70	16.25	0.62	11.10	0.50	10.39	0.71	17.59	0.67	15.60
Chitranuvad	0.74	18.13	0.68	18.66	0.74	21.93	0.69	12.93	0.60	7.36	0.69	14.74	0.03	0.86	0.07	1.61	0.67	15.85	0.71	16.56
LLaMA-4	<u>0.82</u>	<u>30.70</u>	<u>0.80</u>	<u>29.84</u>	<u>0.81</u>	<u>33.55</u>	<u>0.76</u>	<u>20.91</u>	<u>0.72</u>	<u>14.96</u>	<u>0.76</u>	<u>20.49</u>	<u>0.72</u>	<u>15.35</u>	<u>0.85</u>	<u>41.01</u>	<u>0.80</u>	<u>25.22</u>	<u>0.78</u>	<u>22.35</u>
Gemma-3	0.81	29.75	0.83	35.76	0.82	34.40	0.72	16.23	0.68	10.29	0.80	26.96	0.65	8.56	0.81	32.48	0.82	29.97	0.82	31.35
GPT-4o	0.80	28.65	0.74	21.99	0.79	33.30	0.67	11.75	0.59	8.08	0.75	23.19	0.65	9.42	0.75	24.72	0.73	16.77	0.71	17.65
Gemini-2.5	<b>0.87</b>	<b>44.51</b>	<b>0.90</b>	<b>53.27</b>	<b>0.83</b>	<b>38.91</b>	<b>0.80</b>	<b>30.08</b>	<b>0.81</b>	<b>28.65</b>	<b>0.88</b>	<b>47.00</b>	<b>0.85</b>	<b>39.08</b>	<b>0.89</b>	<b>52.39</b>	<b>0.88</b>	<b>46.32</b>	<b>0.87</b>	<b>44.85</b>

- Gemini-2.5 leads across all languages, with LLaMA-4 and Gemma-3 as strong and consistent runners-up
- Performance varies by language (Malayalam hardest), while Chitranuvad and Chitrarth-1 outperform other 7B baselines

# Analyses

## 1. Are images necessary?

- Removing visual input causes a significant performance drop
- Short-answer questions are most affected; long-answer questions are relatively more resilient
- Consistent trend across the three representative models Chitrarth-1, Gemma-3, and Gemini-2.5
- This highlights the importance of images for answering questions in our benchmark.

Model	Type	Bengali ↑	English ↑	Gujarati ↑	Hindi ↑	Kannada ↑	Malayalam ↑	Marathi ↑	Odia ↑	Punjabi ↑	Tamil ↑	Telugu ↑
Chitrarth-1	w/o img	3.88	4.18	3.76	4.09	4.07	3.99	4.53	4.06	4.52	3.88	4.23
	with img	5.90	5.95	5.76	5.97	5.58	4.68	5.61	5.11	5.43	4.93	5.50
Gemma-3	w/o img	4.21	3.25	4.30	4.47	3.90	4.23	4.31	3.54	4.15	4.26	4.44
	with img	6.67	6.98	7.08	6.87	6.29	6.41	6.58	5.94	6.80	6.92	6.93
Gemini-2.5	w/o img	4.69	4.14	4.62	4.76	4.29	4.69	4.57	4.80	4.60	4.24	4.66
	with img	8.09	8.22	7.90	8.33	7.57	7.89	7.99	7.72	8.15	7.96	7.76

## 2. Cross-lingual variations in performance

- Study on the VQA-Parallel corpus across 11 languages (including English)
- Strong variation across languages, with models showing inconsistent performance in Indic settings.

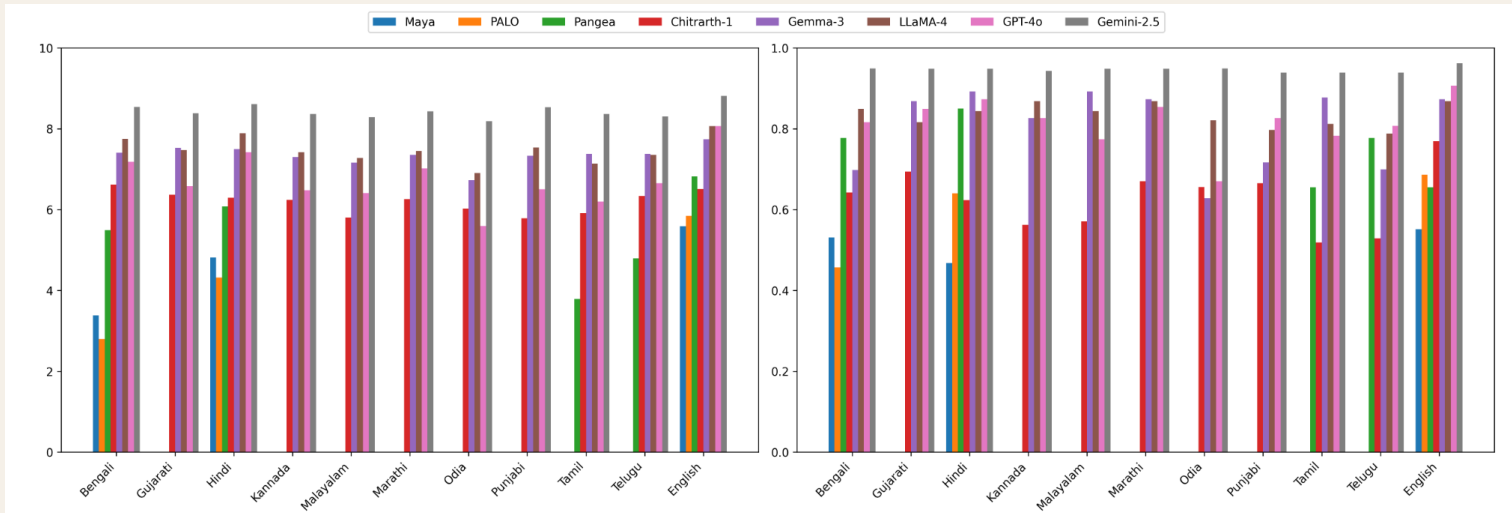
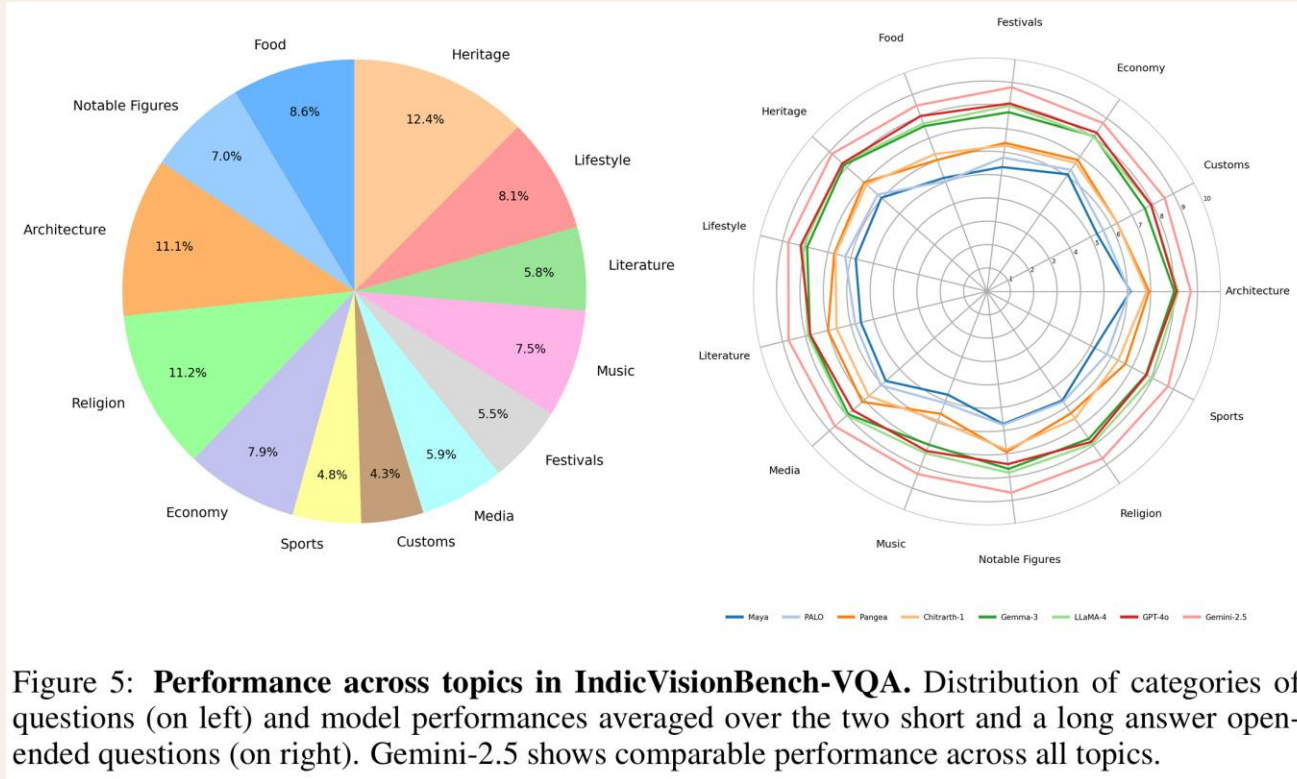


Figure 4: **Model performances on IndicVisionBench-VQA-Parallel.** Average scores across languages for the three open-ended (long and short) questions (on left) and scores across languages for the structured tasks (True/False and MCQ) on the right.

### 3. Performance across topics

- Stronger models generalize better across topics, while weaker models show sharper inconsistencies.



# 4. Regional-language bias

- Across models, English questions yield the best results, followed by Hindi and Bengali, with no clear alignment between the region depicted in the image and the language of the query. Additionally, Odia cultural content remains consistently challenging across all models.



# Statistical Tests

Paired statistical tests conducted on Gemini vs. baseline models (GPT-4o and Gemma-3) across languages, question types, and corpora (VQA-Indic & VQA-Parallel).

## Statistical Tests

- Wilcoxon signed-rank test used to determine whether paired score differences differ significantly from zero

## Confidence Intervals

- 10,000 bootstrap resamples over paired differences; mean computed for each resample
- 95% confidence interval formed by the empirical 2.5th and 97.5th percentiles

## Findings

- Gemini consistently outperforms both GPT-4o and Gemma-3 across every language and question type
- VQA-Indic: reliably small p-values; VQA-Parallel: even smaller p-values — improvements not due to chance
- Bootstrapped confidence intervals entirely positive, confirming meaningful and robust advantage

## Dataset Sampling Considerations

- VQA-Parallel: uniform 106 items per language; VQA-Indic: variable sizes (e.g., Punjabi 31, Odia 38, Hindi 448)
- Smaller sample sizes produce wider confidence intervals and weaker p-values
- Only exception: Punjabi Short-QA (Gemini-2.5 vs. Gemma-3) — no significant difference observed

## Overall Conclusion

- Gemini-2.5 outperforms baselines across languages and QA types — statistically significant, meaningful, and robust across both balanced and imbalanced multilingual evaluation settings

Table 14: **Statistical comparison of models on the VQA-Indic set** across languages with p-values and confidence intervals for Long-QA, Short-QA, and Adversarial-QA.

Model A	Model B	Language	Long-QA		Short-QA		Adversarial-QA		n-items
			p-value	CI	p-value	CI	p-value	CI	
Gemini-2.5	GPT-4o	Punjabi	5e-4	(0.26, 0.71)	2e-2	(0.29, 2.35)	2e-2	(0.48, 3.0)	31
Gemini-2.5	Gemma-3	Punjabi	1e-2	(0.1, 0.65)	0.06	(0.0, 1.68)	3e-3	(1.29, 4.26)	31
Gemini-2.5	GPT-4o	Hindi	1e-31	(0.49, 0.72)	9e-4	(0.26, 0.7)	9e-21	(1.81, 2.6)	448
Gemini-2.5	Gemma-3	Hindi	7e-34	(0.5, 0.68)	2e-12	(0.72, 1.22)	3e-26	(2.36, 3.25)	448
Gemini-2.5	GPT-4o	Odia	1e-4	(0.53, 1.24)	3e-5	(1.26, 3.03)	1e-2	(0.63, 3.66)	38
Gemini-2.5	Gemma-3	Odia	1e-4	(0.58, 1.47)	1e-4	(0.95, 2.61)	8e-4	(1.55, 4.61)	38
Gemini-2.5	GPT-4o	Marathi	1e-4	(0.51, 1.62)	3e-3	(0.7, 2.57)	2e-2	(0.32, 3.54)	37
Gemini-2.5	Gemma-3	Marathi	2e-5	(0.46, 0.97)	1e-3	(0.89, 2.76)	1e-4	(2.73, 5.7)	37

Table 15: **Statistical comparison of models on the VQA-Parallel set** across languages with p-values and confidence intervals for Long-QA, Short-QA, and Adversarial-QA.

Model A	Model B	Language	Long-QA		Short-QA		Adversarial-QA		n-items
			p-value	CI	p-value	CI	p-value	CI	
Gemini-2.5	GPT-4o	Malayalam	2e-14	(1.07, 1.6)	5e-9	(1.33, 2.53)	6e-12	(3.72, 5.43)	106
Gemini-2.5	Gemma-3	Malayalam	1e-8	(0.51, 1.07)	2e-7	(0.94, 2.05)	4e-7	(1.87, 3.66)	106
Gemini-2.5	GPT-4o	Gujarati	2e-15	(1.0, 1.52)	8e-9	(1.33, 2.41)	2e-11	(3.43, 5.1)	106
Gemini-2.5	Gemma-3	Gujarati	1e-11	(0.56, 0.98)	5e-4	(0.33, 1.33)	1e-7	(2.15, 3.95)	106
Gemini-2.5	GPT-4o	Telugu	1e-13	(0.97, 1.49)	1e-9	(1.13, 2.13)	1e-11	(3.42, 5.16)	106
Gemini-2.5	Gemma-3	Telugu	2e-8	(0.47, 1.07)	1e-3	(0.36, 1.3)	1e-5	(1.22, 3.03)	106
Gemini-2.5	GPT-4o	English	5e-12	(0.59, 1.02)	2e-2	(0.12, 1.13)	5e-13	(4.2, 5.88)	106
Gemini-2.5	Gemma-3	English	7e-11	(0.58, 0.94)	9e-5	(0.64, 1.85)	4e-11	(3.15, 4.79)	106

**Thank You**