HinTel-AlignBench: A Framework and Benchmark for Hindi–Telugu with English-Aligned Samples

Rishikant Chigrupaatii[1*] **Ponnada Sai Tulasi Kanishka**[1,*] **Lalit Chandra Routhu**[1,*] **Martin Patel**[1,*]
Sama Supratheek Reddy[1] **Divyam Gupta**[1] **Dasari Srikar**[1] **Krishna Teja Kuchimanchi**[1]
Rajiv Misra[1] **Rohun Tripathi**[2]

[1]Indian Institute of Technology Patna [2]Allen Institute for AI
{rishikant_2101cs66, ponnada_2101cs57, lalit_2101ai17, martin_2101cs43}@iitp.ac.in

Abstract

With nearly 1.5 billion people and more than 120 major languages, India represents one of the most diverse regions in the world. As multilingual VLMs gain prominence, robust evaluation methodologies are essential to drive progress toward equitable AI for low-resource languages. Current multilingual VLM evaluations suffer from four major limitations: reliance on unverified auto-translations, narrow task / domain coverage, limited sample sizes, and lack of cultural and natively sourced QA. To address these gaps, we present a scalable framework to evaluate VLMs in Indian languages and compare it with performance in English. Using the framework, we generate HinTel-AlignBench[1], a benchmark that draws from diverse sources in Hindi and Telugu with English aligned samples. Our contributions are threefold: (1) a semi-automated dataset creation framework combining back-translation, filtering, and human verification; (2) the most comprehensive vision-language benchmark for Hindi and Telugu, including adapted English datasets (VQAv2, RealWorldQA, CLEVR-Math) and native novel Indic datasets (JEE for STEM, VAANI for cultural grounding) with $\sim 4k$ QA per language; and (3) a detailed performance analysis of various SOTA open weight and closed source VLMs. We find a regression in performance for tasks in English vs. in Indian languages for 4 out of 5 tasks across all the models, with an average regression of 8.3 points in Hindi and 5.5 for Telugu. We categorize common failure modes to highlight concrete areas of improvement in multilingual multimodal understanding.

1 Introduction

India is home to an extraordinary diversity of languages, with 122 major languages and 1599 other languages[2].

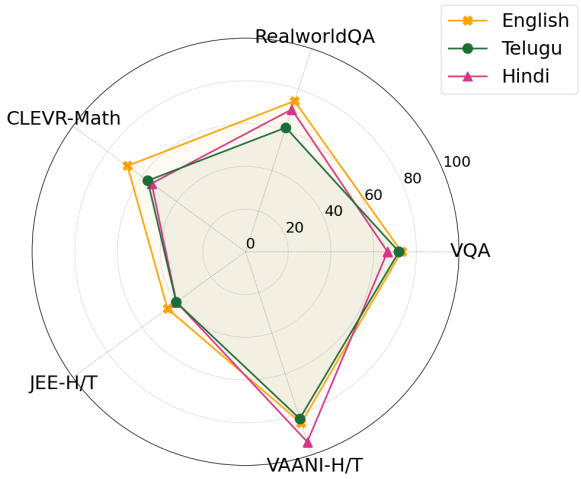


Figure 1: Average Performance of GPT-4.1 and Gemini-2.5-Flash models on English, Hindi and Telugu Languages on data-parallel visual question answering samples. Overall, performance regresses from English to Hindi by 8.3 points and regresses from English to Telugu by 5.5 points.

Despite this linguistic diversity, most state-of-the-art Multimodal Large Language Models (MLLMs) have historically been English-centric or focused on high-resource European languages. There has been a recent surge in multilingual MLLMs that support multiple languages, including frontier models (OpenAI, 2025; Google DeepMind, 2025; Meta Llama, 2025; Anthropic, 2025; Dash et al., 2025). However, there is no unified benchmark per Indian language to evaluate these models. Various benchmarks have been proposed for evaluating multilingual LLMs (Ahuja et al., 2023; Singh et al., 2024) but evaluating MLLMs fairly and comprehensively for Indian and other low-resource languages remains a significant challenge (Romero et al., 2024).

Current evaluation methodologies for multilingual vision-language models suffer from four fundamental limitations. First, many benchmarks rely on automatically translated datasets without human verification e.g., m-WildVision (Dash et al., 2025), inevitably introducing noise and artifacts. As noted in (Wu et al., 2025), auto-translated evaluation sets with and without back-translation often introduce biases, while relying entirely on manual curation is too slow for scalable

*Equal contribution
[1]https://rishikant24.github.io/indicvisionbench.github.io/
[2]https://en.wikipedia.org/wiki/Languages_of_India

benchmarking. Second, existing evaluations suffer from narrow scope, both in task diversity and domain coverage (Salazar et al., 2025). This myopic focus fails to assess whether models can handle different domains and tasks ranging from real-world visual question answering to mathematical reasoning and cultural awareness in the target language. Third, commonly used benchmarks have very limited sample sizes: xChat (Yue et al., 2025) and AyaVisionBench (Dash et al., 2025) have 50 and 135 QA per language, respectively. Fourth, current approaches like BharatBench's (Khan et al., 2024) use of translated LLaVABench (Liu et al., 2023), MM-Vet (Yu et al., 2024) and Pope (Li et al., 2023b) datasets fundamentally miss cultural grounding, linguistic nativity, and domain diversity, evaluating only surface-level translation quality rather than true multilingual competence. These limitations compound and decrease the reliability of benchmarks to measure progress in the support of Indian languages, ultimately slowing the development of equitable multilingual vision systems.

To address these issues, we introduce an efficient framework for generating high-quality, visually diverse multilingual evaluation sets. The framework employs a translation or text only LLM based QA generation stage followed by human review and verification stage. In the human review, each sample is assessed for semantic accuracy, linguistic style, and readability. Our semi-automated approach accelerates benchmark generation while maintaining rigorous linguistic fidelity through manual review. Our framework uses translation followed by verification, which is an easier annotation task and is significantly faster than generating question answer pairs from scratch, with processing speeds 5x faster for more than 79% of the samples as compared to generating question and answer pairs.

We follow this framework to create the most comprehensive vision language benchmark for Hindi and Telugu, featuring: (1) Adapted English benchmarks: Carefully adapted versions of real-world visual QA (VQAv2 (Goyal et al., 2017a)), practical reasoning (RealWorldQA (xAI, 2024a)), and visual mathematical reasoning (CLEVR-Math (Lindström and Abraham, 2022)) and (2) Native Indic evaluation sets: JEE-Vision: STEM competency testing via India's Joint Entrance Exam problems and VAANI (Team, 2025): Cultural and region-specific visual QA covering traditions, artifacts, and local contexts. Our selected subsets provide a multi-domain coverage and represent a broad assessment of multilingual vision language capabilities. For each sample, we also generate and manually verify the English sample. This gives us aligned samples in the target Indian language and English, which then enables an exact comparison of the model across the two languages. As far as we are aware, these are the most diverse collection of Hindi and Telugu VQA evaluation sets with English-aligned samples.

We comprehensively evaluate various open-weight and frontier closed-source models on our benchmark. There is a performance regression of 8.3 points from English to Hindi and 5.5 points from English to Telugu on average in evaluated models, Figure 1. Surprisingly, even for frontier SoTA models such as GPT 4.1, we find a 3.8-point drop in performance from English to Hindi and a 8.6-point drop from English to Telugu. Finally, the performance across Hindi and Telugu on aligned subsets is within 1 point, demonstrating a similar performance gap between English and both Indian languages.

By establishing a reliable evaluation methodology, we aim to enable the creation of robust evaluation datasets for low-resource languages worldwide. Our contributions are as follows.

- **A Framework**: A semi-automated, scalable system for generating high-quality multilingual vision language evaluation sets.
- **A Benchmark**: The large-scale, visually-diverse and publicly available evaluation dataset for Hindi and Telugu VLMs to support rigorous and consistent evaluation.
- **Comprehensive Evaluation**: An in-depth analysis of **SOTA vision LLMs**, highlighting regression in performance from English to Indic language in 4 out 5 domains and up to 15.6 points in the worst case.

2 Related Work

2.1 Multimodal Instruction-Tuned Models

Previous work has shown that large language models can be extended to vision tasks by instruction tuning with multimodal data (Liu et al., 2023; Dai et al., 2023; Li et al., 2023a; Zhu et al., 2024). Many open weight Vision-Language Models such as Gemma3 (Team et al., 2025), Qwen2.5VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), Molmo (Deitke et al., 2025), Llama3.2 Vision (Meta Llama, 2025) and closed source models such as GPT-4.1 (OpenAI, 2025), Claude 3.7 (Anthropic, 2025), Gemini 2.5 (Google DeepMind, 2025)) have since been released, all leveraging some form of multimodal instruction tuning. Many of these recent models are multilingual, but the evaluation of their multilingual abilities is limited and is not reported on a consistent and reliable multimodal multilingual benchmark (Dash et al., 2025; Yue et al., 2025; Rasheed et al., 2025; Alam et al., 2025).

2.2 Multilingual Vision-Language Benchmarks

There have been a number of recent benchmarks which evaluate multi-lingual abilities and cultural robustness of VLMs. CVQA (Romero et al., 2024) contains 10.4k visual QA pairs across 31 languages, demonstrating large performance gaps on low-resource languages. MTVQA (Tang et al., 2024) provides text-centric VQA in 9 languages with human annotations. xGQA (Pfeiffer et al., 2022) automatically translates the GQA (Hudson and Manning, 2019) dataset into 7 languages while

Benchmark	Indic Lang	QA Type	Image count	QA	Indic QA per lang	Human	Culturally Sourced
AyaVisionBench	**hin**	Chat	3.1k	3.1k	135	✗	✗
xMMMU	**hin**	MC	300	3k	291	✗	✗
xGQA	ben	OE	300	77.3k	9.7k	✗	✗
MTVQA	-	OE	8.8k	28.6k	-	✓	✗
M3Exam	-	MC	2.8k	12.3k	-	✓	✗
xChatBench	**hin**	Chat	400	400	50	✓	✗
Kaleidoscope	ben, **hin**, **tel**	MC	20.9k	20.9k	$\sim$ 800	✓	✗
XM100	ben, **hin**, **tel**	Caption	100	3.6k	100	✓	✓
MaRVL	tam	Caption	5.5k	5.7k	1.2k	✓	✓
MaXM	**hin**	OE	1.4k	2.1k	294	✓	✓
CVQA	ben, urd, tam, mar, **hin**, **tel**	MC	5.2k	10.4k	$\sim$ 300	✓	✓
Ours	**hin**, **tel**	OE, MC	5.1k	13.5k	$\sim 4k$	✓	✓

Table 1: Comparison of existing multilingual Visual Question Answering (VQA) benchmarks with ours.**"QA Type"** denotes the question-answering format (Chat = Multimodal chat; OE = open-ended VQA;MC = multiple-choice VQA; Caption = captioning/multi-image captioning); **"QA"** denotes the total question-answer pairs; **"Indic QA per lang"** denotes the question-answer pairs per Indic language; **"Human"** indicates human verified/annotated data; **"Culturally Sourced"** indicates culturally-sourced data. xChatBench, xMMMU and XM100 are part of PangeaBench (Yue et al., 2025). Ours (bottom row) is the only VQA dataset with human-verified annotations, diverse culturally sourced samples and a significant sample size per Indic language.

MaxM (Changpinyo et al., 2023) uses Machine Translation plus lightweight post-editing for 7 languages, with less than 300 samples per language. MarVL (Liu et al., 2021) focuses on culturally sourced reasoning by evaluating whether a caption about an image pair is true or false, but is not a Visual Question Answering dataset. Concurrently with our work, Kaleidoscope (Salazar et al., 2025) is a multi-lingual multimodal benchmark with 800 VQA MCQs per Indic language sourced from regional examinations. However, Kaleidoscope has a narrow scope as it focuses only on exam-based questions and does not evaluate real-world understanding and practical reasoning. Multimodal benchmarks released along with their corresponding models include PangeaBench (Yue et al., 2025) (14 datasets in 47 languages), and AyaVisionBench (Dash et al., 2025) (9 tasks in 23 languages). However, these sets commonly have few manually verified samples per language, making the results less reliable for a particular target language. Table 1 provides an overview and compares them with our benchmark. A significant portion of our images per language (>1K) and our QA (>1K) are using culturally sourced, which is 5 to 20x larger than the culturally relevant subsets of previous datasets in this field.

2.3 Indic and Domain-Specific Datasets

Although there are dozens of English VQA resources (VQAv2 (Goyal et al., 2017a), CLEVR-Math (Lindström and Abraham, 2022), RealWorldQA (XAI.org, 2024)), there is a need for benchmarks in Indian languages. Previous works include (Singh et al., 2024; Arora et al., 2023). IndicGenBench (Singh et al., 2024) consists of 5 diverse tasks in 29 Indic languages but does not include multimodality. JEEBench (Arora et al., 2023) consists of questions from the highly competitive IIT JEE-Advanced exam, but does not include images-related questions. In contrast, our JEEset only includes questions with images. Our work fills the need for an accurate and diverse multimodal benchmark by providing $\sim$4k Visual Question Answering each in Hindi and Telugu.

3 Datasets

3.1 Sources

Our goal is to create a diverse and robust benchmark for Hindi and Telugu by combining samples translated from English VQA datasets with two new India-centric datasets, Figure 3. Concretely, we cover the following five broad tasks. (1) **Real-World Understanding:** We sample 1000 examples from the VQAv2 dataset (Goyal et al., 2017b) and translate it. VQAv2 is a multimodal dataset designed for visual question answering. (2) **Practical Visual Reasoning:** We extend the RealWorldQAdataset (765 QA) using the translation pipeline. RealWorldQA (xAI, 2024a,b) is a benchmark designed to evaluate real-world spatial understanding. (3) **Visual Mathematical Reasoning:** We sample 1000 samples from the CLEVR-Math dataset and translate it. CLEVR-Math (Lindström and Abraham, 2022) is a multimodal dataset designed for basic arithmetic and spatial reasoning. (4) **STEM Competency testing (JEE-Vision):** The Joint Entrance Exam (JEE) is a high-school nation wide exam conducted yearly in multiple languages, and the exam papers are distributed for the general public (JEE Adv and Mains). (5) **Cultural and Region-specific QA:** From the VAANI (Team, 2025) image-caption corpus, we select images sourced from Hindi and Telugu-speaking regions and use GPT-4.1 to generate multiple-choice questions in each language from the original captions to form the VAANI-H and VAANI-T datasets respectively. We share the prompt to generate

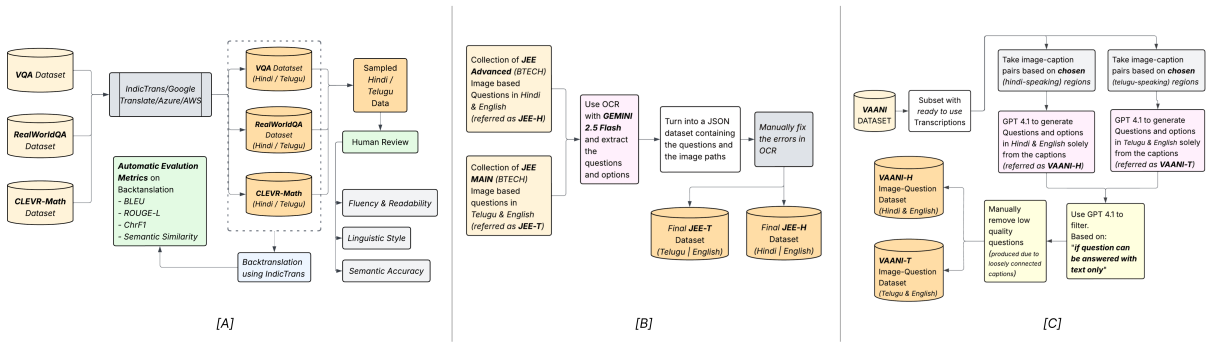


Figure 2: Dataset generation pipeline for **[A]** VQAv2, RealWorldQA & CLEVR-Math **[B]** VAANI-H & VAANI-T **[C]** JEE-H & JEE-T. We use back translation and text only LLMs to reduce human involvement in QA generation.

this QA in the appendix.

We include existing datasets by translating English sets as opposed to creating completely new datasets from scratch based on the motivations in (Singh et al., 2024). Translation-based extension of existing benchmark results in multi-way parallel data, allowing researchers to attribute performance due to task knowledge vs. language understanding, and measure cross-lingual generalization. Using the translation pipeline, we are able to leverage the quality control that went into designing the initial benchmarks.

Language	VQAv2	RealWorldQA	CLEVR-Math	JEE-H	JEE-T	VAANI-H	VAANI-T
Hindi	1,000	765	1,000	192	-	945	-
Telugu	1,000	765	1,000	-	325	-	1,020
English	1,000	765	1,000	192	325	945	1,020
Total	3,000	2,295	3,000	384	650	1,890	2,040

Table 2: Number of QA pairs per task per language in HinTel-AlignBench. The samples used in VQAv2, RealWorldQAand CLEVR-Math are the same across all languages. For the VAANI and JEEsets, the samples are aligned across the Indian language and English.

The distribution of the number of QA pairs per language per subset is reported in Table 2.

3.2 Dataset Generation and translation

Translation Pipeline

Figure 2 [A] illustrates our pipeline for translating VQAv2, RealWorldQAand CLEVR-Math into Hindi and Telugu. For each language, we review the translation with 50 diverse samples using IndicTrans (Gala et al., 2023), Google Translate, Azure and AWS. Based on the performance, we selected the best translation tool per language. For Telugu, we used AWS Translate and for Hindi we used Azure. To ensure an accurate evaluation benchmark, we manually review each sample. During manual review, the annotator verifies and makes minor edits based on *Semantic Accuracy*, *Linguistic Style*, and *Fluency & Readability*, as performed in (KJ et al., 2025). Previous works (KJ et al., 2025) have used back-translation to the source language in combination with automated evaluation metrics (Papineni et al., 2002; Popović, 2015) to verify a portion of the samples or to select samples to verify. However, we found that the subsets selected using back-translation were biased towards high confidence mistakes of the translation system, thus biasing the dataset distribution.

Using the translation pipeline to generate the initial sample reduces the annotator burden per sample from generating a QA pair to only correcting a translated QA pair. Additionally, it leads to a consistent formatting of the translations. In a representative subset such as VQAv2, 79% of the samples that were reviewed were accepted during manual review without any change. Within the subset that was changed, about 42% are minor changes (such as the tense of a word) and only the remaining 58% need deletion and/or addition of new words during the manual verification stage. Based on our early evaluation on a subset, processing all samples, except those that need addition and deletion, is 5× faster than generating samples from scratch and equally accurate. Finally, since our translation pipeline reduces the burden and in order to maximize the size of the dataset given a fixed annotator budget, we use 1 annotator per sample. All verifications were carried out by co-authors who are native speakers or possess fluent proficiency in Hindi and Telugu, with sufficient linguistic and contextual understanding to perform high-quality annotations.

VAANI-H and VAANI-T Dataset Generation

Figure 2 [C] details the creation process for our VAANI-H and VAANI-T image-question datasets. From the original VAANI set, we extract a subset which have text transcriptions. There are no overlapping set that has both Hindi and Telugu transcriptions and hence we create separate sets of both languages. For each of the languages, we use text only GPT 4.1 to generate the questions and options in both the languages from their respective captions, resulting in the VAANI-T and VAANI-H for Telugu and Hindi Image-Question datasets respectively. Both datasets then undergo a refinement process: First, text-only GPT 4.1 is used to filter out the question answer pairs that can be answered using only the text without the requirement of the accompanying image. Second, we employ a human verification process to filter out low-quality questions. These

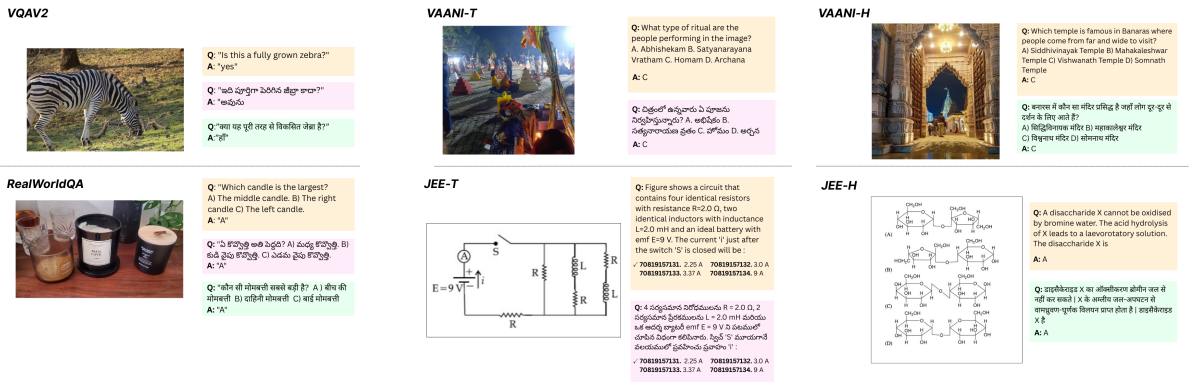


Figure 3: Qualitative Examples for different domains in our dataset. More images are shown in the appendix

were necessary because some of the captions in the VAANI set are relevant to the images and will not be needed for a dataset in which the captions always align with the image.

JEE-H and JEE-T Dataset Creation

We present JEE-Vision, a novel vision-language dataset constructed from India's Joint Entrance Examination (JEE)[3][4], leveraging its unique linguistic distribution (JEE Mains: 13 languages; Advanced: English/Hindi) to source authentic STEM problems authored in the target languages by subject-matter experts. This helps us avoiding any manual steps needed to filter out the translation artifacts that compromise existing benchmarks. Unlike text-only evaluations such as (Arora et al., 2023), we specifically curate diagram-dependent problems (circuit schematics, chemical structures etc.) creating the first benchmark to evaluate: (1) non-translated multilingual STEM reasoning, and (2) joint understanding of technical visuals and linguistic content in native educational contexts. Figure 2 [B] outlines the procedure the JEE datasets. The JEE-H dataset consists of questions from *JEE-Advanced-B.TECH* examination. The dataset has 192 questions (with accompanying images) spanning Math, Physics, Chemistry in both English and Hindi. However, the released questions for the Advanced set are only in English and Hindi. Hence, for the JEE-T dataset, we collect questions from *JEE-Main-B.TECH* examination. The dataset has 325 questions that have accompanying image across Math, Physics, Chemistry in both English and Telugu. For both JEE sets, we extract the questions and options using OCR with Gemini-2.5-Flash (Google DeepMind, 2025). During initial dataset generation, we found some OCR issues in the samples. Hence, all the questions undergo a manual verification to fix any OCR errors. While this set currently requires a manual verification step, as OCR capabilities improve, this manual verification process will become unnecessary.

[3]JEE Advanced:https://jeeadv.ac.in/archive.html
[4]JEE Mains:https://www.nta.ac.in/Downloads

4 Experimental Setup

4.1 Selected Models

We evaluate the performance of current open-weight and frontier proprietary Multimodal Large Language Models (MLLMs) on our benchmark. Some previous works report multilingual results in a target language using models that may not be trained to support that specific language (Dash et al., 2025; Rasheed et al., 2025). In contrast, we include model evaluations for a target Indian language only if that model claims prior proficiency in the target language. This is to capture an accurate representation of the current multilingual multimodal models.

For Hindi, we report results on open-weight models including Google's Gemma3 series (Gemma3-4B, 12B, and 27B) (Team et al., 2025), Chitrarth-8B (Khan et al., 2024), an Indic-focused vision-language model designed for multiple regional languages including both Hindi and Telugu; Aya-8B (Dash et al., 2025), Alibaba's Qwen2.5VL-7B (Bai et al., 2025); and Meta's LLaMA 3.2 Vision 11B (Meta Llama, 2025). These models were chosen for their class leading capabilities and explicit support for Hindi. From the set of proprietary models, we report on the Gemini-2.5 Flash variants (Google DeepMind, 2025), which exhibits strong support for Telugu and Hindi, with notable performance in OCR, and OpenAI's GPT-4.1 (OpenAI, 2025), a state-of-the-art model with documented support for both Hindi and Telugu.

For Telugu, we find that very few VLMs support the language. We report on the Gemini series (Gemini 1.5 Flash, 2.0 Flash, and 2.5 Flash variants) (Google DeepMind, 2025), OpenAI's GPT-4.1 (OpenAI, 2025) and Chitrarth-8B (Khan et al., 2024), all of which support Telugu.

4.2 Evaluation Metrics

The RealWorldQA, VAANI-H/T subsets have only multiple-choice questions and JEE-T has multiple-choice or integer-answer questions. We use accuracy as the evaluation metric for all of these sets. We extract

Model	VQAv2		RealWorldQA		CLEVR-Math		JEE-T		VAANI-T		Ours-T	
	Tel	En	Tel	En	Tel	En	Tel	En	Tel	En	Tel	En
GPT 4.1	68.70	72.00	61.05	**75.29**	46.60	65.00	34.36	40.92	81.57	**82.45**	58.46	67.13
Gemini 2.5 Flash	75.10	74.10	**61.18**	73.20	**66.70**	**71.80**	**45.90**	**54.15**	**82.75**	80.78	**66.33**	**70.81**
Gem 2.0 Flash	70.20	74.20	60.92	69.67	43.60	53.50	42.15	53.23	80.49	79.61	59.47	66.04
Gem 1.5 Flash	68.50	74.40	60.00	67.19	37.40	46.70	29.85	39.38	76.27	79.61	54.40	61.46
Chitrarth	**76.00**	**78.50**	53.59	52.55	53.90	56.90	20.00	18.15	81.57	82.45	57.01	57.71
Model Mean	71.10	74.64	59.35	67.58	49.64	58.78	34.85	41.17	80.53	80.98	59.13	64.63

Table 3: Results (in %) for Telugu (Tel) and English (En). **Bold** indicates the best and underline indicates the second best. There is a performance regression for all sets for most models from English to Telugu, with the primary exception being VAANI-T. Ours-T includes the VAANI-T, JEE-T subsets along with VQAv2, CLEVR-Math and RealWorldQA

Model	VQAv2		RealWorldQA		CLEVR-Math		JEE-H		VAANI-H		Ours-H	
	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En
GPT-4.1	**68.00**	72.00	**70.59**	**75.29**	48.10	65.00	23.18	23.05	93.33	86.88	60.64	64.44
Gemini 2.5 Flash	65.00	74.10	69.54	73.20	**60.70**	71.80	**56.90**	**62.89**	**93.86**	**87.19**	**69.20**	**73.84**
Chitrarth	66.00	**78.50**	52.94	52.55	57.20	56.90	11.72	13.93	84.23	80.14	54.42	56.40
Qwen2.5VL-7B	37.20	74.30	51.11	68.10	29.10	**98.80**	17.84	20.70	81.79	84.76	43.41	69.33
Aya-8B	36.30	47.30	55.42	58.82	46.20	61.40	9.63	16.02	82.22	80.42	45.95	52.79
LLaMA 3.2 11B	35.90	59.80	35.68	61.57	18.90	35.60	14.32	14.45	77.67	83.28	36.49	50.94
Gemma3-27B	64.10	65.50	54.38	61.04	43.50	53.70	19.66	17.58	87.41	82.01	53.81	55.97
Gemma3-12B	63.00	65.70	53.98	58.69	40.20	46.80	14.84	16.80	85.50	82.33	51.50	54.87
Gemma3-4B	55.00	58.20	43.27	50.19	33.20	39.60	14.32	17.19	80.64	77.78	45.29	48.59
Model Mean	53.74	66.26	53.99	62.49	43.01	58.62	20.01	22.29	85.85	83.76	51.19	58.49

Table 4: Results (in %) for Hindi (Hi) and English (En). **Bold** indicates the best and underline indicates the second best. There is a performance regression for all sets for most models from English to Hindi, with the primary exception being VAANI-H. Ours-H includes the VAANI-H, JEE-H subsets along with VQAv2, CLEVR-Math and RealWorldQA

the answers using regex-based parsing, and report the overall accuracy across all the questions.

Hybrid Evaluation for VQA and CLEVR-Math: For VQAv2 and CLEVR-Math subsets, the answers are either a single word or short phrases. We adopt a hybrid evaluation strategy. We first evaluate a sample using exact match. Our exact match evaluation is built using the official VQA evaluation script (Goyal et al., 2017a), with the functionalities also extended to Hindi and Telugu. While exact match is strict and interpretable, it may penalize correct answers with minor surface-level variations (e.g., "yes" and "yes, it is", synonyms, etc.). If exact match fails for a sample, we evaluate that sample using "gpt-4.1-2025-04-14" (OpenAI, 2025) as descr in (**?**). This two-step approach enables both high precision and flexibility, especially in cases in which answers may vary in form but not meaning. Its impact on scores is discussed in the appendix.

JEE-H Evaluation: The JEE-H Dataset has single correct MCQs, multiple correct MCQs, and numeric-type questions. We extend the scoring process used in (Arora et al., 2023). However, instead of the manual step they use, we replace with regex-based parsing to extract answers, and rule based processing to score each question. See appendix for details.

5 Results and Analysis

5.1 English-Telugu Comparison

The results of our evaluated models for Telugu and English are reported in Table 3. In all the tasks, the average performance of the models in English is higher than in Telugu, though the margin on the VAANI-T subset is marginal. Averaging across the models and the tasks, there is a gap of 5.5 points between the English and Telugu subsets.

Results on VAANI-T go against the trend of the other subsets and we explore this in Section 5.4. Overall, the best performing model is Gemini 2.5 Flash (Google DeepMind, 2025) on both Telugu and English. A notable model was Chitrarth (Khan et al., 2024), which has been trained on VQAv2 in multiple languages and hence the leader on the VQAv2 subset in English and Telugu. GPT-4.1 shows leading on the RealWorldQA subset for English and Hindi but much poorer performance on the Telugu version. This data point demonstrates how

Method	VQAv2		RealWorldQA		JEE-H		CLEVR-Math		VAANI-H		Ours-H	
	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En
Standard	**64.10**	**65.50**	54.38	61.04	19.66	17.58	43.50	53.70	**87.41**	82.01	53.81	55.97
CoT	60.50	62.70	**61.96**	**64.31**	**21.88**	**31.38**	**44.3**	**57.1**	84.34	**83.28**	**54.60**	**59.75**
Caption-Only	56.40	64.1	44.05	54.77	–	–	36.4	32.9	80.85	77.14	–	–

Table 5: Results on Gemma-27B across multiple datasets using difference inference methods. **Bold** indicates the best performance, underline indicates the second best. CoT significantly improves results on English but has a much smaller improvement when using Hindi.

models can have a high performance in some languages but fail to generalize and sheds light on the need for thorough evaluation of all the target languages.

5.2 English-Hindi Comparison

The results of the evaluated models for Hindi and English are in Table 4. For 4 out of 5 tasks, the average performance in English is higher than in Hindi. This demonstrates the need to improve the capabilities of our multilingual LLMs. Averaging across the models and the tasks, there is a gap of 8.3 points between the English and Hindi subsets.

Going against the trend, results on VAANI-H are higher in Hindi than in English, further discussed in Section 5.4. Among the proprietary and open-weight models, the open source models lag the proprietary ones in both Hindi and English in most cases. Gemini 2.5 Flash (Google DeepMind, 2025) is the overall best performing model on both languages in this paired dataset. Among the open-weight models, the largest gap is seen on Qwen2.5VL-7B (Bai et al., 2025), dropping from 25.92 points from English and Hindi.

5.3 Telugu-Hindi Comparison

We have aligned samples on the VQAv2, CLEVR-Mathand RealWorldQAsets for Hindi, Telugu and English along with evaluations with GPT-4.1, Gemini 2.5 Flash and Chitrarth for all three languages. On this subset of datasets and models, the average performance on Hindi, Telugu and English is 61.51, 62.53 and 68.6 respectively. This comparison demonstrates that the performance regresses similarly from English to both Indian languages and underlines a systemic issues when using the current frontier model with Indian languages.

5.4 Task-Specific Results

The average gaps across models for each of the tasks is visualized in Figure 4. This graph motivates the need for overall, multi-dimensional improvement needed in the capabilities of multilingual LLMs to catch up to the performance in English. CLEVR-Math and RealWorldQA have the largest deltas in performance between the English and the Indic Language sets. In comparison, the smallest delta in performance is on the VAANI dataset. On VAANI-T, average performance is marginally worse than English but on VAANI-H, the average performance is 2.09 points higher on the Hindi

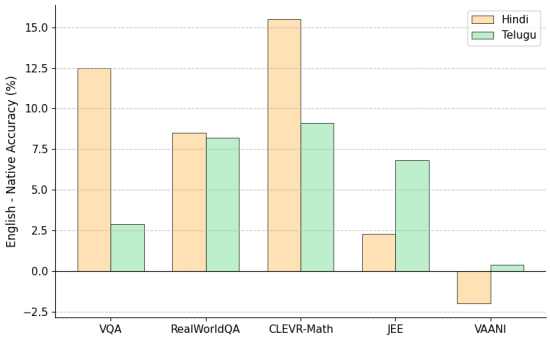


Figure 4: Average performance regression from English to Indic Language per domain across all models. In all but one scenario, there is a regression in performance from English to the target Indian language

subset. On reviewing the QA that have the correct prediction for Indic Languages and incorrect predictions for English, we found two possible reasons. Firstly, we found samples where the English QA generated did not completely capture the meaning of the options or translated the options into English for OCR tasks where the text in the Indic language was visible in the image. This pattern made the questions harder for the model in English sets. Secondly, we find that text-only LLMs are poor at generating distractors when generating MCQA from VAANI captions. Thus, the models may guess the correct answer by exploiting statistical patterns, which inflates metrics. In future work, we'll explore using LLM based quality control to further align questions across languages and using Multi-Binary Accuracy (Cai et al., 2024) for VAANI subsets.

5.5 Inference Techniques and Baselines

We ablate prompting and other baselines using the Gemma-27B model on the Hindi subsets. We report comparisons for Direct Inference (Standard Prompting), Chain-of-Thought (Wei et al., 2022), and a Caption-Only (Deitke et al., 2025) baseline, Table 5.

Standard prompting provides a strong baseline performance across all datasets, especially on VAANI and VQA, indicating that the model benefits significantly from both modalities in relatively straightforward recognition tasks. In contrast, Chain-of-Thought (CoT) prompting notably improves performance on reasoning-heavy datasets. On JEE-H, CoT boosts performance

in English and Hindi by 13.8 and 2.22 points respectively. Similar gains are observed in CLEVR-Math, where CoT surpasses the standard setup in both languages. These improvements highlight the effectiveness of step-by-step reasoning prompts in tasks that require deeper cognitive processing. However, CoT does not benefit all tasks, with marginal gains or regression on VAANI and VQAv2, where reasoning depth is less critical. CoT may be introducing unnecessary verbosity or distraction when the task primarily relies on perception. Overall, CoT benefits the English subset by 3.78 points in comparison to the benefit on Hindi at 0.79 points. This indicates a possible bias in the training data, with possibly more training done on English CoT data.

Next, to explore the need for visual embedding for QA, we evaluate a caption-only baseline. Here, the Gemma-27B model is asked to first generate a generate a rich textual summary of the image. Then the generated caption and the original question are given to the model without the original image. This approach is inferior to the direct image-based inputs in all tasks.

5.6 Failure Analysis of Visual Language Models

We categorized the failures of GPT 4.1 on VAANI-T subset for both English and Telugu. We analyze results on VAANI-T as it sources images and QA from the Indian context and failures on this set point to common reasons for errors for current multimodal multilingual LLMs. We found 4 major categories of errors that appeared in both English and Telugu questions. They are: (1) Missing Indian context: Failed predictions in this category were due to missing knowledge specific to India needed to answer the question. (2) Visual Grounding error: Failures in this category were due to the model's inability to associate objects in the picture with the question or options. (3) Visual Perception Failure: Errors in this category are due to the model failing to interpret the content of the image correctly. (4) Failure to Ground in Indian Context: Samples in this category accurately identified visual elements but they failed to analyze them in the appropriate socio-cultural context. Table 6 reports the percentages of the different categories on this dataset.

Error Category	English	Telugu
Lack of Knowledge about India	17%	16%
Visual Grounding Error	49%	50%
Visual Perception Failure	19%	18%
Failure to Ground in Indian Culture	15%	16%

Table 6: Error category distribution for GPT 4.1 on VAANI-T for English and Telugu. Percentages indicate the proportion of errors falling into each category.

5.7 Discussion and Future Work

Using our framework, benchmarks can be created in a number of domains. We aim to extend the framework as follows. First, the proprietary models achieve high scores on the VAANI-H/T. We hypothesize that the text only LLMs we use for generating QA from captions do not design good distractors. Thus, the VLMs may guess the correct answer by exploiting statistical patterns. A future work is using Multi-Binary Accuracy (Cai et al., 2024) for VAANI subsets. Second, the JEE-Advanced examination data source only contains 2 image-based Mathematics questions and we plan to include the questions in JEE-Mains. Our current scope is limited to images but can be extended to video understanding similar to (**?**). Finally, there are 122 major Indian languages, and we cover only Hindi and Telugu. We hope this benchmark gets extended to all other Indic languages with contributions from native speakers from those languages.

6 Conclusion

This paper introduces HinTel-AlignBench, a framework for developing benchmarks to evaluate multimodal large language models in Hindi and Telugu, addressing critical gaps in existing multilingual evaluations. We combined semi-automated dataset creation with rigorous human verification and sourced culturally grounded native datasets to assess diverse capabilities. Evaluations of state-of-the-art VLMs reveal significant performance regressions in Indic languages compared to English, emphasizing the need for targeted improvements in multilingual visual understanding. The regression highlights the need for rigorous evaluation for all Indic languages, beyond Hindi and Telugu.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *Preprint*, arXiv:2303.12528.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, and 1 others. 2025. Behind maya: Building a multilingual vision language model. *arXiv preprint arXiv:2505.08910*.

Anthropic. 2025. Claude 3.7 sonnet: The first hybrid reasoning model. Anthropic Company Website. Available at: https://www.anthropic.com/news/claude-3-7-sonnet.

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards multilingual visual question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, arXiv:2505.08751.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. Google DeepMind Blog. Released on March 25, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaharukh Khan, Ayush Tarun, Abhinav Ravi, Ali Faraz, Akshat Patidar, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Chandra Khatri, and Shubham Agarwal. 2024. Chitrarth: Bridging vision and language for a billion people. In *NeurIPS Multimodal Algorithmic Reasoning*.

Sankalp KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding. *arXiv preprint arXiv:2501.15747*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual, and mathematical reasoning. *arXiv preprint*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Meta Llama. 2025. Llama 3.2-vision: Instruction-tuned image reasoning generative models. Model release by Meta. Available at: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision.

OpenAI. 2025. Gpt-4.1: A new series of gpt models with major improvements on coding, instruction following, and long context. OpenAI Company Website. Released on April 14, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A polyglot large multimodal model for 5b people. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 56 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, and 25 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *Preprint*, arXiv:2504.07072.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, and 1 others. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

VAANI Team. 2025. Vaani: Capturing the language landscape for an inclusive digital india (phase 1). https://vaani.iisc.ac.in/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*, arXiv:2504.15521.

xAI. 2024a. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v.

xAI. 2024b. Realworldqa dataset. Hugging Face Dataset Repository.

XAI.org. 2024. Realworldqa: A benchmark for real-world spatial understanding capabilities of multimodal ai models. https://huggingface.co/datasets/xai-org/RealworldQA. RealWorldQA is a benchmark designed for real-world understanding with 765 multiple-choice questions requiring recognition of details in high-resolution images.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*. JMLR.org.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Jinguo Zhu and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for

open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Qualitative Examples

Refer to Fig. 5, with examples from each of the datasets, i.e VQAv2, RealWorldQA, CLEVR-Math, VAANI-H, JEE-T, VAANI-T, JEE-H

A.2 Compute and Costs

We primarily ran all open-source models on H100 and A100 GPUs, rented via the Akash Console Network[5], incurring approximately 100 USD in compute costs. For proprietary models, we used APIs, leveraging Gemini's free tier and spending around 60 USD on OpenAI's APIs. The total expenditure amounts to approximately 160 USD.

A.3 Prompts

JEE-H COT Inference

```
eng_prompt = """You are an expert at solving
physics, chemistry and math questions that
involve both text and images. Analyze the text
and the associated image to answer the question
above.
First, think step-by-step and provide your
detailed reasoning. Explain how you interpret the
text and the image, the principles or formulas
you are using, and the intermediate calculations
or logical steps you take to arrive at the
solution.
After your reasoning, provide the final answer on
a new line.
Your entire response, including the reasoning and
the final answer, MUST end with the following
line exactly:
{ answer: }"""
```

JEE-H/T Standard Inference

```
eng_prompt = "You are an expert at solving
physics, chemistry and math questions that
involve both text and images.
Analyze the text and the associated image to
answer the question above. Provide only the final
answer, without any explanations or intermediate
steps. Your response MUST end with the following
line: { answer: }"
```

VQA LLM Eval

```
System_Prompt = "You are an expert judge
evaluating semantic similarity between two
answers.
Respond only with '1' for similar or '0' for not
similar."

User_Prompt = "Are the following two answers
semantically similar?

Answer 1: {a}
Answer 2: {b}

Respond with only '1' if they are similar in
meaning, or '0' if they are not similar."
```

[5]https://akash.network/

CLEVR LLM Eval

```
System_Prompt = "You are an expert judge
evaluating semantic similarity between two
answers.
Respond only with '1' if they are semantically
similar, or '0' if they are not."

User_Prompt = "Are the following two {lang}
answers semantically similar?

Answer 1: {ans1}
Answer 2: {ans2}

Ignore minor differences in phrasing. Respond
with '1' if they express the same meaning, or '0'
otherwise."
```

A.4 Inference Strategies

This appendix details the various inference techniques employed for evaluating the multimodal models, along with the specific prompt setups used for English and Hindi. Below are setups in English.

Caption-Only Baseline

This technique involves a two-step process:
1) Dense Caption Generation: The model first generates a detailed, factual description (dense caption) of the provided image.
2) Question Answering with Caption: The model then answers the original question using *only* the generated dense caption as context, without access to the original image.

Prompt for Step 1 (Dense Caption Generation):

```
eng_prompt = Don't forget these rules:
1. Be Direct and Concise: Provide straightforward
descriptions without adding interpretative or
speculative elements.
2. Use Segmented Details: Break down details
about different elements of an image into
distinct sentences, focusing on one aspect at a
time.
3. Maintain a Descriptive Focus: Prioritize
purely visible elements of the image, avoiding
conclusions or inferences.
4. Follow a Logical Structure: Begin with the
central figure or subject and expand outward,
detailing its appearance before addressing the
surrounding setting.
5. Avoid Juxtaposition: Do not use comparison or
contrast language; keep the description purely
factual.
6. Incorporate Specificity: Mention age, gender,
race, and specific brands or notable features
when present, and clearly identify the medium if
it's discernible.
When writing descriptions, prioritize clarity
and direct observation over embellishment or
interpretation. Write a detailed description of
this image, do not forget about the texts on it
if they exist. Also, do not forget to mention the
type/style of the image. No bullet points. Start
with the words, "This image displays:"
```

Prompt for Step 2 (Question Answering with Caption):

```
Context: {generated_caption}
Question: {question_text}
Answer:
```

Chain of Thought (CoT) Prompting

This technique guides the model to generate a step-by-step thought process (rationale) before arriving at the final answer. This is intended to improve reasoning capabilities for complex questions.

The image is provided along with language-specific system and user prompts.

English System Prompt:

```
When provided with an image and a question,
generate a rationale first and then derive an
answer.
Your rationale should include detailed visual
elements in order to derive the answer.
```

English User Prompt:

```
Answer the question with following
instruction:
1. Generate a rationale first and then derive
an answer.
2. For your final answer, provide a Correct
Option Number Only.
Question:
{question}
# Output Format #
<rationale>
### Answer: <your answer>
```

Standard (Direct Inference)

The model is provided with both the image and the question and is expected to generate a direct answer without explicit intermediate reasoning steps requested in the prompt.

The image and the question are provided to the model. The prompt typically instructs the model to answer directly, often in a specific format (e.g., option number for multiple-choice questions).

Example input (for a multiple-choice question):

```
Image: [Image Data]
Question: {question_text_with_options}
Instruction: Provide the Correct Option Number
Only.
Answer:
```

(Note: This often shares the final answer formatting instruction with CoT, but without the explicit rationale generation step.)

JEE-H Scoring Rules

- For single correct Multi-Choice Questions (MCQs) and integer-type questions, a binary score was assigned: 1 if the model answered with the correct option/integer and 0 otherwise.
- For multi-correct MCQs, a score of 1 is given when the generated response consists of all and only the correct options. If any of the wrong option is chosen, the response is given a score of 0. If the response contains no incorrect option and some of the correct options, a score of 0.25 is given for each of the correct options.
- For numeric-type questions, answers in the range of ±0.01 with the gold answer are given a score of 1, and 0 otherwise.
- In Chain-of-Thought (CoT) inferencing, if a Hindi response included reasoning in English, 0 was given irrespective of the answer.
- In standard zero-shot inference, where the model was explicitly instructed to return only a direct answer, responses containing any explanation or reasoning were also scored 0, irrespective of the answer.

Impact of LLM Judge on Evaluation Scores

As detailed in Section **??**, our hybrid evaluation strategy for CLEVR-Math and VQA incorporates an LLM judge to re-evaluate answers initially marked incorrect by the exact match protocol. This approach revealed a notable trend: the score increase attributed to the LLM judge was considerably more for the responses in Hindi and Telugu as compared to English. For instance, when evaluating the VQA dataset, the average scores for English responses increased by $\approx 4\%$ after LLM judging. In contrast, Hindi scores on the same dataset increased by $\approx 12\%$ after the initial exact match scores. This disparity suggests that the models demonstrate a greater proficiency in adhering to strict answer formatting conventions in English, while responses in Hindi and Telugu, though often semantically correct, are more frequently penalized by the exact match criteria due to variations in phrasing or formatting. The LLM judge effectively mitigates this by recognizing a broader range of correct expressions.

A.5 Licenses

The datasets used in our benchmark are released under the following licenses:

- VQAv2, VAANI and ClevrMath are released under Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- RealWorldQA is released under Creative Commons Attribution No Derivatives 4.0.
- JEE-Vision is built from publicly available JEE questions. We will be releasing links under CC BY-NC 4.0 license along with the QA pair. The exam papers are distributed for the general public (https://www.jeeadv.ac.in/archive.html) used by various publishers and institutes for commercial purposes already. When collecting this data, we followed the steps by previous works on JEE Bench (Arora et al., 2023).

VQAV2

Q: "Is this a fully grown zebra?"
A: "yes"

Q: "ఇది పూర్తిగా పెరిగిన జీబ్రా కాదా?"
A: "అవును

Q:"क्या यह पूरी तरह से विकसित जेब्रा है?"
A:"हाँ"

Q: How aggressive is the cat?
A: Not at all

Q: పిల్లి ఎంత దూకుడుగా ఉంది?
A: అస్సలు కాదు

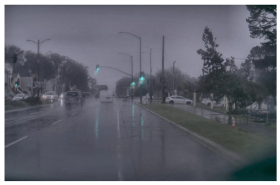
Q: बिल्ली कितनी आक्रामक है?
A: बिलकुल नहीं

RealWorldQA

Q: "Which candle is the largest? A) The middle candle. B) The right candle C) The left candle.
A: "A"

Q: "ఏ కొవ్వొత్తి అతి పెద్దది? A) మధ్య కొవ్వొత్తి. B) కుడి వైపు కొవ్వొత్తి. C) ఎడమ వైపు కొవ్వొత్తి.
A: "A"

Q: "कौन सी मोमबत्ती सबसे बड़ी है? A) बीच की मोमबत्ती B) दाहिनी मोमबत्ती C) बाईं मोमबत्ती
A: "A"

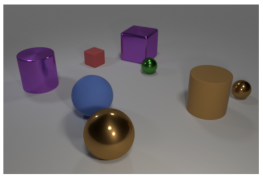


Q: What color is the traffic light?
A: Green

Q: ట్రాఫిక్ లైట్ ఏ రంగు?
A: ఆకుపచ్చ

Q: ट्रैफ़िक लाइट किस रंग की है?
A: हरा

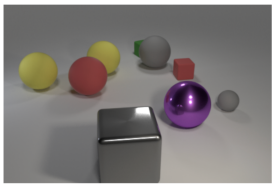
Clevr-Math



Q: Do the big blue object and the purple thing behind the red block have the same material?
A: No

Q:పెద్ద నీలం వస్తువు మరియు రెడ్ బ్లాక్ వెనుక ఉన్న ఊదా రంగు వస్తువు ఒకే పదార్థాన్ని కలిగి ఉన్నాయా?
A: లేదు

Q: क्या लाल ब्लॉक के पीछे बड़ी नीली वस्तु और बैंगनी चीज़ एक ही सामग्री की हैं?
A: नहीं



Q: How many yellow things are the same size as the red ball?
A: 2

Q:ఎరుపు బంతి మాదిరిగానే ఎన్ని పసుపు వస్తువులు ఉంటాయి?
A: 2

Q: कितनी पीली चीजें लाल गेंद के समान आकार की हैं?
A: 2

VAANI-H

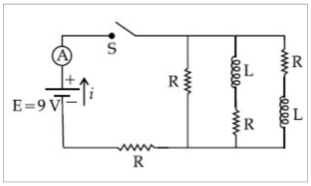
Q: How many trains are visible in this picture? A) 0 B) 3 C) 2 D) 1
A: D

Q: इस चित्र में कितनी रेलगाड़ियाँ दिखाई दे रही हैं? A) 0 B) 3 C) 2 D) 1
A: D

Q: Which animal is seen standing inside the house A) goat B) cow C) horse D) sheep
A: B

Q: कौन सा जानवर घर के अंदर खड़ा दिखाई दे रहा है? A) बकरी B) गाय C) घोड़ा D)भेड़
A: B

JEE-T

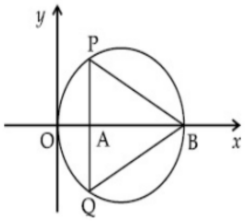


Q: Figure shows a circuit that contains four identical resistors with resistance R=2.0 Ω, two identical inductors with inductance L=2.0 mH and an ideal battery with emf E=9 V. The current 'i' just after the switch 'S' is closed will be :

✓ **70819157131.** 2.25 A **70819157132.** 3.0 A
70819157133. 3.37 A **70819157134.** 9 A

Q: 4 సర్వసమాన నిరోధములను R = 2.0 Ω, 2 సర్వసమాన ప్రేరకములను L = 2.0 mH మరియు ఒక ఆదర్శ బ్యాటరీ emf E = 9 V ని పటములో చూపిన విధంగా కలిపినారు. స్విచ్ 'S' మూయగానే వలయములో ప్రవహించు ప్రవాహం 'i' :

✓ **70819157131.** 2.25 A **70819157132.** 3.0 A
70819157133. 3.37 A **70819157134.** 9 A



Q: In the circle given below, let OA = 1 unit, OB = 13 unit and PQ ⊥ OB. Then, the area of the triangle PQB (in square units) is :

✓ **70819168091.** 24√2 **70819168092.** 24√3
70819168093. 26√2 **70819168094.** 26√3

Q: క్రింద ఇవ్వబడిన వృత్తములో OA = 1 యూనిట్, OB = 13 యూనిట్లు మరియు PQ⊥OB అయితే త్రిభుజము PQB వైశాల్యము (చదరపు యూనిట్లలో) :

✓ **70819168091.** 24√2 **70819168092.** 24√3
70819168093. 26√2 **70819168094.** 26√3

VAANI-T

Q: How many windows are mentioned in the described scene? A) 4 windows B) 2 windows C) 3 windows D) 5 windows.

A: C

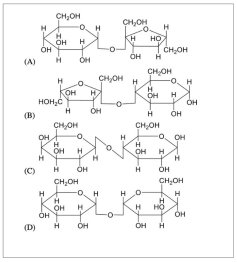
Q: వివరిస్తున్న దృశ్యంలో ఎంత విండోలను ప్రస్తావించారు?A) 4 విండోలు B) 2 విండోలు C) 3 విండోలు D) 5 విండోలు
A: C

Q: What type of place is prominently visible in the scene described? A. Hospital B. Temple C. Market D. School

A: D

Q: అక్కడ వివరించబడిన దృశ్యంలో ప్రధానంగా కనిపించే స్థలం ఏది? A. ఆసుపత్రి B. దేవాలయం C. మార్కెట్ D. పాఠశాల
A: D

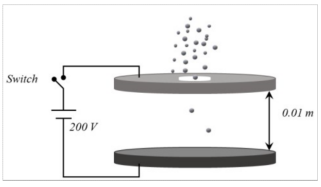
JEE-H



Q: A disaccharide X cannot be oxidised by bromine water. The acid hydrolysis of X leads to a laevorotatory solution. The disaccharide X is

A: A

Q: డైసకరైడ్ X ను బ్రోమిన్ జలతో ఆక్సీకరణ చెందించలేము. X ను ఆమ్ల జలవిశ్లేషణ చేసినపుడు వామభ్రమణ చక్కెర ఏర్పడుతుంది. X అనే డైసకరైడ్ ఏది?
A: A



Q: Two large circular discs separated by a distance of 0.01 m are connected to a battery via a switch as shown in the figure. Charged oil drops of density 900 kg m^{-3} are released through a tiny hole at the center of the top disc. Once some oil drops achieve terminal velocity, the switch is closed to apply a voltage of 200 V across the discs. As a result, an oil drop of radius 8×10^{-5} m stops moving vertically and floats between the discs. The number of electrons present in this oil drop is _____. (neglect the buoyancy force, take acceleration due to gravity = 10 ms^{-2} and charge on an electron (e) = 1.6×10^{-19} C)

A: 6

Q: दो बड़ी वृत्ताकार चकिकाएं (discs), जिनके बीच की दूरी 0.01 m है, एक स्विच के द्वारा एक बैटरी से चित्रानुसार जुड़ी हुई हैं। ऊपरी चकिका के केंद्र में स्थित एक लघु छिद्र से आवेशित तेल की बूँदें, जिनका घनत्व 900 kg m^{-3}, छोड़ी जाती हैं। जब तेल की कुछ बूँदें अंतिम वेग (terminal velocity) प्राप्त कर लेती हैं, तब चकिकाओं के बीच 200 V की वोल्टता लगाने के लिए स्विच को बंद (on) कर दिया जाता है। इसके परिणामस्वरूप, 8×10^{-5}m की त्रिज्या वाली तेल की एक बूँद ऊर्ध्वाधर दिशा में गमन बंद कर देती है तथा चकिकाओं के बीच में तैरने लगती है। तेल की इस बूँद में विद्यमान इलेक्ट्रॉनों की संख्या _____ है (उत्प्लावन (buoyancy) बल को नगण्य मानें, गुरुत्वीय त्वरण का मान 10 ms^{-2} तथा इलेक्ट्रॉन का आवेश (e) = 1.6×10^{-19} C लें।)

A: 6

Figure 5: Qualitative Examples from ***VQAv2, RealWorldQA, CLEVR-Math, VAANI-H, JEE-T, VAANI-T, JEE-H***