Systematic Performance Degradation in Indic Vision-Language Models: Evidence from Hindi and Telugu

Rishikant Chigrupaatii[1,*], Ponnada Sai Tulasi Kanishka[1,*], Lalit Chandra Routhu[1,*]
Martin Patel[1,*], Sama Supratheek Reddy[1], Divyam Gupta[1], Rajiv Misra[1], Rohun Tripathi[2]
[1]Indian Institute of Technology Patna [2]Allen Institute for AI

Abstract

With 1.5 billion people speaking over 120 major languages, India exemplifies the challenges of multilingual AI evaluation. Current multilingual VLM benchmarks suffer from unverified auto-translations, narrow task coverage, small sample sizes, and lack of culturally grounded content. We present HinTel-AlignBench, a comprehensive evaluation framework and benchmark for Hindi and Telugu vision-language models with English-aligned samples. Our framework combines semi-automated translation with human verification to generate ∼4k QA pairs per language across five domains: adapted English datasets (VQAv2, RealWorldQA, CLEVR-Math) and native Indic sets (JEE for STEM, VAANI for cultural grounding). Evaluation of state-of-the-art open and closed-source VLMs reveals consistent performance regression from English to Indic languages, with average drops of 8.3 points for Hindi and 5.5 points for Telugu across four of five tasks. We identify key failure modes and establish reproducible baselines for multilingual multimodal evaluation.

1 Introduction

India's 122 major languages and 1599 other languages[1] present unique challenges for multilingual AI. While recent multimodal large language models (MLLMs) such as ChatGPT (OpenAI, 2025), Gemini 2.5 (Google DeepMind, 2025), and open-weight variants (Meta Llama, 2025; Dash et al., 2025) claim multilingual support, comprehensive evaluation benchmarks for Indian languages remain scarce.

Current evaluation methodologies suffer from critical limitations in quality, scope, and scale. First, many benchmarks rely on unverified automatic translations (Wu et al., 2025), inevitably introducing noise. While text-only benchmarks like IndicGenBench (Singh et al., 2024) exist, they lack multimodal coverage. Second, existing vision-language benchmarks often suffer from insufficient sample sizes; for instance, xChat (Yue et al., 2025) and AyaVisionBench (Dash et al., 2025) contain only 50 and 135 QA pairs per language, respectively, preventing statistically significant analysis. Third, domain coverage is often narrow. Concurrent work such as Kaleidoscope (Salazar et al., 2025) provides ∼800 samples per Indic language but focuses exclusively on exam-based multiple-choice questions, neglecting real-world reasoning. Finally, adapted benchmarks often lack cultural grounding, evaluating surface-level translation rather than native competence (Khan et al., 2024). We elaborate on previous works in the appendix.

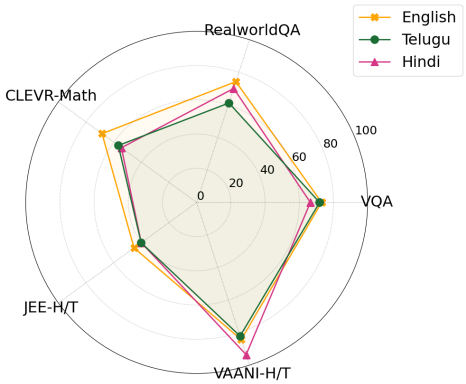

Figure 1: Average performance of GPT-4.1 and Gemini-2.5-Flash on English, Hindi, and Telugu across data-parallel visual question answering samples. Performance regresses from English to Hindi by 8.3 points and from English to Telugu by 5.5 points.

To address these gaps, we introduce HinTel-AlignBench, a scalable framework and benchmark for evaluating VLMs in Hindi and Telugu. Our semi-automated pipeline combines translation or LLM-based QA generation with strict human verifi-

*Equal contribution.

[1]https://en.wikipedia.org/wiki/Languages_of_India

Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR), pages 272–277

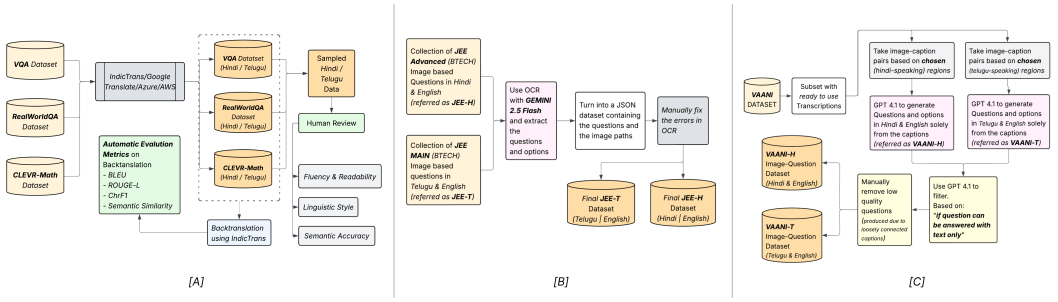

Figure 2: Dataset generation pipeline for (A) VQAv2, RealWorldQA, and CLEVR-Math using translation and human verification; (B) JEE-H and JEE-T using OCR extraction and verification; (C) VAANI-H and VAANI-T using LLM-based question generation from captions with filtering and verification.

cation, achieving 5x faster processing than manual creation for 79% of samples while maintaining linguistic fidelity. The benchmark comprises ∼4k QA pairs per language—significantly larger than prior manually verified sets—spanning five domains: real-world understanding (VQAv2 (Goyal et al., 2017a)), practical reasoning (RealWorldQA (xAI, 2024a)), visual mathematics (CLEVR-Math (Lindström and Abraham, 2022)), STEM competency (JEE-Vision from India's Joint Entrance Exam), and cultural grounding (VAANI (Team, 2025)). Crucially, each sample includes manually verified English translations, enabling direct cross-lingual comparison.

Evaluation of state-of-the-art models on our benchmark reveals systematic performance degradation. Across all models, we observe average regressions of 8.3 points (Hindi) and 5.5 points (Telugu) relative to English, with gaps appearing in four of five tasks (Figure 1). Even frontier models like GPT-4.1 exhibit 3.8-point (Hindi) and 8.6-point (Telugu) performance drops. Performance on aligned Hindi and Telugu subsets differs by less than 1 point, indicating comparable gaps between English and both Indic languages.

Our contributions are: (1) a semi-automated framework for generating multilingual vision-language evaluation sets; (2) the largest human-verified Hindi and Telugu VLM benchmark to date, featuring culturally sourced content and English-aligned samples; and (3) a comprehensive evaluation of state-of-the-art models, highlighting significant performance regressions across diverse domains.

2 Datasets

2.1 Data Sources

We construct HinTel-AlignBench by combining translated English VQA datasets with native Indic evaluation sets across five domains. The translated sets include 1000 samples from VQAv2 (Goyal et al., 2017b) for real-world visual understanding, 765 samples from RealWorldQA (xAI, 2024a,b) for practical spatial reasoning, and 1000 samples from CLEVR-Math (Lindström and Abraham, 2022) for visual mathematical reasoning.

For native Indic content, we develop JEE-Vision from India's Joint Entrance Examination, sourcing 192 Hindi questions (JEE-H) from the Advanced exam and 325 Telugu questions (JEE-T) from the Mains exam. These diagram-dependent STEM problems span mathematics, physics, and chemistry, providing the first benchmark for non-translated multilingual technical reasoning with visual content. We generate culturally grounded evaluation sets (VAANI-H and VAANI-T) by sampling 945 Hindi and 1020 Telugu images from the VAANI corpus (Team, 2025), using GPT-4.1 to create multiple-choice questions from original captions, then filtering out text-only answerable questions.

Translation-based extension enables multi-way parallel data, allowing attribution of performance to task knowledge versus language understanding (Singh et al., 2024). This approach also leverages the quality control invested in designing the original English benchmarks. Table 1 shows the distribution of QA pairs per language and task. Figure 3 showcases a few examples.

2.2 Dataset Generation Framework

Figure 2 illustrates our three-stage generation framework tailored to different data sources.

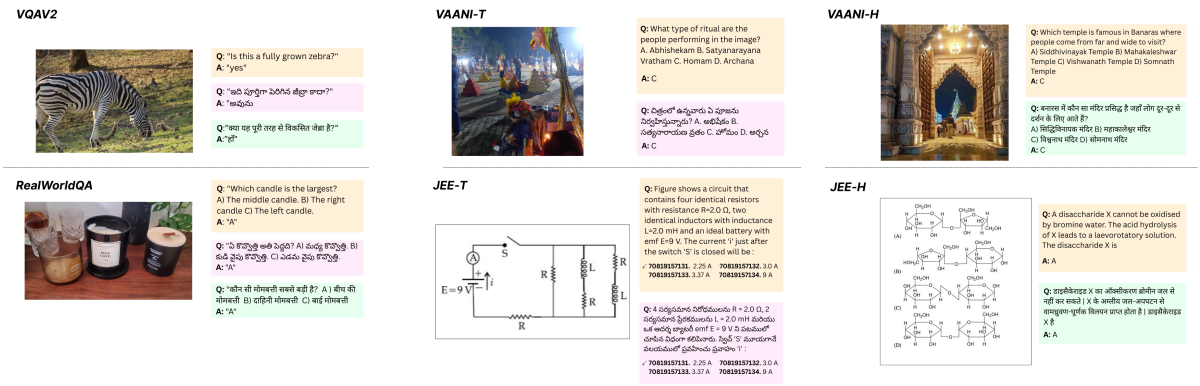

Figure 3: Qualitative Examples for different domains in our dataset. More images are shown in the appendix

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 1: 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.

Translation Pipeline. For VQAv2, RealWorldQA, and CLEVR-Math, we evaluate four translation systems (IndicTrans (Gala et al., 2023), Google Translate, Azure, AWS) on 50 diverse samples per language, selecting Azure for Hindi and AWS for Telugu. All translations undergo manual verification for semantic accuracy, linguistic style, and readability (KJ et al., 2025). We avoid back-translation-based sample selection, which introduces bias toward high-confidence translation errors. Manual review accepts 79% of VQAv2 translations without modification; among modified samples, 42% require only minor changes (verb tense), while 58% need word addition or deletion. Samples requiring only minor edits process 5x faster than generation from scratch. All verification is performed by co-author native speakers. We use one annotator per sample to maximize dataset size within budget constraints.

JEE-Vision Creation. India's Joint Entrance Examination provides authentic STEM problems authored by subject-matter experts in target languages (JEE Mains: 13 languages; Advanced: English/Hindi), avoiding translation artifacts. We curate diagram-dependent problems, evaluating joint understanding of technical visuals and linguistic content. Questions and options are extracted using Gemini-2.5-Flash OCR (Google DeepMind, 2025), then manually verified to correct OCR errors.

VAANI Generation. From the VAANI corpus (Team, 2025), we extract images with text transcriptions from Hindi and Telugu speaking regions. Since no images have both Hindi and Telugu transcriptions, we create separate language-specific sets. Text-only GPT-4.1 generates multiple-choice questions from captions, which undergo two-stage refinement: automated filtering removes questions answerable without images, followed by human verification to eliminate low-quality questions. This process addresses cases where VAANI captions do not perfectly align with images.

3 Experimental Setup

Models. We evaluate open-weight and proprietary models claiming Indic language support. For Hindi, we test Gemma3 (4B, 12B, 27B) (Team et al., 2025), Qwen2.5VL-7B (Bai et al., 2025), Llama3.2-Vision-11B (Meta Llama, 2025), Aya-8B (Dash et al., 2025), Chitrarth-8B (Khan et al., 2024), GPT-4.1 (OpenAI, 2025), and Gemini-2.5-Flash (Google DeepMind, 2025). For Telugu, fewer models provide support; we evaluate the Gemini variants, GPT-4.1, and Chitrarth-8B.

Metrics. For multiple-choice tasks (RealWorldQA, VAANI, JEE), we report standard accuracy. For open-ended generation (VQAv2, CLEVR-Math), we utilize a hybrid evaluation protocol combining exact match with a GPT-4.1 judge to account for linguistic variations, following standard VQA practices (complete details in appendix).

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 2: Results (in %) for Telugu and English. **Bold** indicates the best and underline indicates the next best.

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 3: Results (in %) for Hindi and English. **Bold** indicates the best and underline indicates the next best.

4 Results and Analysis

4.1 Main Results

Tables 2 and 3 present Telugu-English and Hindi-English comparisons. Across all models and tasks, average performance regresses 5.5 points from English to Telugu and 8.3 points from English to Hindi. Performance drops occur in four of five tasks, with VAANI showing smaller gaps. On aligned samples spanning VQAv2, CLEVR-Math, and RealWorldQA evaluated with GPT-4.1, Gemini-2.5-Flash, and Chitrarth, Hindi, Telugu, and English achieve 61.51, 62.53, and 68.6 points respectively, demonstrating systematic degradation from English to both Indic languages.

Gemini-2.5-Flash achieves best overall performance on both language pairs. Chitrarth leads on VQAv2 due to multilingual VQAv2 training. However, models show substantial cross-language variance: GPT-4.1 excels on RealWorldQA in English and Hindi but underperforms on Telugu, highlighting the need for comprehensive evaluation across all target languages. Qwen2.5VL-7B exhibits the largest Hindi-English gap at 25.92 points.

4.2 Task-Specific Analysis

Figure 4 shows average performance regression per task. CLEVR-Math and RealWorldQA exhibit the largest English-Indic gaps, while VAANI shows the smallest. VAANI-H performance exceeds English by 2.09 points on average. Analysis reveals two factors: first, some English questions fail to capture Indic-script option meanings in images with visible Indic text. Second, text-only LLM-generated distractors may enable statistical pattern exploitation.

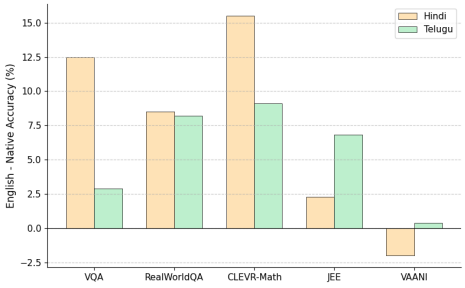

Figure 4: Average performance regression from English to Indic languages per domain. All tasks except VAANI show consistent regression.

Chain-of-Thought prompting improves reasoning tasks but benefits English more than Hindi. On JEE-H, CoT gains 13.8 points in English versus 2.22 points in Hindi, suggesting training bias toward English CoT data. Detailed ablationions and error analysis are reported in the appendix.

5 Conclusion

This paper introduces HinTel-AlignBench, a framework for developing benchmarks to evaluate multimodal large language models in Hindi and Tel-

ugu, 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.

Limitations

While our benchmark introduces a diverse evaluation set it has limitations. First, the proprietary models we evaluated achieve high scores on the VAANI-H/T. We use text only LLMs for generating QA from VAANI captions and they often do not design good distractors. Thus, the models may guess the correct answer by exploiting statistical patterns, which inflates metrics. A future work is using Multi-Binary Accuracy (Cai et al., 2024) for VAANI subsets. Second, the JEE-H benchmark contains only 2 Mathematics questions, due to lack of image-based mathematics questions in the JEE-Advanced examination. Finally, there are 22 official Indian languages and we cover English and 2/22 (Hindi and Telugu) with this work. We hope this benchmark gets extended to all other Indic languages with contributions from native speakers from those languages.

References

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*.

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.

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.

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*.

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

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.

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.

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/.

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.

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*.