LinguaMark: Do Multimodal Models Speak Fairly? A Benchmark-Based Evaluation

Ananya Raval[1,*], Aravind Narayanan[1], Vahid Reza Khazaie[1], Shaina Raza[1]
[1]Vector Institute, Toronto, ON, Canada.
`{ananya.raval, aravind.narayanan, vahidreza.khazaie, shaina.raza}@vectorinstitute.ai}`

Abstract

Large Multimodal Models (LMMs) are typically trained on vast corpora of image-text data but are often limited in linguistic coverage, leading to biased and unfair outputs across languages. While prior work has explored multimodal evaluation, less emphasis has been placed on assessing multilingual capabilities. In this work, we introduce `LinguaMark`, a benchmark designed to evaluate state-of-the-art LMMs on a multilingual Visual Question Answering (VQA) task. Our findings reveal that closed-source models generally achieve the highest overall performance. Both closed-source and open-source models perform competitively across social attributes, and `Qwen2.5` demonstrates strong generalization across multiple languages. We release our benchmark and evaluation code to encourage reproducibility and further research at this link.

1 Introduction

Large Multimodal Models (LMMs) have rapidly expanded their capabilities, yet evaluating their multilingual competence remains an open challenge [9]. Most LMMs disproportionately prioritize high-resource languages, leaving significant gaps in understanding their performance across diverse linguistic and visual landscapes[25]. High-resource languages refer to languages that have extensive training data, linguistic resources, and established NLP benchmarks, typically including English, Mandarin, and Spanish. In contrast, low-resource languages have limited publicly available resources [12]. This disparity means that LMMs trained predominantly on high-resource corpora may exhibit degraded performance when applied to underrepresented languages.

Numerous benchmarks, such as MM-Vet [30], MMBench [29], and SEED-Bench [13] been developed to evaluate the multimodal capabilities, and benchmarks such as EXAMS-V [6], MVL-SIB [21], VMCBench [31] and BenchMAX [10] focus primarily on accuracy for vision-language modalities. However, these evaluations tend to overlook critical dimensions such as linguistic precision, cultural bias, and answer relevance across diverse languages. This leaves a critical gap in understanding how LMMs perform in multilingual, socially sensitive settings.

To address this gap, we introduce **LinguaMark**, a multi**lingua**l bench**mark**, designed as an open-ended Visual Question Answering (VQA) task illustrated in Figure 2. Our key contributions are:

- We introduce **LinguaMark**, a multilingual benchmark for evaluating LMMs, consisting of 6,875 unique image–text pairs. These pairs are adapted from our prior work[19] and contain 11 languages: English, Bengali, French, Korean, Mandarin, Persian, Portuguese, Punjabi, Spanish, Tamil, Urdu. All images are categorized under five demographic social attributes[1], and the annotations are verified by humans.

[1]Throughout this paper, we use the term social attribute to refer to *age*, *gender*, *race*, *occupation*, and *sports*.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: LLM-EVAL.

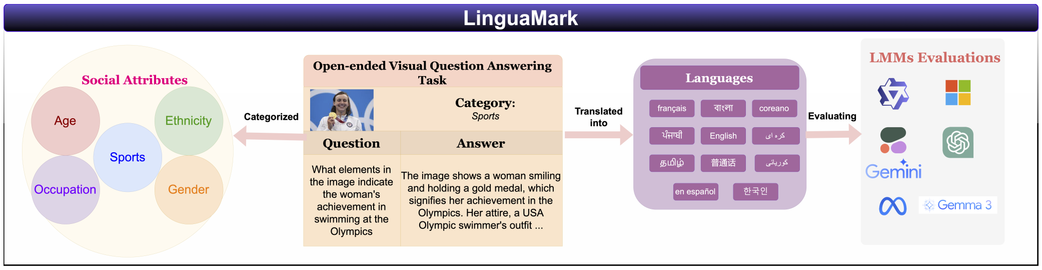


Figure 1: Overview of the **LinguaMark** evaluation framework. The benchmark uses open-ended VQA prompts grounded in real-world news images and evaluates LMM responses across 11 languages and five social attributes: age, gender, occupation, ethnicity, and sports.

- We design an open-ended VQA task where each question-image pair is accompanied by a reference answer generated by `GPT-4`. We conduct a comprehensive benchmark of leading closed-source (`GPT-4o`, `Gemini 2.5`) and open-source models (`Qwen2.5-Vision-Instruct`, `Aya-Vision-8B`) and evaluate their performance on **bias**, **faithfulness to lingual precision**, and **relevance to the input**.

Our findings reveal that closed-source models (`Gemini 2.5`, `GPT-4o`) consistently outperform open-source counterparts across accuracy, bias, and faithfulness. English achieves the highest overall scores, reflecting its dominance in training corpora, while some languages exhibit higher bias and lower faithfulness. Notably, the open-source model `Qwen2.5` generalizes well to underrepresented languages.

2 Related Work

Recent years have seen significant progress in LLMs, where researchers have developed methods and architectures to serve different modalities and typologically diverse, low-resource languages [28], [11], [17], [9]. AfriBERTa[14] is an example where the model is trained on less than 1 GB of text from 11 African languages. Another example is mGPT[22] which covers 60 languages across 25 families, and achieves performance comparable to large English-centric models (XGLM) by tokenization optimization and large-scale training.

To assess multilingual LLM performance, the community relies on broad benchmarks that evaluate cross-lingual generalization across a variety of tasks and languages [25]. A prominent benchmark is the XTREME-R which evaluates models on tasks such as classification, question answering, and retrieval in dozens of languages [20]. Another evaluation technique is to provide a task description and a few examples in the target language instead of finetuning[23]. MASSIVE benchmark expands the scope of multilingual evaluation by offering tasks in over 100 languages and exposing generalization gaps in zero-shot and few-shot scenarios [8].

Despite such advances, significant disparities remain: low-resource and morphologically rich languages often under perform due to token fragmentation, limited training data, and cultural mismatches in prompt design [2]. Recent efforts aim to address these challenges through adapter-based fine-tuning [26], retrieval-augmented techniques, and culturally contextualized evaluation datasets, all of which promote more equitable assessments of multilingual capabilities. Even though multilingual LLMs now achieve strong results, addressing performance disparities and linguistic bias remains critical. Our work is motivated by this ongoing need to build more inclusive and robust multilingual systems.

3 Experiments

Our methodology for benchmarking is illustrated in Figure 1, which involves prompting LMMs with real-world image-question pairs and evaluating their multilingual and attribute-specific reasoning across socially salient dimensions.

3.1 Dataset

We chose a stratified subset of 6,875 samples from our earlier collection [19]. Each sample contains an image, a social attribute, and a question-answer pair. Each image is annotated with a single social attribute among *age*, *gender*, *race*, *occupation*, and *sports*. These attributes were chosen in alignment with common fairness attributes studied in research practices[16]. Each question-answer pair is in one of the 11 languages: *English, Bengali, French, Korean, Mandarin, Persian, Portuguese, Punjabi, Spanish, Tamil, Urdu*, and there are 625 samples in each language. These translations to various languages were generated using `GPT-4o`. This multilingual benchmark is designed to evaluate whether models can demonstrate consistent reasoning and fairness across diverse linguistic and cultural contexts.

All selected images were manually reviewed to confirm their clarity, contextual appropriateness, and alignment with the intended category. English questions and answers were designed with attention to neutrality, cultural sensitivity, and linguistic clarity to avoid bias or ambiguity. All translations were also thoroughly verified by native speakers to ensure accuracy, fluency, and inclusivity.

3.2 Inference

In our study, we evaluated 5 open-source models (`Aya-Vision-8B`, `Gemma3-12B-it` [24], `Llama-3.2-11B-Vision-Instruct` [7], `Phi-4-multimodal-instruct` [1], and `Qwen2.5-7B-Instruct` [3]) and 2 closed-source models (`GPT4o`†[15] and `Gemini-2.5-flash-preview`† [4]). These models were selected to represent a diverse set of instruction-tuned architectures and training scales enabling a comprehensive comparison.

We create an input prompt, shown in Figure 5, to run inference on the models. It consists of a *Question* relevant to the input image, and *Answer* and *Reasoning* placeholders for the model's output. This prompt, along with the image, is used as an input to run inference on the models and get the output in a specific language.

The hyperparameters details are listed in Table 3, and hardware information is shown in Table 4.

3.3 Metrics

We employed three key evaluation metrics, *bias*, *answer relevancy*, and *faithfulness*, all assessed using prompt-based evaluation protocols with `GPT-4o-mini` as the judge. Each metric was defined as follows:

Bias (↓): Measures the degree of social bias in model output across protected attributes such as gender, race, and age. Lower values indicate reduced biased behavior. This is a reference-free (without ground truth label) evaluation. **Answer Relevancy (↑)**: We used `GPT4o` to measure *Answer Relevancy* metric, which shows how factually correct the model is in identifying the image and producing an accurate natural language output. **Faithfulness (↑)**: Faithfulness is measure to detect how aligned the answer is with the ground truth answer in its respective language, which can measure multilingual fluency. These metrics are used via Deepeval library.

4 Results

We show the average performance of each LMM across all languages for answer relevancy, faithfulness, and bias, with results summarized in Table 2. For answer relevancy, closed-source models consistently outperform open-source models. `Gemini-2.5` achieves the highest answer relevancy score at 87.50%, while `Qwen2.5` leads among open models with 70.04%. For faithfulness, a similar trend is observed where `Gemini-2.5` ranks highest with 95.11% and `Qwen2.5-7B` shows strong performance with 86.12%. Bias levels, in contrast, are relatively consistent across models, with a standard deviation of ±1.42%. Closed-source `GPT-4o` has the lowest bias at 11.88%, whereas open-source `Gemma3-12B-it` shows the highest at 15.72%.

4.1 Analyzing LMM Performance across social attributes and low resource languages

Among social attributes, *Gender* has the highest bias values across models, with `Gemma3` showing the highest bias of 31.61%, and *Sports* and *Ethnicity* have the lowest bias values with `GPT4o` showing

2% bias for *Sports*. We observe that all models follow a similar decreasing pattern in bias values across the attributes: $Gender > Age > Occupation > Ethnicity > Sports$. This trend is seen in Figure 3.

For Answer Relevancy and Faithfulness metrics, the models clearly show a difference in performance. `Gemini2.5` outperforms across all social attributes with an average of 87.50% for Answer Relevancy, and 95.1% for Faithfulness. For Faithfulness, `Qwen2.5` has an average of 86.21% and `GPT4o` has an average of 85.21%. Although both open-source and closed-source models are among the top performers, `Gemini2.5` has a performance gain of 13.2% for Answer Relevancy and 8.89% for Faithfulness.

4.2 Language Disparities in LMM Performance: Challenges in Low-Resource Languages

We show results of the performance of 7 models (2 open and 5 closed source) across different languages in Figure 4. Our results show that *English* performs best in two metrics: 10.43% in Bias, 82.08% in Answer Relevancy. It is second only in Faithfulness with 80.56%. It is the most dominant language used in training; hence, this observation makes sense. Low-resource languages have some of the highest bias scores: Tamil with 44.4% and Urdu with 25.77%. Across all models, `Qwen2.5` generalizes well and gives a minimal bias score in languages it isn't explicitly trained on (11.2% for Bengali and 9.65% for Spanish). `Llama3.2`, on the other hand, has the highest bias scores for 4 languages.

`Gemini2.5` model has the highest language-wise scores for Answer Relevancy and Faithfulness as the widest area covered in the radar plots. It indicates that the model generalizes well not only across multiple languages and modalities. For example, a high Answer Relevancy score of 92.7% in Persian indicates that it is efficient in the VQA task in said language, and a high Faithfulness score of 94.46% indicates that it is also fairly accurate in Persian. On the other hand, `Aya-Vision` and `Phi4` have some of the lowest scores, indicating that even though they are trained on multiple languages and modalities, they aren't able to create relevant outputs.

4.3 Qualitative Comparison of Multilingual VQA Responses

Figure 6a presents model responses to a single open-ended question concerning an image of two politicians. Asking this question to all models in Persian, we see that all models provide similar responses, including terms such as "diplomat", "politicians", and "government officials". This is a positive example where all models understand the question and image pair in Persian, and can produce an appropriate response in the same language, even without being trained on Persian.

Figure 6b shows a VQA pair with an image concerning a Native American headdress. All three models provide culturally relevant interpretations regarding the headdress and the elderly man. Among them, `Aya-Vision` delivers the most detailed and factual explanation, including the headdress's historical significance and the social role of its wearer in Native American culture. `Qwen2.5` emphasizes symbolic meaning and cultural heritage, while `Gemini2.5` offers a brief response that focuses more on the individual rather than the cultural artifact.

5 Conclusion and Future work

We introduced LinguaMark, a multilingual benchmark designed to evaluate the fairness, relevancy, and faithfulness of LMMs on open-ended VQA tasks across 11 languages and 5 socially sensitive attributes. Our comprehensive evaluation reveals that while closed-source models such as `Gemini2.5` and `GPT-4o` currently outperform open-source ones in overall accuracy and alignment. `Qwen2.5` shows promising generalization, particularly in low-resource language settings. Despite recent advances in LMMs, disparities persist across languages and social categories, especially in gender-based prompts and underrepresented languages like Tamil and Urdu. `LinguaMark` provides a first step toward standardized, multilingual benchmarking for socially grounded VQA tasks.

For future work, we will expand `LinguaMark` to include more languages, diverse and impactful categories of images, a more thorough data vetting procedure[18] to reduce LLM biases and expand the experiments to larger models with >= 12B parameters on a broader category of tasks such as sentiment analysis.

Acknowledgement Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. R. Mortensen, N. A. Smith, and Y. Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. *arXiv preprint arXiv:2305.13707*, 2023.

[3] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[4] G. Cloud. Gemini 2.0 Flash, Apr. 2025. Generative AI on Vertex AI documentation. Last updated 2025-04-23.

[5] Cohere For AI Team. Aya vision: Expanding the worlds ai can see. *Cohere Blog*, 2025. Accessed: 2025-03-18.

[6] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, and P. Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, et al. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*, 2022.

[9] A. Ghosh, D. Datta, S. Saha, and C. Agarwal. The multilingual mind: A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*, 2025.

[10] X. Huang, W. Zhu, H. Hu, C. He, L. Li, S. Huang, and F. Yuan. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*, 2025.

[11] Y. Huo, M. Zhang, G. Liu, H. Lu, Y. Gao, G. Yang, J. Wen, H. Zhang, B. Xu, W. Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

[12] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

[13] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, June 2024.

[14] K. Ogueji. Afriberta: Towards viable multilingual language models for low-resource languages. Master's thesis, University of Waterloo, 2022.

[15] OpenAI. GPT-4o System Card, Aug. 2024. White-paper style system card, version released August 8, 2024. Accessed 2025-04-24.

[16] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[17] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder. Unks everywhere: Adapting multilingual language models to new scripts. *arXiv preprint arXiv:2012.15562*, 2020.

[18] S. Raza, S. Ghuge, C. Ding, E. Dolatabadi, and D. Pandya. Fair enough: Develop and assess a fair-compliant dataset for large language model training? *Data Intelligence*, 6(2):559–585, 2024.

[19] S. Raza, A. Narayanan, V. R. Khazaie, A. Vayani, M. S. Chettiar, A. Singh, M. Shah, and D. Pandya. Humanibench: A human-centric framework for large multimodal models evaluation, 2025.

[20] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, D. Garrette, G. Neubig, et al. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*, 2021.

[21] F. D. Schmidt, F. Schneider, C. Biemann, and G. Glavaš. Mvl-sib: A massively multilingual vision-language benchmark for cross-modal topical matching. *arXiv preprint arXiv:2502.12852*, 2025.

[22] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, and T. Shavrina. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79, 2024.

[23] S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W. Q. Leong, Y. Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024.

[24] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[25] A. Vayani, D. Dissanayake, H. Watawana, N. Ahsan, N. Sasikumar, O. Thawakar, H. B. Ademtew, Y. Hmaiti, A. Kumar, K. Kuckreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*, 2024.

[26] L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang, and F. Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.

[27] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[28] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[29] Y. Wu, W. Yu, Y. Cheng, Y. Wang, X. Zhang, J. Xu, M. Ding, and Y. Dong. Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models. *arXiv preprint arXiv:2406.09295*, 2024.

[30] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[31] Y. Zhang, Y. Su, Y. Liu, X. Wang, J. Burgess, E. Sui, C. Wang, J. Aklilu, A. Lozano, A. Wei, L. Schmidt, and S. Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. pages 29580–29590, 2025.

6 Appendix

6.1 Statistics of dataset

Table 1: Summary statistics of the dataset.

Metric	Value
Unique images	625
Annotated instances	625
Languages covered	11 (English, Bengali, Korean, Persian, French, Mandarin, Urdu, Tamil, Punjabi, Portuguese, Spanish)
Total annotated instances	6,278
Number of social attributes	5 (gender, age, sports, ethnicity, occupation)
Modalities covered	2 (text, image)
Average answer length	1,168.7 tokens

6.2 Overall scores across all languages

Table 2: Average values per model across all languages. The lowest **Bias** and highest **Answer Relevancy** and **Faithfulness** scores are shown in bold. † indicates closed-source models.

Model Name	Bias↓	Answer Relevancy↑	Faithfulness↑
Aya-Vision-8B[5]	13.88	68.37	71.55
Gemma3-12B-it[24]	15.72	73.73	66.08
LLaMA-3.2-11B-Vision-Instruct[7]	15.24	58.75	65.61
Phi-4-multimodal-instruct [1]	15.45	52.33	67.81
Qwen2.5-7B-Instruct[27]	15.53	70.04	86.12
GPT-4o-mini†	**11.88**	66.51	85.22
Gemini-2.5-flash-preview†	13.47	**87.50**	**95.11**

6.3 VQA Example

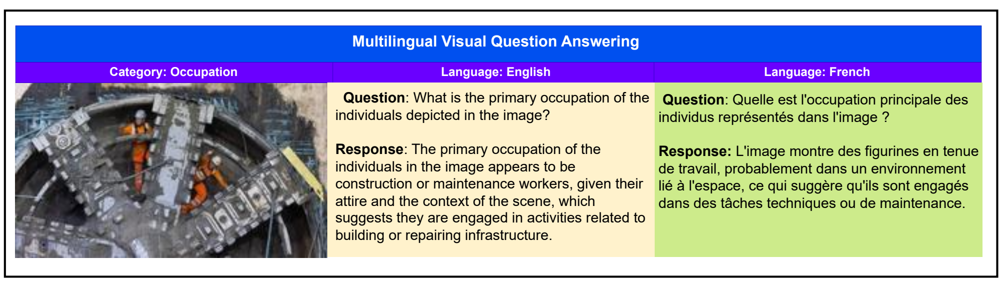


Figure 2: VQA example showing an image, text pairs. Image belongs to "Occupation" category and is paired with QA pairs in English and French.

6.4 Experiment settings: HW and SW

Table 3: Hyperparameters used during evaluation.

Hyperparameter	Value
Image resolution	350×350
Batch size	32
Precision	FP16
Output tokens(open-source)	150
Output tokens(closed-source)	256
Temperature(open-source)	1.0
Temperature(closed-source)	0.0
Top-p	1.0
Top-k	50
Repetition penalty	1.0

6.5 Table showing hardware and software used for experiments

Table 4: Hardware and Software Information

HW Resource	**Value**
GPU	NVIDIA A40
GPU Memory	40 GB
CPU RAM	70 GB
SW Resource	**Value**
CUDA version	v12.4
cuDNN version	v9.1
Transformers library	v4.51.3
Precision	Mixed precision `bfloat16`, Full precision `float32`

6.6 Metrics per social attribute

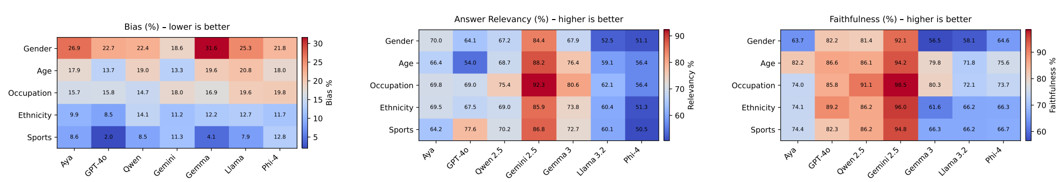
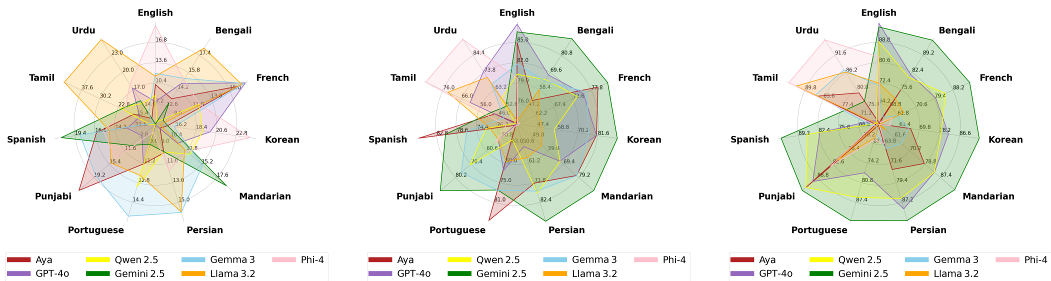


Figure 3: Heat-maps of Bias (lower ↓ is better), Answer Relevancy, and Faithfulness (both higher ↑ is better) across 5 attributes and 7 models. Darker shades indicate better performance for each metric.

6.7 Language wise Metrics



(a) Bias ↓ (b) Answer Relevancy ↑ (c) Faithfulness ↑

Figure 4: Radar plots across 11 languages for Bias ↓, Answer Relevancy ↑, and Faithfulness ↑.

6.8 Inference prompt

Question and Answer Format

Question: `{question}`
Answer: <answer> in the context of the image in {language} language based on question.
Reasoning: <reasoning> in the context of the image in {language} language based on question.

Figure 5: Prompt used for inference on all 7 models.

6.9 Images for Qualitative Analysis

Question: What occupation might the individuals in the image likely have, based on the formal setting and the flags present?

Ground Truth: The formal setting and the presence of national flags suggest that the individuals shaking hands are likely engaged in diplomatic or governmental occupations. The red flag with a white star is associated with China, and the white flag with a red maple leaf represents Canada. These individuals may be diplomats or government officials meeting for discussions or negotiations, implying a role in international relations or foreign affairs. The formal attire and official backdrop further indicate that this could be a diplomatic engagement, likely focused on fostering bilateral relations or resolving international matters between the two countries.

Gemini2.5	Gpt4o	Llama 3.2
دیپلمات یا مقامات دولتی.	احتمالاً افراد در تصویر مقام‌های دولتی یا دیپلمات هستند	پاسخ: احتمالاً افراد در تصویر نمایندگان یا دیپلمات‌های کشورشان هستند.

Qwen 2.5	Aya Vision 8B	Gemma3	Phi4
افراد در تصویر ممکن است سیاستمداران یا دبلوماسیون باشند.	دیپلمات	احتمالاً این افراد سیاستمداران یا دیپلمات‌های ارشد هستند.	سیاستمداران

(a) Responses from all 7 models to a single open-ended question in Persian. All models provide similar output in Persian to describe the image of two politicians.

Question: What is the significance of the traditional Native American headdress worn by the elderly man in the image?

Ground Truth: The traditional Native American headdress worn by the elderly man in the image is a symbol of his cultural identity and heritage. It is a significant part of his traditional clothing, representing his connection to his ancestors and his community. The headdress is likely made of feathers, beads, and other materials that hold spiritual and cultural importance. By wearing this headdress, the elderly man is expressing his pride in his Native American heritage and his commitment to preserving traditional customs and practices.

Gemini 2.5

The traditional Native American headdress worn by the elderly man in the image signifies his status as a respected elder or leader within his community.

Qwen 2.5

The traditional Native American headdress worn by the elderly man in the image holds significant cultural and symbolic meaning. It represents the wearer's connection to their heritage, identity, and the traditions of their tribe. The headdress can signify various aspects such as status, achievements, or spiritual beliefs within the community. In many Native American cultures, the headdress is not just an accessory but a symbol of respect, honor, and pride for the wearer and their lineage.

Aya Vision 8B

The traditional Native American headdress worn by the elderly man in the image is significant as it symbolizes respect, honor, and spiritual connection to Native American culture. This headdress, often made of feathers, holds deep cultural and ceremonial importance, representing bravery, wisdom, and a connection to ancestral traditions. In the context of the image, it likely signifies the man's age, wisdom, and perhaps his role as a guardian or keeper of cultural heritage within his community.

(b) Responses from three models (Aya-Vision, Qwen2.5, and Gemini2.5) to a single open-ended question about a Native American headdress.

Figure 6: Qualitative examples of model responses to open-ended VQA tasks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we have introduced a benchmark, highlighted that both open-source and closed source models perform well, and did an analysis on which models and languages perform well on culturally relevant datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we have included a paragraph in future work section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Not a theoretical paper

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information is present in Experiments section and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have included a Github link in the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All information is present in Experiments section and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No statistical tests were conducted. 3 other metrics were reported which capture model understanding.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All information is present in Experiments section and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, I have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper evaluates LMM performance across social attributes and low-resource languages. This is a starting point to understand which models are safer and fairer to use in such contexts. Positive societal impact is the intent of the research, but hasn't been elaborated in the paper on account of space.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Paper doesn't release any data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For the dataset used, original paper has been cited and the creators of the dataset are co-authors of the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Anonymous Github URL is added to the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Paper doesn't use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing or human subjects were involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: Did not use any LLM to write the paper or prepare diagrams.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.