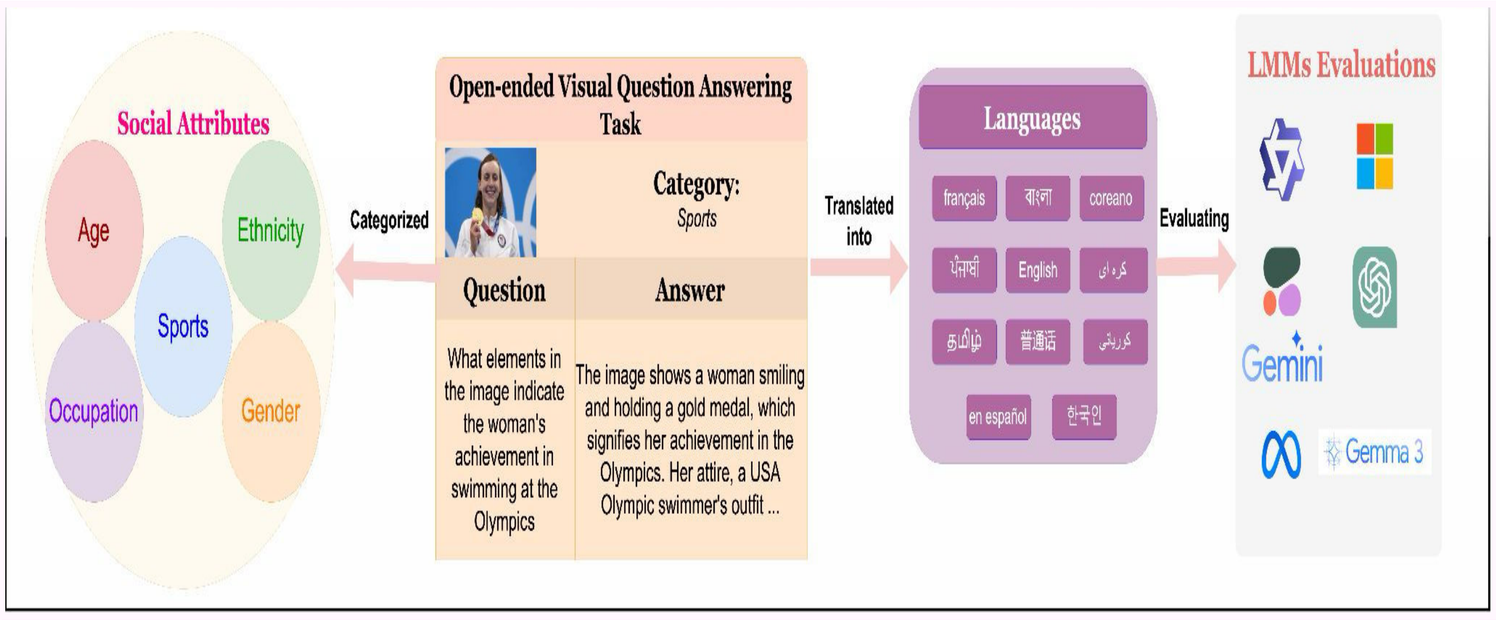
LinguaMark: Do Multimodal Models Speak Fairly? A Benchmark-Based Evaluation

Ananya Raval*, Aravind Narayanan*, Vahid Reza Khazaie*, Shaina Raza
Vector Institute of Artificial Intelligence

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Many LMMs disproportionately prioritize high-resource languages, leaving significant gaps in performance understanding across diverse linguistic and visual landscape. Linguistic precision, cultural bias, and answer relevance across diverse languages are a few gaps to be looked at.



QUANTITATIVE ANALYSIS

Contributions

- A multilingual benchmark of **6,875** unique image-text pairs in **11** languages adopted from HumaniBench for evaluating LMMs.
- Comprehensive experiments with leading closed and open-source models evaluating their performance on 3 metrics.

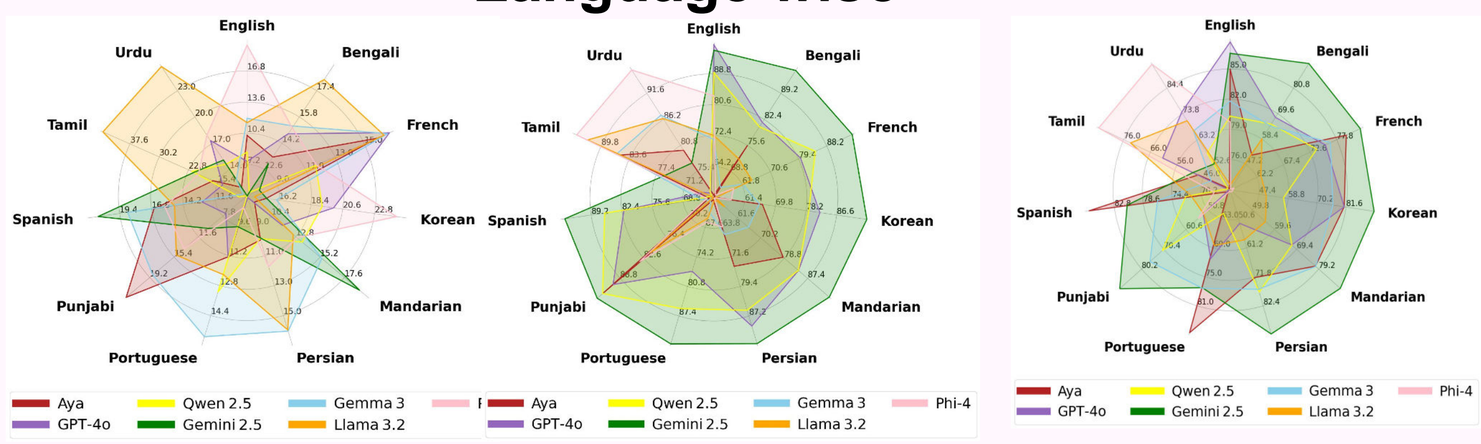
Experiments

- We run inference on **7** models: Aya-Vision-8B, Gemma3-12B-it, Llama-3.2-11B-Vision-Instruct, Phi-4-multimodal-instruct, Qwen2.5-7B-Instruct, GPT4o, Gemini-2.5-flash-preview.
- Input prompt used for inference: (i) *Question* about the input image, (ii) *Answer placeholder*, (iii) *Reasoning* placeholder.

Evaluation Metrics

Bias: Degree of social bias across protected attributes.
Answer Relevancy: How factually correct the model is in identifying the image and producing an accurate natural language output.
Faithfulness: Detects how aligned the answer is with the ground truth answer.

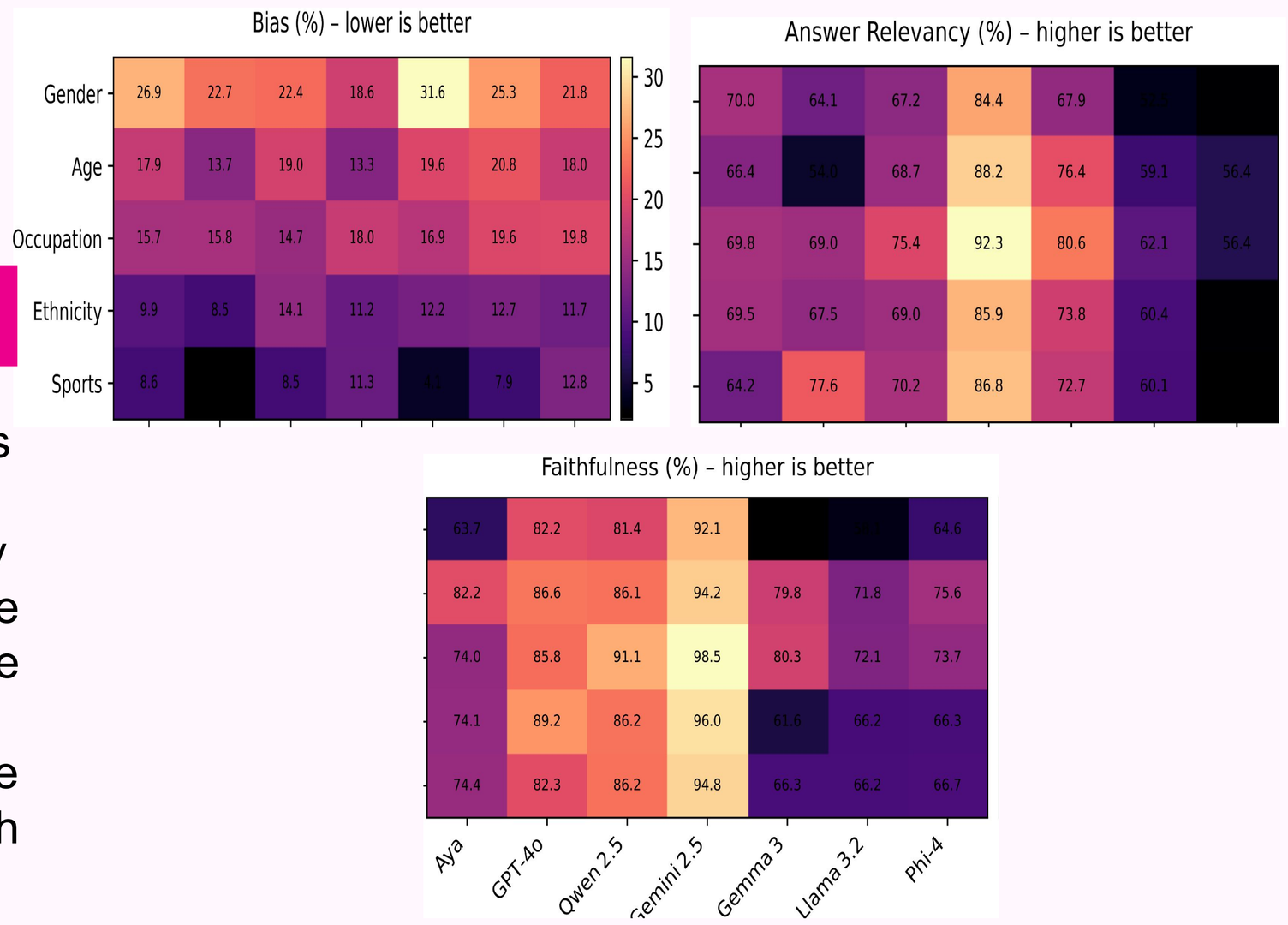
Language wise



Bias **Answer Relevancy** **Faithfulness**

- **English** performs best in Answer Relevancy and Faithfulness.
- **Qwen2.5** generalizes well and gives a minimal bias score in languages it isn't explicitly trained on.
- **Gemini2.5** model has the highest language-wise scores for Answer Relevancy and Faithfulness indicating that it generalizes across multiple languages and modalities.

Demographic wise



- **Gemini2.5** outperforms across all social attributes.
- All models follow a similar decreasing pattern in bias values
  across the attributes: ***Gender > Age > Occupation > Ethnicity > Sports.***