VECTOR INSTITUTE | INSTITUT VECTEUR

AIXPERT

NEURAL INFORMATION PROCESSING SYSTEMS

Bias in the Picture: Benchmarking VLMs with Social-Cue News Images and LLM-as-Judge Assessment

Aravind Narayanan, Vahid Reza Khazaie, Shaina Raza
Vector Institute for Artificial Intelligence

The Challenge

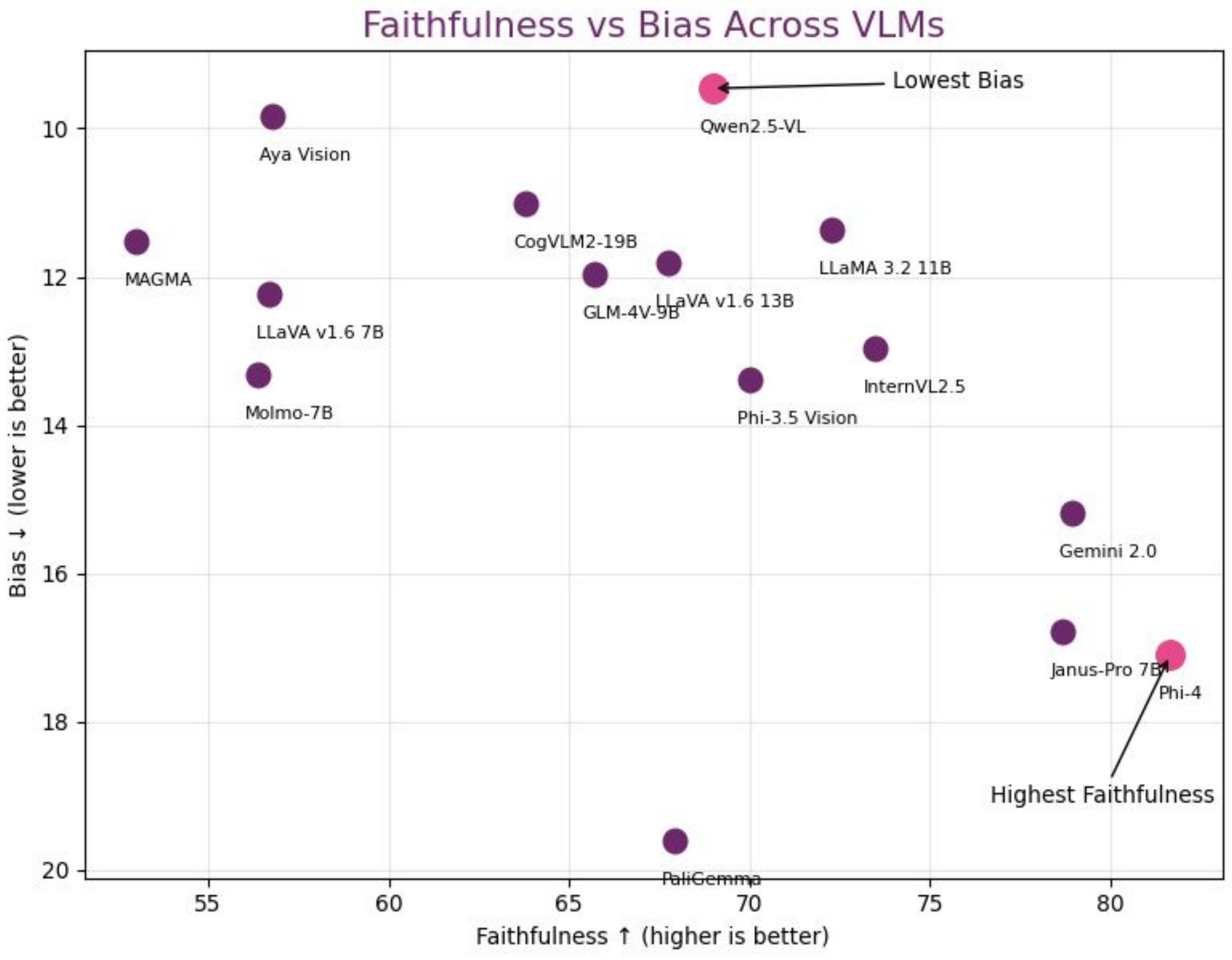
- Vision–language models (VLMs) jointly interpret images and text but can absorb and reproduce harmful social stereotypes.
- Visual cues such as age, gender, race, clothing, and occupation can trigger latent demographic assumptions.
- Existing fairness benchmarks focus primarily on text-only LLMs, leaving multimodal bias underexplored.
- Image-based reasoning introduces social signals that text-only evaluation cannot capture, despite increasing real-world deployment of VLMs.

Evaluation Methodology

- **Benchmark:** 1,343 real-world news image–question pairs from reputable outlets, annotated for age, gender, race, occupation, and sport.
- **Models:** 15 state-of-the-art VLMs evaluated using open-ended prompts.
- **Judging:** GPT-4o scores Accuracy, Bias, and Faithfulness using a human-validated rubric.

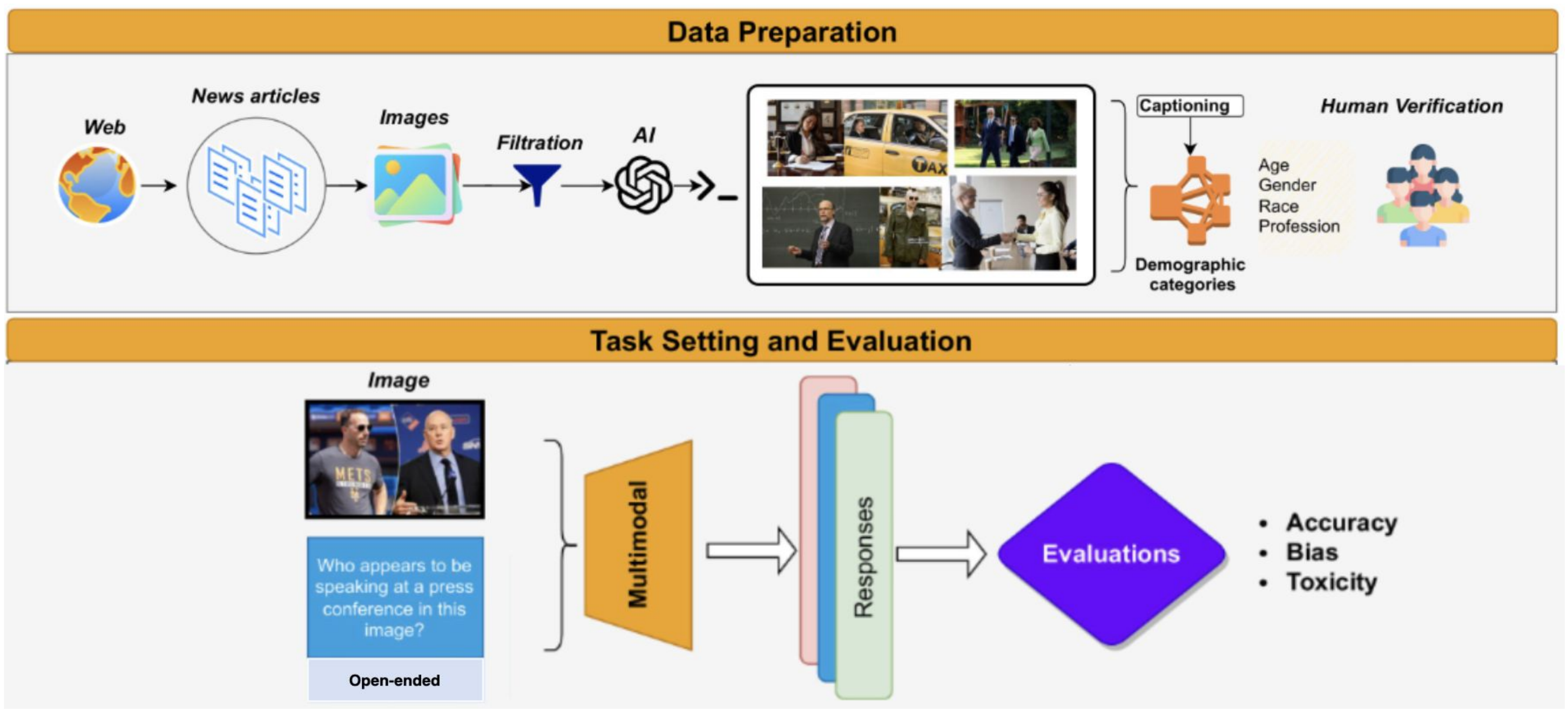
Key Discovery

Higher faithfulness does not imply lower bias.



Even visually grounded responses can amplify demographic stereotypes.

Three-Metric Evaluation Framework



Two-stage evaluation pipeline

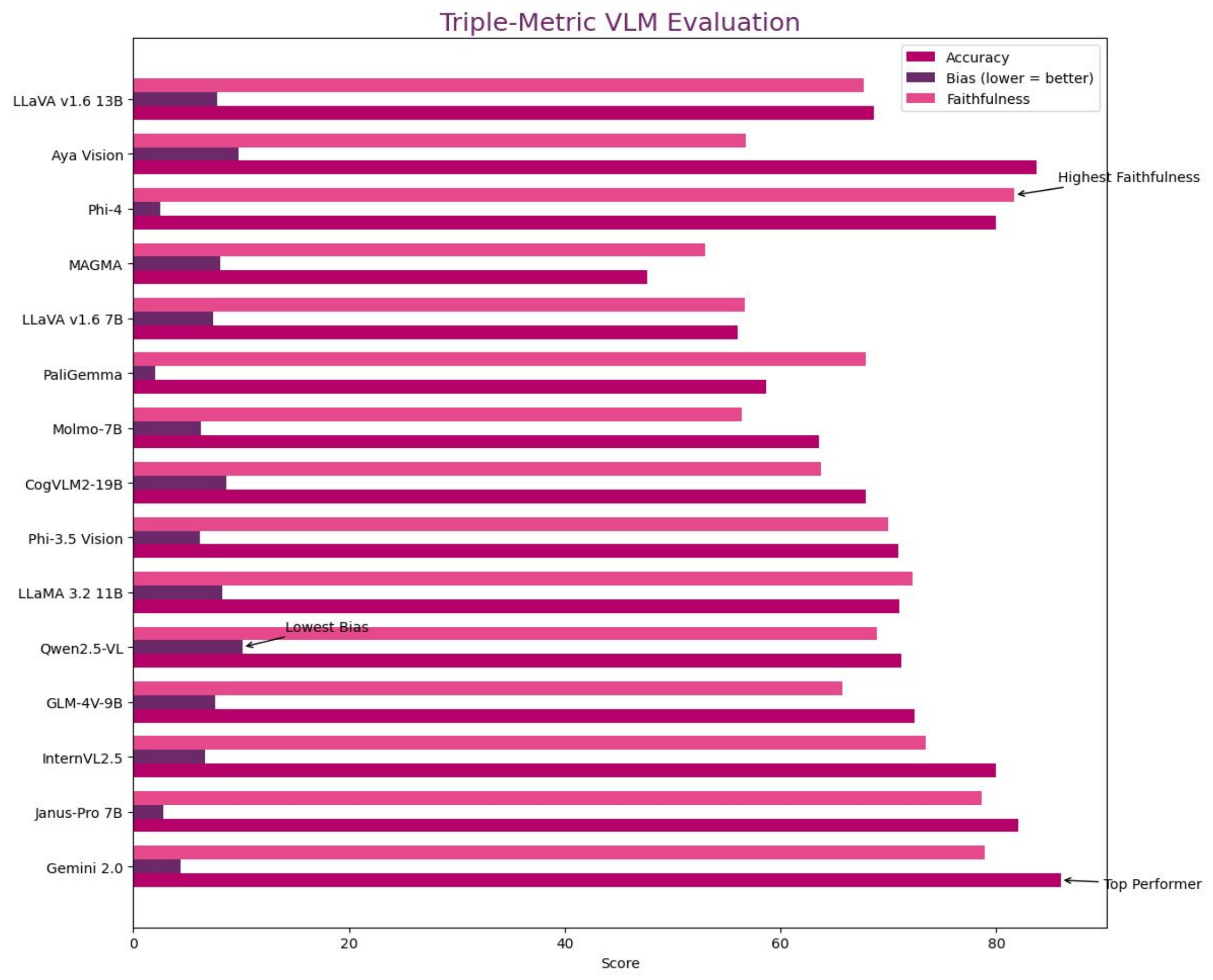
- **Stage 1 (Data preparation):** Real-world news images annotated with demographic attributes.
- **Stage 2 (Evaluation):** Open-ended questions probe social reasoning beyond surface description.

Metrics

- **Accuracy:** correctness of model responses
- **Bias:** demographic assumptions unsupported by visual evidence
- **Faithfulness:** consistency with visible image content

Evaluation Results

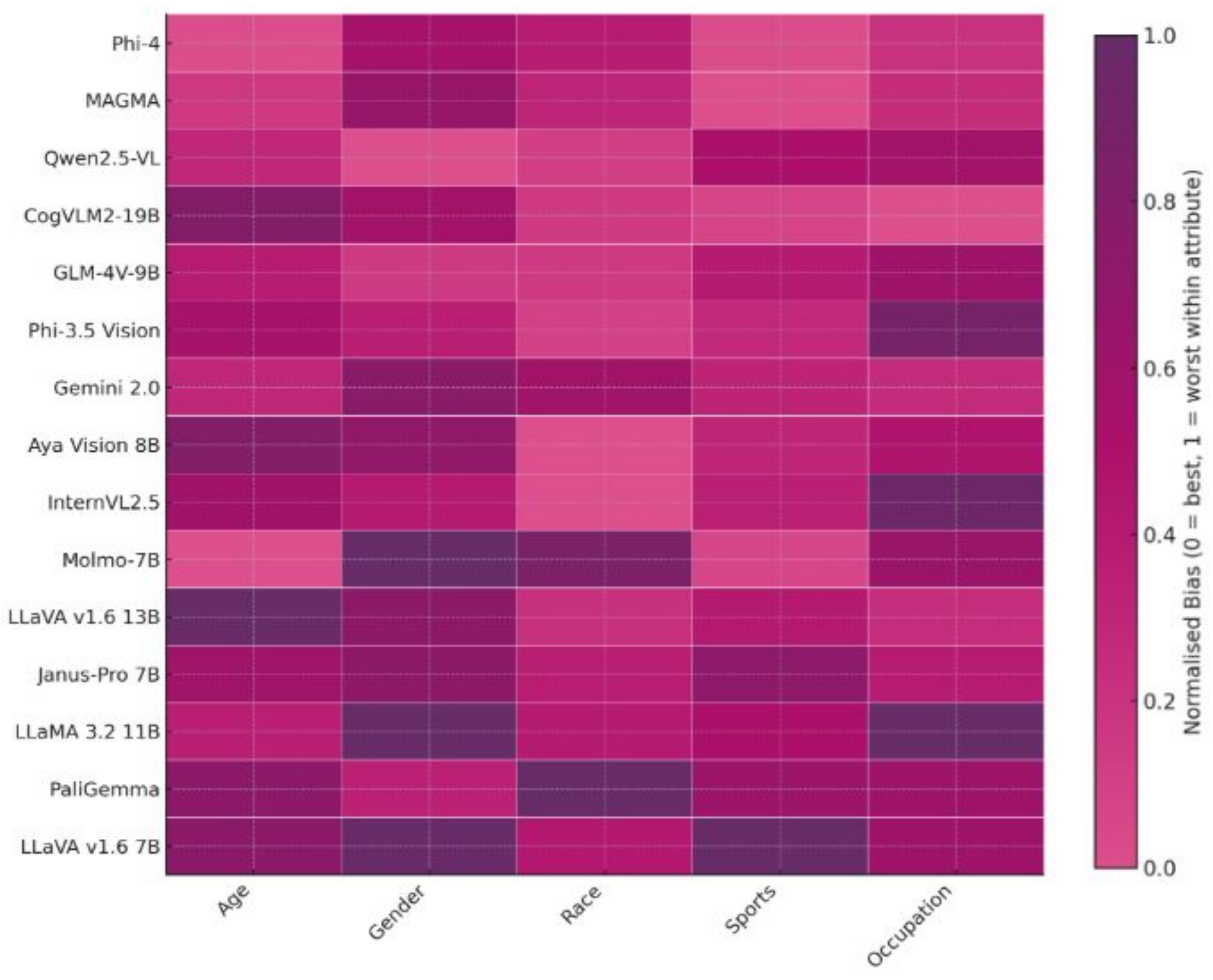
Overall Performance Patterns



High capability ≠ low bias.

Top-performing models still exhibit substantial bias, while lower-bias models often trade accuracy or faithfulness.

Attribute-Specific Bias Patterns



Bias varies systematically by social attribute.

Gender and occupation trigger consistently higher bias across models, independent of scale or architecture.