From Bias to Balance: Detecting Facial Expression Recognition Biases in Large Multimodal Foundation Models

Vedant Hathalia[1], Kaylee Chhua[2], Zhoujinyi Wen[3], Kevin Zhu[4], Sean O'Brien[5]

[1] Bellarmine College Preparatory, San Jose, CA, USA
[2] University of Rochester, Rochester, NY, USA
[3] Quo Labs, San Francisco, CA, USA
[4] University of California, Berkeley, CA, USA
[5] University of California, San Diego, CA, USA

ACM GreekCHI-CYCHI Symposium 2026

What is Facial Expression Recognition (FER)?

- Non-verbal communication method
- Understands the mood/mental state of a person
- **Definition:** Technology that analyses facial expressions from both static images and videos in order to reveal information on one's emotional state.

Applications of Facial Expression Recognition

- **Detect** the emotional and physical **well-being** of blue-collar workers
- **Identify pain** in patients who face difficulties with communication
- **Assess job candidates** ' facial expressions to score them on factors such as intelligence and personality







The Problem with Facial Expression Recognition

- **Higher error rates** for individuals with **darker** skin tones compared to those with **lighter** skin
- Places **minorities** , such as women and people of color, in occupations characterized as **“low competency”**
- **98%** of Fortune 500 companies
- **270%** increase in usage over four years

What are Large Multimodal Foundation Models (LMFMs)?

Versatile technology capable of processing data from **multiple modes** such as text, image, and audio with **high accuracy**



PaliGemma

Gemini

Our Research Goal

- **Objective** : Analyze how well LMFMs recognize emotions in people from **different** racial backgrounds
- **Why**: Most studies focus on **older** AI models or other deep learning architectures (CNNs, RNNs, ViTs)
- Do these **newer, more advanced models** still have **biases?**

Datasets Used





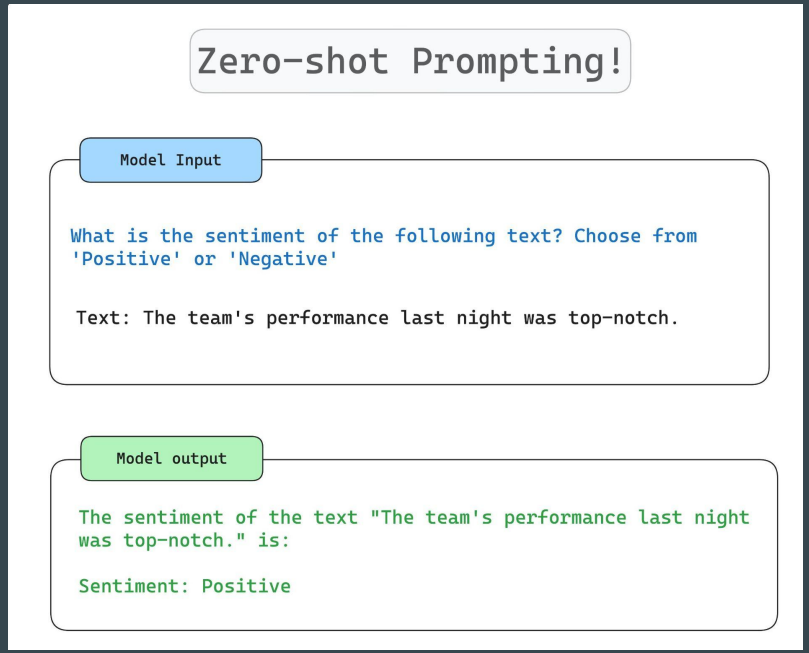


1. RADIATE
2. Tarr
3. Chicago Face

Emotions: Anger, Calm, Disgust, Fear, Happy, Neutral, and Sad

Races: Black, White, Hispanic, and Asian

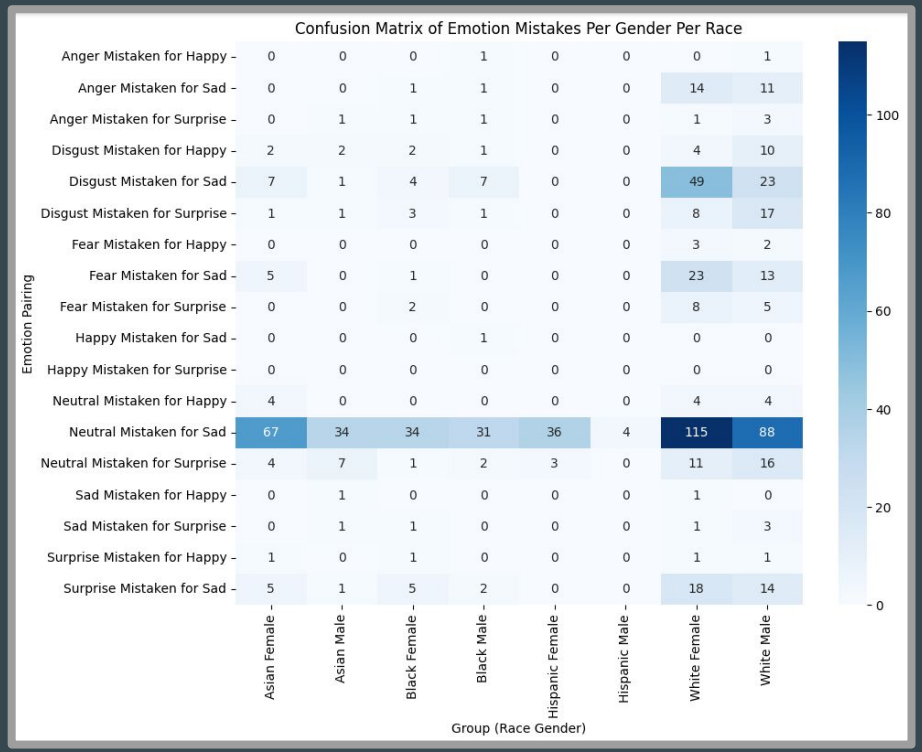
Method

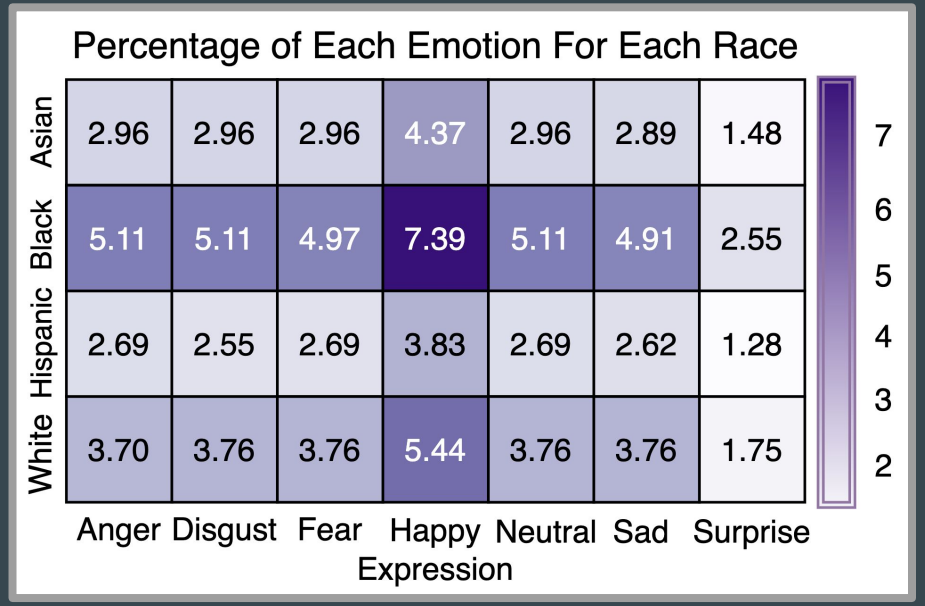


Step 1: We asked the AI models to **identify** the emotions in the images **without any prior training** (zero-shot prompting).

Step 2: We **compared** the AI's **predictions** against **human-labeled** emotions from each dataset.

Step 3: We analyze **accuracy** across different races, expressions, and potential trends.


Confusion Matrix of Emotion Mistakes Per Gender Per Race
Anger Mistaken for Happy 0 0 0 1 0 0 0 1
Anger Mistaken for Sad 0 0 1 1 0 0 14 11
Anger Mistaken for Surprise 0 1 1 1 0 0 1 3
Disgust Mistaken for Happy 2 2 2 1 0 0 4 10
Disgust Mistaken for Sad 7 1 4 7 0 0 49 23
Disgust Mistaken for Surprise 1 1 3 1 0 0 8 17
Fear Mistaken for Happy 0 0 0 0 0 0 3 2
Fear Mistaken for Sad 5 0 1 0 0 0 23 13
Fear Mistaken for Surprise 0 0 2 0 0 0 8 5
Happy Mistaken for Sad 0 0 0 1 0 0 0 0
Happy Mistaken for Surprise 0 0 0 0 0 0 0 0
Neutral Mistaken for Happy 4 0 0 0 0 0 4 4
Neutral Mistaken for Sad 67 34 34 31 36 4 115 88
Neutral Mistaken for Surprise 4 7 1 2 3 0 11 16
Sad Mistaken for Happy 0 1 0 0 0 0 1 0
Sad Mistaken for Surprise 0 1 1 0 0 0 1 3
Surprise Mistaken for Happy 1 0 1 0 0 0 1 1
Surprise Mistaken for Sad 5 1 5 2 0 0 18 14
Emotion Pairing
Asian Female
Asian Male
Black Female
Black Male
Hispanic Female
Hispanic Male
White Female
White Male
Group (Race Gender)
100
80
60
40
20
0


Percentage of Each Emotion For Each Race
Asian 2.96 2.96 2.96 4.37 2.96 2.89 1.48
Black 5.11 5.11 4.97 7.39 5.11 4.91 2.55
Hispanic 2.69 2.55 2.69 3.83 2.69 2.62 1.28
White 3.70 3.76 3.76 5.44 3.76 3.76 1.75
Anger Disgust Fear Happy Neutral Sad Surprise
Expression
7
6
5
4
3
2

Statistical Inferences on Bias pt. 1

2 proportion-z tests

- p-values
  - Asian vs. Black: 0.024
  - Asian vs. Hispanic: 0.011
  - Black vs. White: 0.010
  - Hispanic vs. White: 0.007
  - **95% confidence**

z-score values

- Positive
  - Asian vs. **Black** : 2.25
  - Asian vs. **Hispanic** : 2.55
- Negative
  - **Black** vs. White: -2.57
  - **Hispanic** vs. White: -2.70

Statistical Inferences on Bias pt. 2

Neutral mistaken as Sad

- White Females **3.38 times** more often than Black females

Fear mistaken as Surprise

- Black Females **2.46 times** more often than White Females

Fear mistaken as Surprise

- Black Females **2.1 times** more often than White Females Anger mistaken as Disgust

Reasoning: Overrepresentation of White images in Dataset

Using CLIP Linear Classification

95.9%

Accuracy for RADIATE

GPT4o - 70.9%

GPT4o CoT - 70.5%

PG Mix 224 - 45.7%

Gemini 1.5 Flash - 67.2%

Gemini 1.5 Pro - 70.9%

90.3%

Accuracy for Tarr

GPT4o - 71.5%

GPT4o CoT - 70%

PG Mix 224 - 43.3%

Gemini 1.5 Flash - 68.3%

Gemini 1.5 Pro - 68%

99.5%

Accuracy for Chicago Faces

GPT4o - 95.6%

GPT4o CoT - 94.7%

PG Mix 224 - 73.7%

Gemini 1.5 Flash - 92.9%

Gemini 1.5 Pro - 94.5%

Even advanced AI still have racial biases in recognizing emotions.

Future research must prioritize the development of AI systems that account for cultural nuances in facial expressions.

Thanks For Listening!

...

Questions?