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

Abstract

This study addresses the racial biases in facial expression recognition (FER) systems within Large Multimodal Foundation Models (LMFMs). Despite advances in deep learning and the availability of diverse datasets, **FER systems often exhibit higher error rates for individuals with darker skin tones**. Existing research predominantly focuses on traditional FER models (CNNs, RNNs, ViTs), leaving a **gap in understanding racial biases in LMFMs**. We benchmark four leading LMFMs: **GPT 4o, PaliGemma, Gemini, and CLIP** to assess their performance in facial emotion detection across different racial demographics. A linear classifier trained on CLIP embeddings obtains accuracies of **95.9% for RADIATE, 90.3% for Tarr, and 99.5% for Chicago Face**. Furthermore, we identify that **Anger is misclassified as Disgust 2.1 times more often in Black Females than White Females**. This study highlights the need for fairer FER systems and establishes a foundation for developing unbiased, accurate FER technologies.

Introduction

- **Advancements in FER**
  - Increased real-world applications, including:
  - Assessing blue-collar workers' well-being
  - Detecting pain in non-verbal patients
  - Evaluating job candidates' expressions for intelligence, personality, and criminality
- **Challenges**
  - Pervasive racial bias with higher error rates for darker-skinned individuals
  - FER systems used to select candidates may disadvantage minorities
- **Research Limitations**
  - Limited studies on racial bias in Large Multimodal Foundation Models (LMFMs), which process text, image, and audio data
- **Objectives**
  - Benchmark four leading LMFMs for facial emotion detection across racial groups
  - Demonstrate self-supervised learning's effectiveness in reducing biases
  - Conduct statistical analyses to assess misclassification rates and racial biases

Datasets

- **RADIATE**
  - 1,721 images
  - Anger, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprise
  - Black, White, Hispanic, and Asian backgrounds
  - Removed Other and Calm category

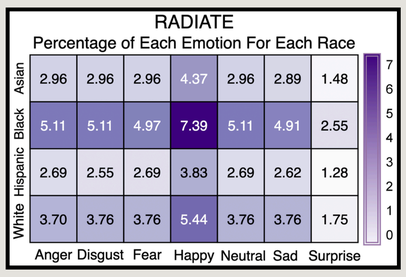


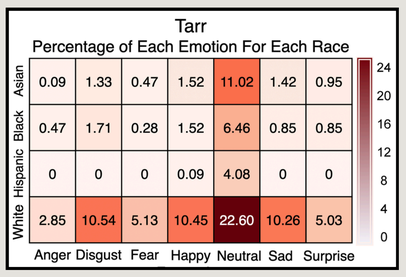
- **Tarr**
  - 1,226 images
  - Anger, Confusion, Disgust, Fear, Happy, Neutral, Sad, and Surprise
  - Black, White, Asian, and Hispanic backgrounds
  - Removed Multiracial and Confusion category

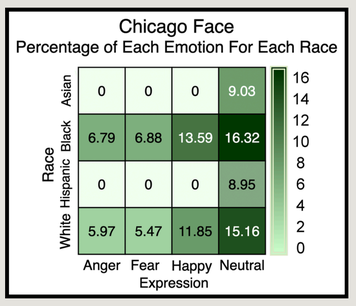


- **Chicago Face**
  - 1,208 images
  - Neutral, Happy, Anger, and Fear
  - Asian, Black, Latino, and White adults
  - Contains primarily Neutral expressions, with all Asian and Hispanic images showing this emotion









Methods

We **benchmark four models** to evaluate performance and detect racial biases in FER.

- **CLIP**
  - Utilized self-supervised learning with logistic regression applied to CLIP embeddings.
  - The embeddings were extracted from high-resolution input images, standardized to 224x224 pixels.
  - Numerical labels for expressions were used for training the logistic regression classifier, followed by decoding predictions back into emotion categories.
- **GPT-4o**
  - Used zero-shot prompting with Chain-of-Thought (CoT) reasoning to improve logical response generation.
  - Sampling settings: maj@5 (majority at 5 attempts) to mitigate hallucinations in emotion classification.
  - Prompt structure: "Classify the image into one of the eight expressions: [Happy, Anger, Sad, Neutral, Calm, Disgust, Fear, Surprise]."
- **Gemini**
  - Benchmarked Gemini 1.5 Flash and Gemini 1.5 Pro versions with zero-shot image classification.
  - Adjustments included reducing temperature values to 1.0 for consistent response output and enabling multi-prompt retry mechanisms for failed classifications.
- **PaliGemma**
  - Evaluated performance using three variants: pt-224, mix-224, and pt-448.
  - Lowercase tagging was used for prompts to reduce overfitting on specific features.
  - Sampling methods ensured comprehensive evaluation across dataset splits.

Results, Validation, and Analysis

- **Performance Overview**
  - Significant performance variations in zero-shot prompting accuracy across three datasets among leading Large Multimodal Foundation Models
  - The linear classifier on CLIP embeddings achieved the highest accuracy: RADIATE (95.9%), Tarr (90.3%), and Chicago Face (99.5%)
  - GPT-4o and GPT-4o with Chain-of-Thought (CoT) showed strong performance but did not consistently outperform when evaluated with ma@5 statistics
  - PaliGemma mix-224 had comparatively lower accuracy compared to other LMFMs
- **Model Performance by Race and Emotion**
  - CLIP outperformed GPT-4o and Gemini 1.5 Pro across all racial categories, with accuracies between 94.2% and 97.4% compared to:
    - GPT-4o: 76.2% to 84.5%
    - Gemini 1.5 Pro: 74.1% to 83.4%
  - CLIP excelled in recognizing emotions such as Happy and Neutral, achieving near-perfect scores (98.1% to 100%)
  - CLIP misclassified emotions of Black females in RADIATE and Tarr 2.1 times more often than the next highest demographic, Asian males
- **Challenges in Emotion Recognition**
  - GPT-4o and Gemini 1.5 Pro exhibited variable performance, especially struggling with recognizing Fear and Sadness, particularly among Asian and White subjects
    - Recognition rates as low as 16.3% for Fear and 27.6% for Sadness
  - Emotions like Happy and Surprise showed higher recognition rates across all models and racial groups, indicating their relatively straightforward identification compared to more nuanced emotions
- **Statistical Analysis of Misclassification**
  - Positive z-scores indicate more misclassifications for Asian demographics (Asian vs. Black: 2.25; Asian vs. Hispanic: 2.55)
  - Negative z-scores indicate more misclassifications for Black demographics (Black vs. White: -2.57; Hispanic vs. White: -2.70)

Table 1: Model (Temperature 1) and Method Comparison for Zero-shot Prompting Dataset Accuracy

Model and Zero-Shot Prompting Method	RADIATE		Tarr		Chicago Face	
	single	maj@5	single	maj@5	single	maj@5
GPT-4o	70.9	70.5	71.5	**71.7**	95.6	95.4
GPT-4o with CoT	70.5	**71.0**	70.0	70.8	94.7	**95.9**
PaliGemma mix-224	45.7	N/A	43.3	N/A	73.7	N/A
Gemini 1.5 Flash	67.2	N/A	68.3	N/A	92.9	N/A
Gemini 1.5 Pro	70.9	N/A	68.0	N/A	94.5	N/A
Linear Classifier on CLIP Embeddings	**95.9**	N/A	**90.3**	N/A	**99.5**	N/A

Table 2: GPT-4o, Gemini 1.5 Pro, and CLIP Overall Accuracy By Emotion and Race

	GPT-4o				Gemini 1.5 Pro				CLIP			
	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White	Asian	Black	Hispanic	White
Anger	57.8	77.9	70	68.2	44.4	73.6	60	56.1	91.1	96.9	92.5	90.4
Disgust	72.4	89.4	97.4	85.0	81.0	83.0	92.1	73.1	84.5	88.3	89.5	95.8
Fear	16.3	57.5	40.0	44.9	36.7	63.1	47.5	49.4	91.8	95.0	100	85.8
Happy	88.9	87.9	87.9	92.0	91.4	87.6	91.4	93.1	100	99.3	100	99.7
Neutral	93.7	95.9	97.4	93.9	97.4	96.5	99.0	95.6	100	99.7	99.5	98.1
Sad	46.6	36.7	56.4	43.9	27.6	30.5	46.2	32.3	82.8	90.2	92.3	88.4
Surprise	75.0	83.0	100	55.7	75.0	78.7	94.7	43.8	93.8	89.4	100	82.3
Accuracy	76.2	81.1	84.5	77.1	77.9	80.2	83.4	74.1	95.1	96.6	97.4	94.2

- **Specific Misclassification Patterns**
  - For zero-shot PaliGemma, Neutral was mistaken as Sad for White females 3.38 times more often than for Black females across all datasets.
  - In zero-shot GPT-4o on the RADIATE dataset, Fear was misclassified as Surprise 2.46 times and Anger as Disgust 2.1 times more often for Black females than for White females.
  - No other significant statistical trends or patterns observed due to the overrepresentation of White images, which were more frequently misclassified.

Evaluation Metrics:

- **Accuracy Analysis**
  - Measured model performance across racial and emotional categories.
  - Focused on misclassification patterns to identify demographic-specific biases.
- **Statistical Testing**
  - Conducted two-proportion z-tests to compare misclassification rates between racial groups.
  - Calculated p-values and z-scores for significant disparities in emotion recognition.

Conclusion

Bias in Large Multimodal Foundation Models (LMFMs) presents significant challenges for real-world applications, particularly in critical fields such as healthcare, security, and social interaction analysis. The methodologies we proposed—such as leveraging self-supervised learning with CLIP embeddings and incorporating Chain-of-Thought reasoning for GPT-4o—demonstrated measurable improvements in reducing misclassification.

- CLIP reduced the overall racial bias gap, showing consistent accuracy across all racial groups, ranging from 94.2% to 97.4%.
- Statistical analyses (two-proportion z-tests) confirmed that models adjusted using our framework significantly decreased biases, such as fewer instances of Anger being misclassified as Disgust or Fear being misclassified as Surprise.

Our results highlight a clear path forward for creating fairer FER systems:

1. Scaling to Broader Demographics: Expanding datasets to include more balanced and nuanced representations of racial and emotional categories will further enhance model fairness and generalizability.
2. Dynamic Adaptation: Developing adaptive AI systems that account for cultural and contextual nuances will reduce biases in real-time applications.
3. Fine-Tuning for Specific Use Cases: Tailoring LMFMs for domain-specific applications (e.g., healthcare or hiring) will address unique challenges, such as detecting pain in non-verbal patients or evaluating emotional responses in high-stakes environments.
4. Bias Detection and Mitigation Pipelines: Embedding bias detection frameworks into model training and deployment workflows will help identify and resolve disparities before they impact end users.

This study demonstrates that addressing biases in FER is not only possible but also critical for ensuring equitable outcomes in AI applications. Our ongoing work focuses on refining these models to achieve universal fairness and reliability. By building on the advancements presented here, future research can drive transformative improvements in AI systems, ultimately enabling inclusive and trustworthy technologies.

References

kvjvhub.github.io/FERRacialBias