EMONET-FACE: An Expert-Annotated Benchmark for Synthetic Emotion Recognition

Christoph Schuhmann*
LAION e.V.
christoph.schuhmann@laion.ai

Robert Kaczmarczyk*
LAION e.V.
Technical University of Munich

Gollam Rabby
L3S Research Center
Leibniz University of Hannover

Felix Friedrich
TU Darmstadt
Hessian.AI

Maurice Kraus
TU Darmstadt

Krishna Kalyan
LAION e.V.

Kourosh Nadi
LAION e.V.

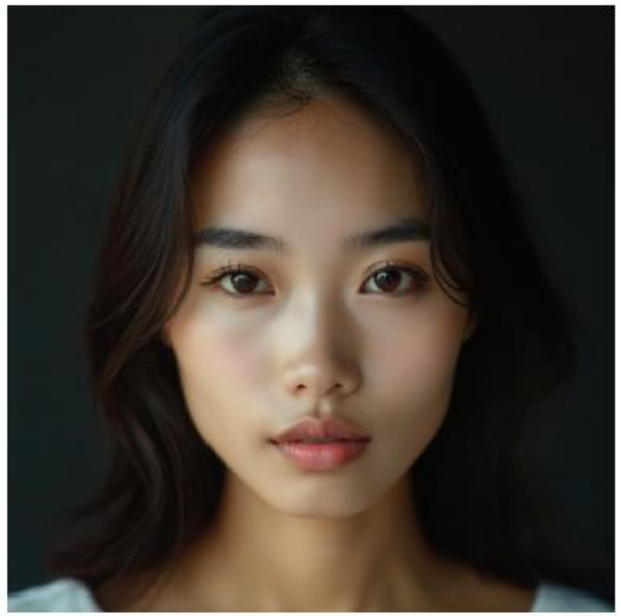
Huu Nguyen
Ontocord
LAION e.V.

Kristian Kersting
TU Darmstadt
Hessian.AI & DFKI

Sören Auer
TIB–Leibniz Information Centre for Science and Technology
L3S Research Center
Leibniz University of Hannover

(a) **Prompt:** An authentic, realistic closeup image of a Hispanic or Latino woman of 70 years who seems to experience triumph, superiority. Strong facial expression of triumph, superiority, [...]
(Midjourney v6)

(b) **Prompt:** A focused close-up photo of a Black or African American man of 30 years who seems to genuinely experience lighthearted fun, amusement, mirth, joviality, laughter, playfulness, silliness, [...]
(Flux Pro)



(c) **Prompt:** A closeup sharp, focused photo of a Southeast Asian woman of 40 years who seems to genuinely experience mild yearning, longing, pining, wistfulness, nostalgia, Craving, desire, Envy, [...]
(Flux Dev)

Figure 1: Samples from our EMONET-FACE datasets generated with different sota T2I models.

Four Key Contributions of EmoNet-Face

01

40-Category Emotion Taxonomy

We introduce a 40-category emotion taxonomy derived from foundational psychological research, capturing a broader and more detailed array of human emotional states beyond basic emotions.

02

Three Synthetic Datasets

We construct three large-scale synthetic datasets with controlled demographic balance across ethnicity, age, and gender, ensuring high-quality and diverse facial expressions for training and evaluation.

03

Expert Annotations

Our datasets feature multi-expert annotations, including both **binary labels** and **continuous ratings**, providing high-fidelity training and evaluation data for fine-grained emotion recognition.

04

Baseline Models

We develop **Empathic Insight-Face** models that achieve **human-expert level performance** on our benchmark, demonstrating the potential for AI systems to match human capabilities in emotion recognition.

Why We Need a Finer-Grained Emotion Benchmark

01

Limitations of Current Benchmarks

Current facial emotion recognition benchmarks are limited to 6–8 basic emotions, often using uncontrolled imagery with occluded faces and lacking demographic diversity, which risks significant bias and overlooks nuanced emotional states.

02

Motivation for Improvement

To enable AI systems to accurately interpret nonverbal emotional cues and generate legible synthetic emotions, a more comprehensive and diverse benchmark is needed to capture the full spectrum of human emotions, including subtle states like shame, embarrassment, and intoxication.

3 Datasets

EmoNet-Face Big

EmoNet-Face Big contains **203,201** synthetic images with **Gemini Flash 2.0 – generated labels**, designed for large-scale pre-training of emotion recognition models to capture a wide range of emotional expressions.

EmoNet-Face HQ

EmoNet-Face HQ comprises **2,500 images** with **continuous expert ratings** across all 40 emotions, serving as a high-quality evaluation benchmark to assess model performance on fine-grained emotion recognition.

EmoNet-Face Binary

EmoNet-Face Binary includes 19,999 images with over 62,000 expert binary labels, providing a fine-tuning dataset with high-quality annotations for training robust emotion recognition models.

Taxonomy Construction

Our 40-category emotion taxonomy extends beyond basic emotions to include social, cognitive, and bodily states such as elation, pride, teasing, shame, jealousy, pain, fatigue, numbness, and intoxication. This taxonomy is **grounded** in the **Handbook of Emotions** and refined through **expert consultation**.

Controlled T2I Pipeline & Expert Labeling Workflow

Data Generation

We use state-of-the-art text-to-image models like Flux Pro/Dev and Midjourney v6, along with carefully designed prompts to ensure demographic diversity and explicit full-face expressions, generating high-quality synthetic images for our datasets.

Expert Annotation

Psychology experts provide rigorous annotations, including binary labels and continuous ratings, ensuring high-quality and reliable emotion labels. Inter-annotator agreement is analyzed using Krippendorff's alpha to validate the annotation quality.

How We Benchmark Humans & Models

We evaluate model performance using correlation and agreement metrics, comparing against **human expert annotations on the EmoNet-Face HQ benchmark**. Metrics include **Spearman's rho** and **weighted kappa** to assess the alignment between model predictions and human judgments.

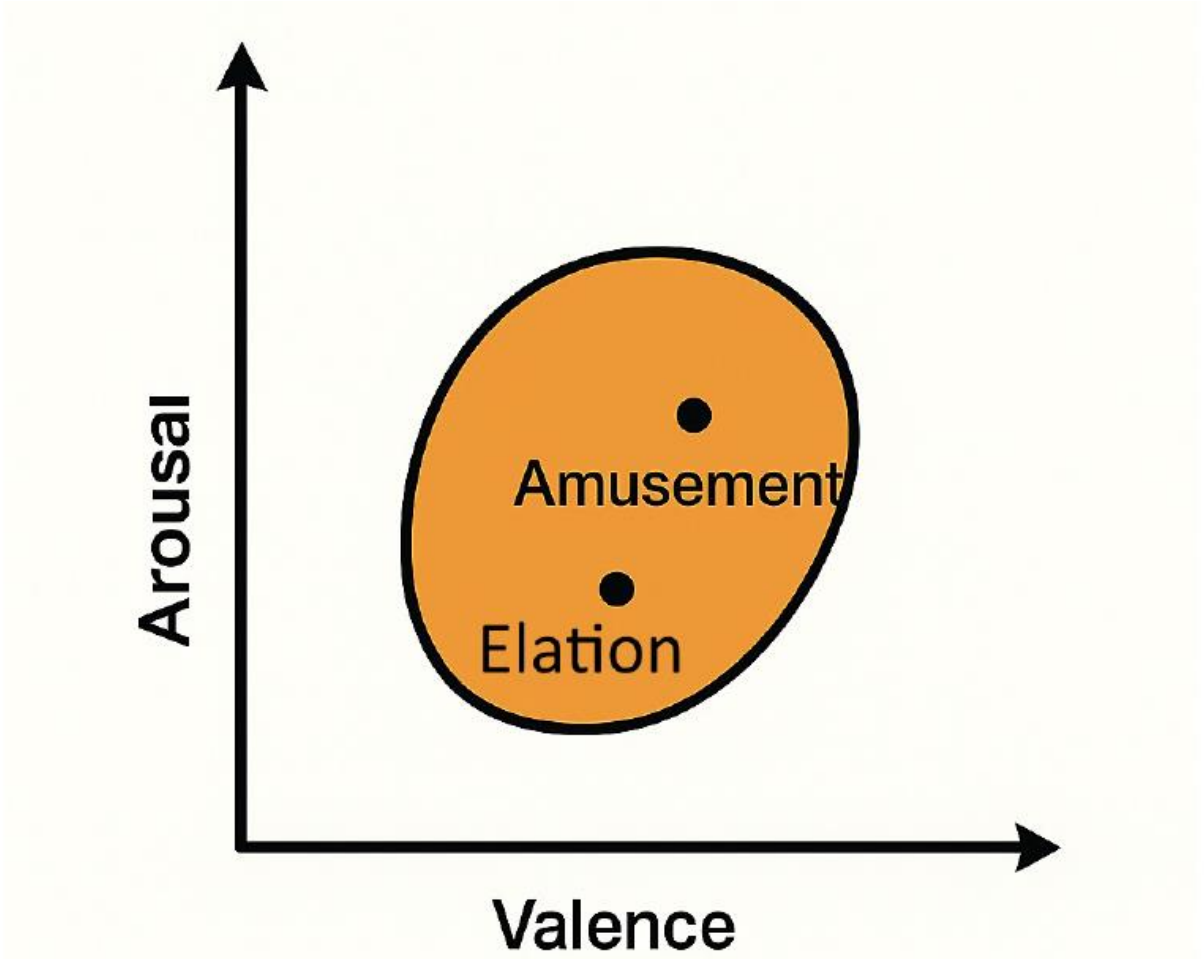
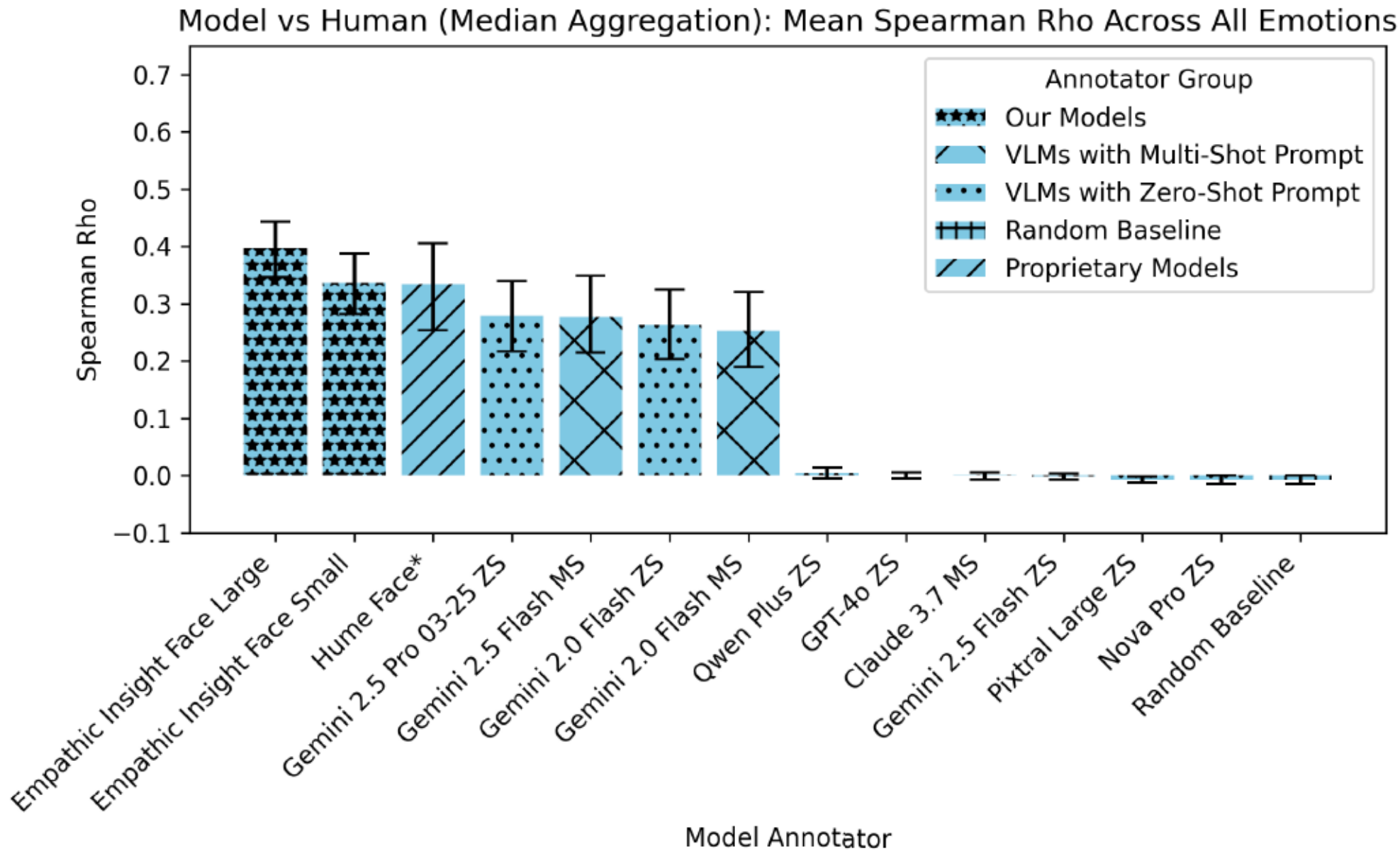


Figure 5: Rather than mapping each face to a single label, we estimate a distribution over plausible emotion categories.


Model vs Human (Median Aggregation): Mean Spearman Rho Across All Emotions
Spearman Rho
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
Annotator Group
Our Models
VLMs with Multi-Shot Prompt
VLMs with Zero-Shot Prompt
Random Baseline
Proprietary Models
Empathic Insight Face Large
Empathic Insight Face Small
Hume Face*
Gemini 2.5 Pro 03-25 ZS
Gemini 2.5 Flash MS
Gemini 2.0 Flash ZS
Gemini 2.0 Flash MS
Qwen Plus ZS
GPT-4o ZS
Claude 3.7 MS
Gemini 2.5 Flash ZS
Pixtral Large ZS
Nova Pro ZS
Random Baseline
Model Annotator

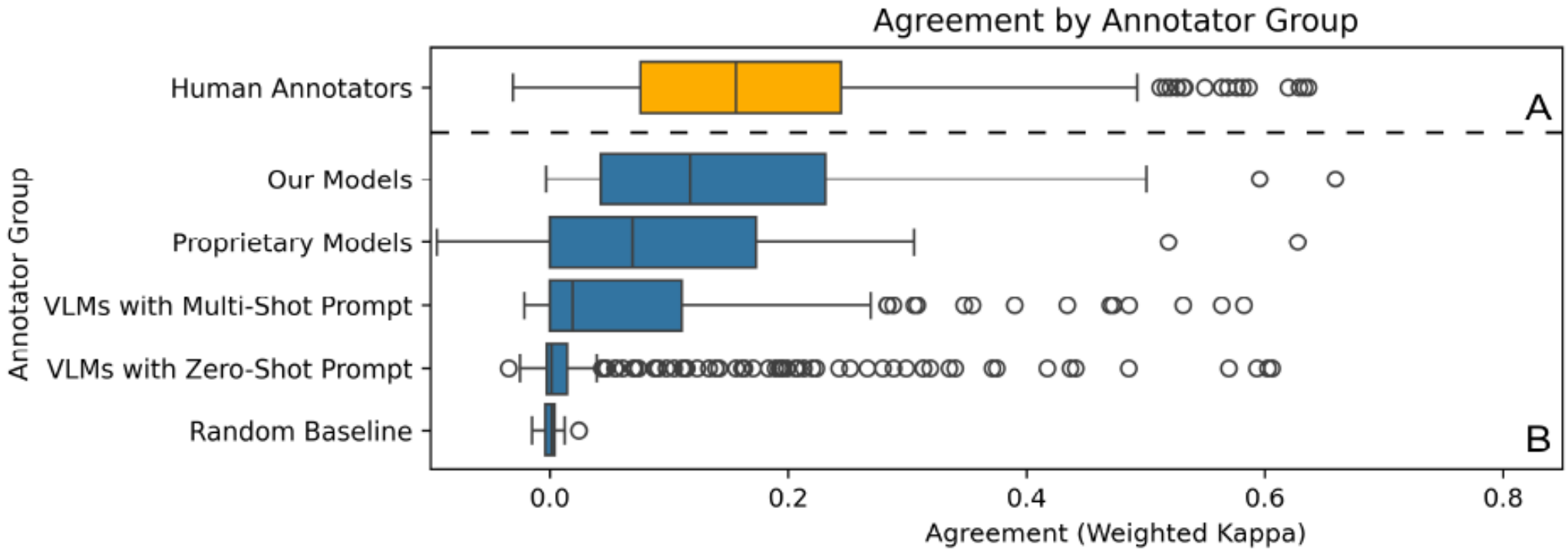


Figure 3: Weighted Kappa (κ_w) agreement scores by annotator group. **A** (top): Pairwise agreement between human annotators. **B** (below): Pairwise agreement between each human annotation and other sources, including 'Our Models' (EMPATHICINSIGHT-FACE), 'Proprietary Models' (HumeFace), 'VLMs (Multi-Shot and Zero-Shot Prompts)', and a 'Random Baseline'. Each box represents the interquartile range (IQR) of κ_w scores, with the median as center line.

Empathic Insight-Face Architecture & Training

EmpathicInsight-Face models use a **SIGLIP 2 400M** encoder with **40 regression heads** to predict emotion intensities on a 0–7 scale. The models are **pre-trained** on EmoNet-Face Big and **fine-tuned** on EmoNet-Face Binary, achieving human-expert level performance on the evaluation benchmark.

Future Goals: Multimodal Integration

Future work should integrate multimodal cues such as **speech**, **body language**, and **context** to enhance emotion recognition capabilities, moving towards more comprehensive and context-aware models.