JEONBUK NATIONAL UNIVERSITY

The 1st International Online Conference on Behavioral Sciences

EMOLAB.

Comparing GPT and Human Affective Evaluations of Social Images

Jongwan Kim (jongwankim80@jbnu.ac.kr)
Department of Psychology, Jeonbuk National University, South Korea

Introduction

Multimodal LLMs increasingly process emotional content, but standardized benchmarking against human affective norms is lacking. This study introduces a framework comparing model-generated emotion ratings with validated human normative data.

Methods

Dataset: ESISCA social image database (274 images).Human ratings: Valence & arousal (9-point SAM scale).GPT-4o produced eight ratings per image.

Analyses: valence-arousal structure, mean bias, variability, and confusion matrix.

Results: Valence–arousal distributions

GPT-4o broadly tracked human affective structure but systematically rated scenes as more positive and more arousing. Emotion clusters were more diffuse than human ratings.

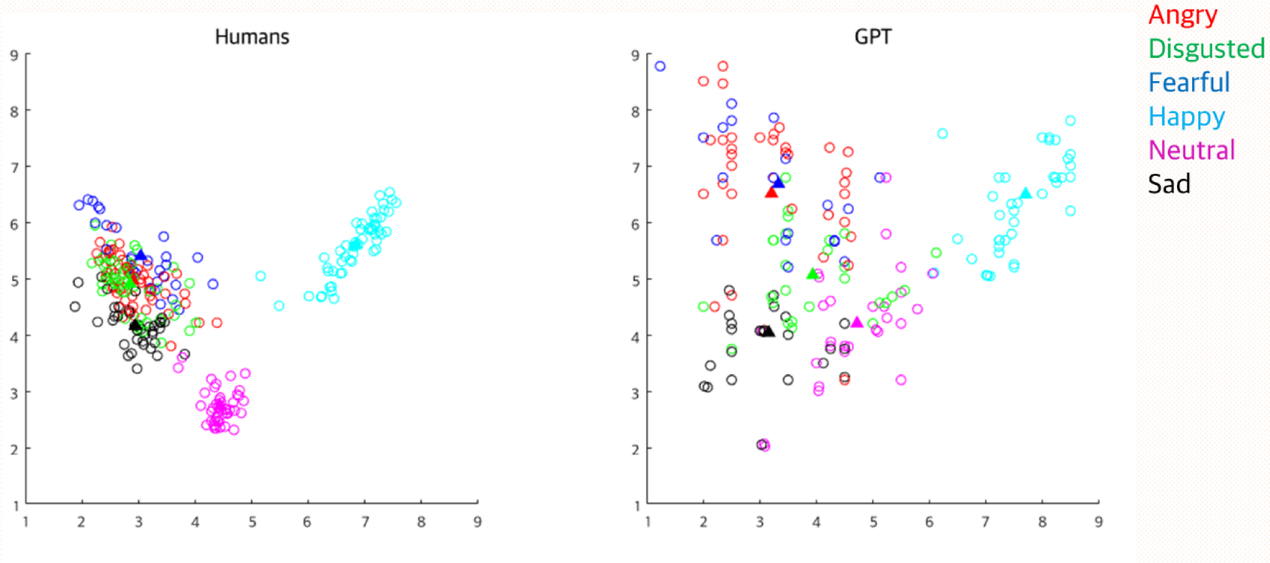

Figure 1. Valence–arousal distributions of human (left) and GPT-4o (right) ratings across six emotion categories from the Social Interaction Image Database (Zhang et al., 2024). Each point represents an individual image. GPT's ratings are shifted toward higher valence and arousal and show less distinct clustering among emotion categories.

Results: Emotion Classification

High agreement for prototypical emotions (Happy, Angry, Sad), but confusion among negative categories such as Fear and Disgust. Indicates coarse emotional differentiation.

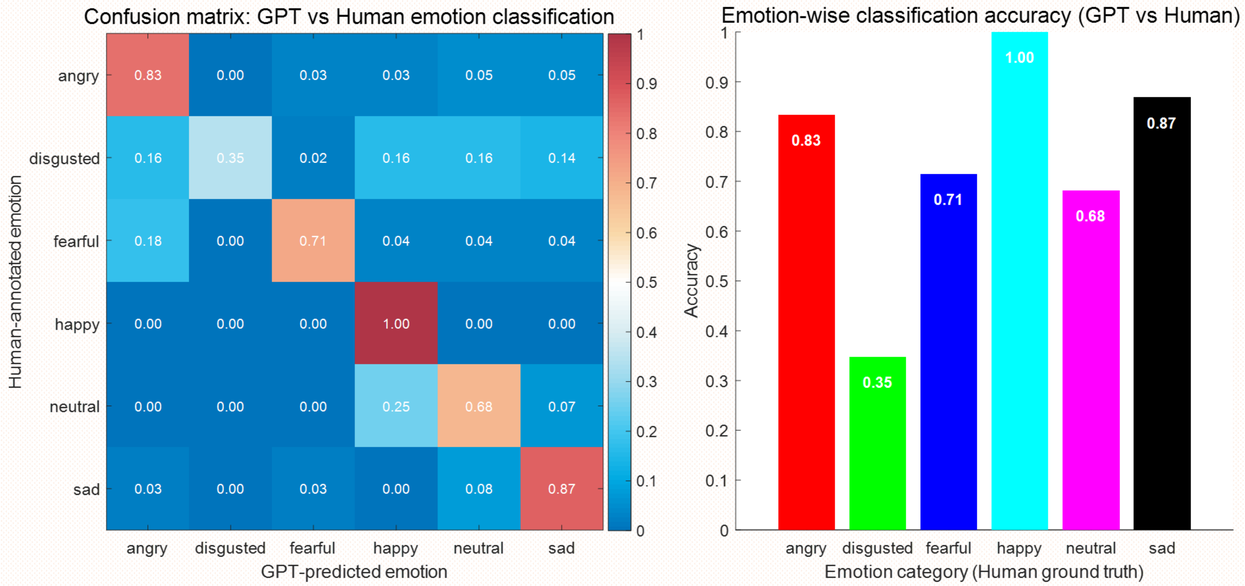

Figure 2. Confusion matrix (left) and emotion-wise classification accuracy (right) comparing human and GPT-4o emotion categorization for 274 social images. The confusion matrix displays the proportion of GPT-predicted emotions (columns) for each human-annotated category (rows), normalized by row. Higher values along the diagonal indicate stronger agreement between GPT and human judgments. The bar graph on the right shows per-emotion classification accuracy, calculated as the proportion of correct GPT predictions within each human-labeled category. Colors correspond to the six emotion categories (red = Angry, green = Disgusted, blue = Fearful, cyan = Happy, magenta = Neutral, black = Sad).

Discussion

Introduces a reproducible norm-referenced framework for benchmarking affective representations in multimodal LLMs across dimensional and categorical analyses.

References

Kim, J. (in press). Comparing ChatGPT and human ratings of affective images. Perception. https://doi.org/10.1177/03010066251391729
Park, C., & Kim, J. (2024). Exploring affective representations in emotional narratives: An exploratory study comparing ChatGPT and human responses. Cyberpsychology, Behavior, and Social Networking, 27(10), 736–741. https://doi.org/10.1089/cyber.2024.0100 X-MOL
Zhang, Z., Peng, Y., Jiang, Y., & Chen, T. (2024). The pictorial set of Emotional Social Interactive Scenarios between Chinese Adults (ESISCA): Development and validation. Behavior Research Methods, 56(3), 2581-2594.