npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01985-5

Evaluating the performance of general purpose large language models in identifying human facial emotions

Benjamin W. Nelson[1,2] ✉, Ari Winbush[3], Steven Siddals[1], Matthew Flathers[1], Nicholas B. Allen[3,4] & John Torous[1,4]

We evaluated the ability of three leading LLMs (GPT-4o, Gemini 2.0 Experimental, and Claude 3.5 Sonnet) to recognize human facial expression using the NimStim dataset. GPT and Gemini matched or exceeded human performance, especially for calm/neutral and surprise. All models showed strong agreement with ground truth, though fear was often misclassified. Findings underscore the growing socioemotional competence of LLMs and their potential for healthcare applications.

Generative artificial intelligence (GenAI) based on large language models (LLMs) is becoming central to human–computer interactions (HCIs), demonstrating impressive capabilities in interpreting human intentions[1,2] and understanding human cognitive, social, and emotional processes. Facial expressions are a key aspect of social–emotional functioning and provide valuable information about human goals, emotions, and psychological states[3].

LLMs have expanded their capabilities beyond traditional text-based tasks, enabling them to process and integrate multimodal inputs such as vision, speech, and text. They have shown promise in social cognition such as "theory of mind" tasks, sometimes matching or exceeding human performance on mentalistic inference[2]. However, these results are largely based on text-only examples and are less robust in assessments where context is critical[2,4,5]. Studies evaluating LLMs' visual emotion recognition have mixed results, with some models performing no better than chance[6].

GenAI's ability to interpret facial expressions holds promise for HCI applications, particularly in behavioral healthcare[7–9]. Subtle expression changes may indicate mental health conditions like depression, anxiety, or even suicidal ideation[10,11]. AI-powered systems trained to recognize these nuanced expressions could potentially enable earlier diagnosis, real-time monitoring, and adaptive interventions.

Facial expressions and interpretation can vary by culture[12] and context[13], highlighting the importance of using diverse stimuli with validated ground truth labels and normative human performance data. Moreover, the need to evaluate performance across diverse actors (i.e., sex/racial/ethnicity) is well recognized[6,14].

Results

Agreement

Cohen's Kappa (κ) across all stimuli and expressions was 0.83 (95% CI: 0.80–0.85) for ChatGPT 4o, 0.81 (95% CI: 0.77–0.84) for Gemini 2.0 Experimental, and 0.70 (95% CI: 0.67–0.74) for Claude 3.5 Sonnet. Specific Kappas by emotion class can be found in Table 1 and Fig. 1b.

Confusion matrix

Overall accuracy across all actors and expressions was 86% (95% CI: 84–89%) for ChatGPT 4o, 0.84% (95% CI: 81–87%) for Gemini 2.0 Experimental, and 74% (95% CI: 71–78%) for Claude 3.5 Sonnet. Accuracy by emotion class can be found in Table 1 and Fig. 1a. For ChatGPT 4o and Gemini 2.0 Experimental, there was little variability in the performance across different emotion categories, except for fear, which was misclassified as surprise 52.50% and 36.25% of the time, respectively (see Figs. 2a-b and 3a-b). For Claude 3.5 Sonnet, there was more variability in the performance across different emotion categories with sadness being misclassified as disgust 20.24% of the time and fear being misclassified as surprise 36.25% of the time (see Figs. 2c and 3c).

Lastly, there were no significant differences in model performance for accuracy, recall, or kappa based on the sex or race of the actor (see Table 2).

Discussion

This study evaluated three leading LLMs, ChatGPT 4o, Gemini 2.0 Experimental, and Claude 3.5 Sonnet, on facial emotion recognition using the NimStim dataset. ChatGPT 4o and Gemini 2.0 Experimental demonstrated "almost perfect"[15,16] agreement and high accuracy with ground truth labels overall, with ChatGPT 4o and Gemini 2.0 Experimental performance comparable to or exceeding human raters on some emotions. Claude 3.5 Sonnet exhibited lower overall agreement and accuracy as compared to the other two models.

There was significant variability in Cohen's Kappa and Recall within and between emotion classes. All models performed relatively well on Happy, Calm/Neutral, and Surprise, but showed difficulty recognizing Fear,

[1]Division of Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [2]Verily Life Sciences, San Francisco, CA, USA. [3]University of Oregon, Eugene, OR, USA. [4]These authors contributed equally: Nicholas B. Allen, John Torous.
✉e-mail: bnelson9@bidmc.harvard.edu

Table 1 | Description of validity ratings for LLM emotional expression estimates across all emotions

LLM	Emotion	Cohen's Kappa (95% CI)	Accuracy (95% CI)	Recall (95% CI)	Precision (95% CI)	F1 (95% CI)
ChatGPT 4o	Overall	0.83 (0.80–0.85)	0.86 (0.84–0.89)	0.85 (0.82–0.87)	0.86 (0.83–0.88)	0.83 (0.80–0.85)
	Angry	0.88 (0.82–0.93)		0.88 (0.81–0.95)	0.94 (0.88–0.99)	0.91 (0.87–0.95)
	Calm/Neutral	0.96 (0.95–0.99)		0.98 (0.96–1.00)	0.95 (0.92–0.98)	0.97 (0.95–0.98)
	Disgust	0.85 (0.78–0.92)		0.85 (0.78–0.93)	0.91 (0.84–0.97)	0.88 (0.83–0.93)
	Fear	0.45 (0.36–0.54)		0.42 (0.32–0.53)	0.87 (0.77–0.98)	0.57 (0.47–0.67)
	Happy	0.93 (0.90–0.96)		0.93 (0.88–0.97)	1.00 (1.00–1.00)	0.96 (0.94–0.99)
	Sad	0.84 (0.78–0.91)		0.87 (0.8–0.94)	0.88 (0.81–0.95)	0.87 (0.82–0.92)
	Surprise	0.91 (0.89–0.93)		1.00 (1.00–1.00)	0.45 (0.35–0.55)	0.62 (0.53–0.71)
Gemini Experimental 2.0	Overall	0.81 (0.77–0.84)	0.84 (0.81–0.87)	0.83 (0.8–0.85)	0.84 (0.82–0.87)	0.81 (0.79–0.84)
	Angry	0.76 (0.69–0.84)		0.76 (0.67–0.85)	0.93 (0.87–0.99)	0.84 (0.78–0.9)
	Calm/Neutral	0.90 (0.87–0.93)		0.95 (0.91–0.98)	0.86 (0.82–0.91)	0.9 (0.87–0.93)
	Disgust	0.87 (0.82–0.93)		0.89 (0.82–0.96)	0.85 (0.77–0.92)	0.87 (0.82–0.92)
	Fear	0.58 (0.49–0.69)		0.56 (0.45–0.67)	1 (1–1)	0.72 (0.63–0.81)
	Happy	0.94 (0.91–0.98)		0.94 (0.89–0.98)	1 (1–1)	0.97 (0.94–0.99)
	Sad	0.68 (0.60–0.77)		0.7 (0.6–0.8)	0.79 (0.69–0.88)	0.74 (0.67–0.81)
	Surprise	0.91 (0.89–0.94)		0.98 (0.93–1)	0.48 (0.38–0.58)	0.64 (0.55–0.73)
Claude 3.5 Sonnet	Overall	0.70 (0.67–0.74)	0.74 (0.71–0.78)	0.74 (0.7–0.77)	0.72 (0.69–0.75)	0.71 (0.69–0.74)
	Angry	0.86 (0.81–0.91)		0.88 (0.81–0.95)	0.83 (0.76–0.91)	0.86 (0.81–0.91)
	Calm/Neutral	0.69 (0.64–0.74)		0.71 (0.64–0.78)	0.92 (0.88–0.97)	0.8 (0.76–0.85)
	Disgust	0.70 (0.63–0.77)		0.73 (0.64–0.83)	0.72 (0.63–0.82)	0.73 (0.66–0.8)
	Fear	0.52 (0.44–0.61)		0.54 (0.43–0.65)	0.62 (0.51–0.74)	0.58 (0.5–0.66)
	Happy	0.88 (0.85–0.91)		0.91 (0.85–0.96)	0.88 (0.83–0.94)	0.89 (0.86–0.93)
	Sad	0.53 (0.45–0.63)		0.58 (0.48–0.69)	0.65 (0.55–0.76)	0.62 (0.54–0.69)
	Surprise	0.72 (0.63–0.81)		0.82 (0.71–0.93)	0.39 (0.29–0.49)	0.53 (0.44–0.62)

Cohen's Kappa = The agreement between ground truth and estimation; Accuracy = The proportion of correctly classified samples out of the total dataset; Recall = The fraction of samples belonging to a given emotion that the model correctly identifies as that emotion; Precision = The fraction of samples predicted as a given emotion that truly belong to that emotion; F1 = The harmonic mean of precision and recall, providing a single measure that balances both. At the class level, accuracy and recall are identical, therefore accuracy is only calculated across all classes overall.

often misclassifying it as Surprise. ChatGPT 4o achieved the best performance across emotions and significantly outperformed Claude 3.5 Sonnet on several emotions, including Calm/Neutral, Sad, Disgust, and Surprise. Gemini 2.0 Experimental also outperformed Claude 3.5 Sonnet for Calm/Neutral, Disgust, and Surprise. When comparing these models' performance to human observers in the NimStim dataset, the overall 95% confidence intervals for kappa overlapped for humans, ChatGPT, and Gemini, indicating similar levels of reliability across all emotion categories. In contrast, Claude's 95% CI did not overlap with that of humans, suggesting lower overall reliability. At the level of individual model-by-emotion comparisons, most 95% CIs overlapped; however, three exceptions emerged such that ChatGPT 4o showed higher reliability than humans for Surprise and Calm/Neutral, Gemini 2.0 Experimental outperformed humans for Surprise, and Claude 3.5 Sonnet was less reliable than humans for Calm/Neutral.

Literature has previously shown LLM biases, but current findings indicate that facial emotion recognition did not differ by sex or race. Furthermore, prior CNN models on this dataset achieved moderate classification performance (42% accuracy overall, with large emotion-specific variability[17]. In contrast, zero-shot vision-language models without training, fine-tuning, or architectural customization may offer stronger generalization.

Although these findings show promise for foundation models in affective computing, limitations remain. All stimuli featured static images[18], actors aged 21–30, and most images were European American, which may limit generalizability. The context of verbal signals can modify facial expression meaning, highlighting the need for future multimodal emotion classification with auditory stimuli[19]. Furthermore, although we selected the NimStim dataset because it is accessible only to researchers upon request and has not appeared in LLM publications, thereby minimizing the likelihood it was included in model training and positively biasing results, relying on a single dataset may limit the generalizability of our findings. While we tested three general-purpose models, specialized large models designed for facial expression and micro-expression recognition (e.g., ExpLLM, MELLM) are also available. Future research should evaluate these models on this dataset to compare their performance with general LLMs. Prompt wording varied slightly across models due to interface constraints, potentially affecting results. Specific healthcare applications may want to fine-tune models or incorporate the Facial Action Coding System into retrieval-augmented generation frameworks to improve recognition of more subtle or complex emotions, such as fear. Understanding when and why models succeed or fail will be critical for guiding responsible integration. Future research should evaluate open-weight models like Llama or DeepSeek, which can support more transparent evaluation, local deployment, and stronger privacy protections, important model considerations for clinical applications.

Overall, this study provides an initial benchmark for evaluating LLMs' socioemotional capabilities. Although ChatGPT and Gemini demonstrated reliability comparable to human observers across emotion categories, caution is warranted when translating these findings and using general-purpose LLMs in applied settings, as Claude, by contrast, showed lower overall reliability. Further testing with ecologically valid, multimodal, and demographically diverse stimuli is essential to understand their limitations and potential.

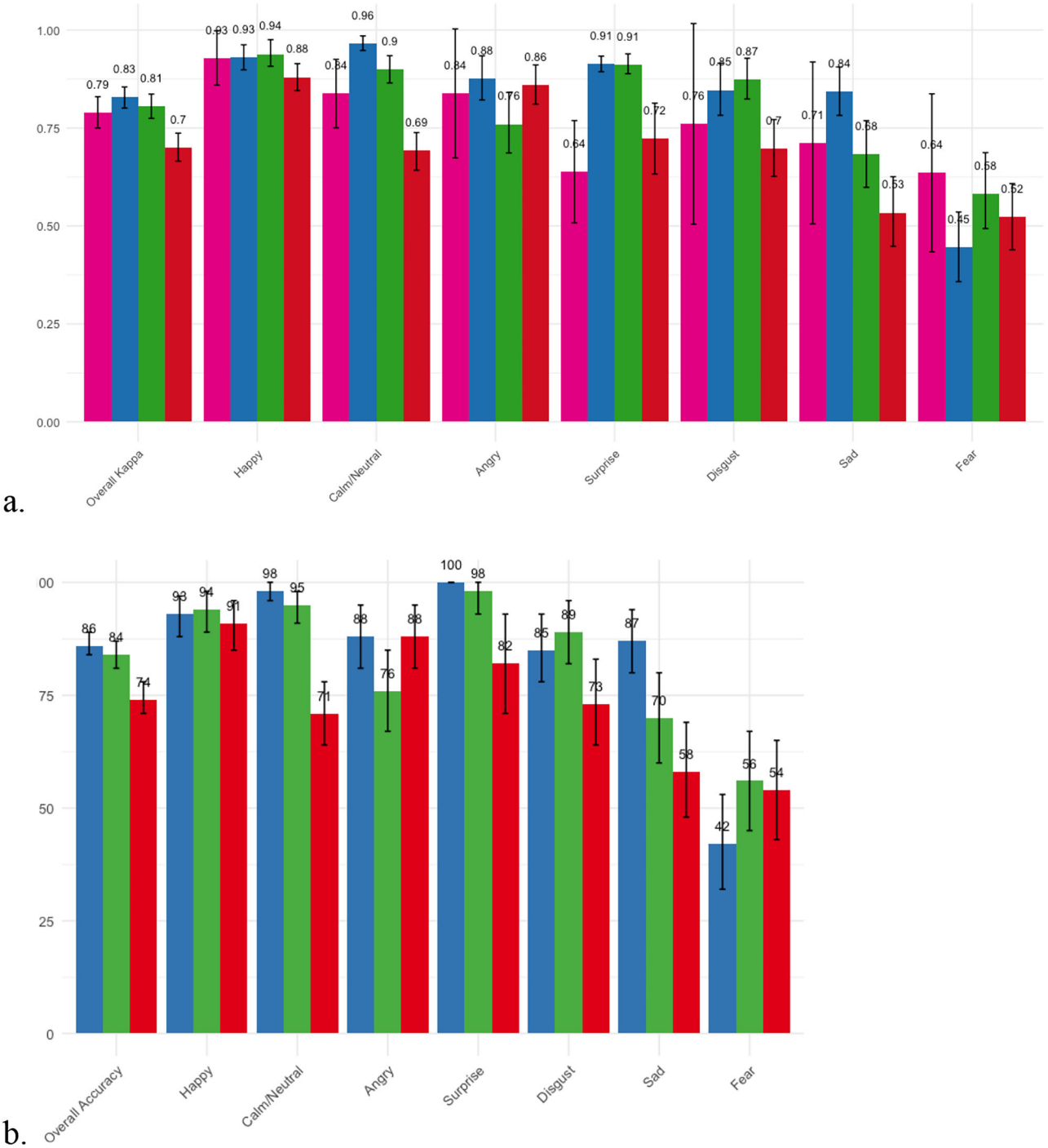

Fig. 1 | LLM model. a Agreement with NimStim human performance benchmark and **b** overall accuracy and recall by emotion class. Pink = NimStim Benchmark; Blue = ChatGPT 4o; Green = Gemini 2.0 Experimental; Red = Claude 3.5 Sonnet.

Methods

Study design

The current study was IRB-exempt from Beth Israel Deaconess Medical Center (2025P000198).

Facial expression stimuli. The NimStim, a large multiracial image dataset, was used as facial expression stimuli[15]. The NimStim Set of Facial Expressions is a comprehensive collection of 672 images depicting facial expressions posed by 43 professional actors (18 female, 25 male) aged between 21 and 30 years. The actors represent diverse racial backgrounds, including African-American (10 actors), Asian-American (6 actors), European-American (25 actors), and Latino-American (2 actors). Each actor portrays eight distinct emotional expressions: neutral, happy, sad, angry, surprised, fearful, disgusted, and calm. Psychometric evaluations with naive observers have demonstrated a high proportion correct at 0.81 (SD = 0.19; 95% CI: 0.77–0.85), high agreement between raters (kappa = 0.79, SD = 0.17; 95% CI = 0.75–0.83), and high test-retest reliability at 0.84 (SD = 0.08; 95% CI: 0.82–0.86)[15]. This dataset has been extensively utilized in various research studies with over 2000 citations[20–23]. The authors have obtained written consent to publish images of models #01, 03, 18, 21, 28, 40, and 45.

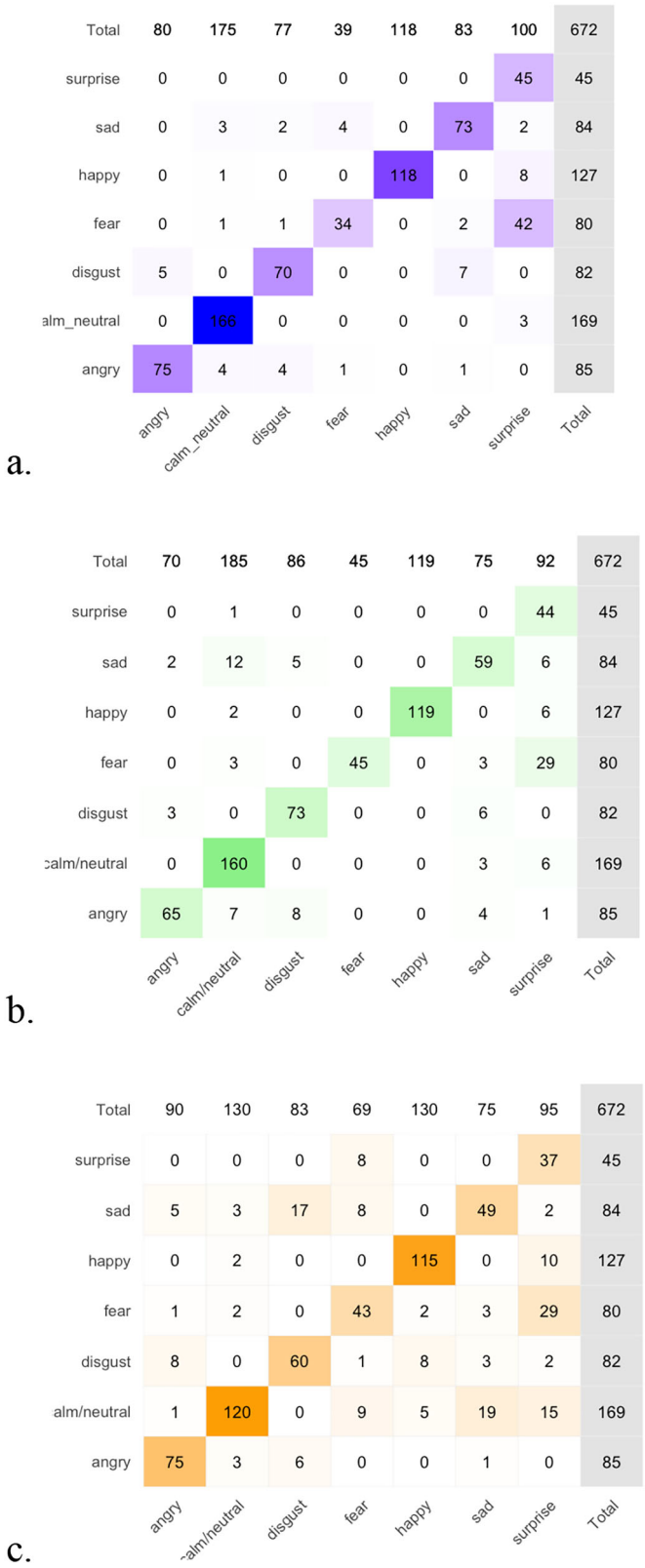

Fig. 2 | Confusion matrix. a ChatGPT 4o, **b** Gemini 2.0 Experimental, and **c** Claude 3.5 Sonnet.

The NimStim dataset provides an independent benchmark, as it is proprietary and restricted to authorized research institutions through licensing agreements that explicitly prohibit public distribution. Our verification process, including extensive web searches, found no public availability of the NimStim data, suggesting it was unlikely to have been included in LLM training datasets. NimStim calm and neutral expressions were recoded as *calm_neutral*, consistent with Tottenham et al.[15], who noted minimal perceptual differences between the two and treated either label as correct. Results separating calm and neutral are provided in the Supplementary Table 2.

Large language models

OpenAI GPT-4o Google Gemini 2.0 Experimental, and Anthropic Claude 3.5 Sonnet were used for facial expression recognition.

Procedures. All NimStim 672 images were individually uploaded twice to each LLM model for facial emotion processing using the user-facing interface, rather than the API, due to the fact that at the time of testing, only OpenAI offered the ability to batch multiple image inputs through the API for the selected models. Standardizing the methodology with the user interface ensured that the model's response remained grounded in the initial instruction. Prompts varied slightly across LLM models due to initial model responses indicating an inability to follow the prompt, likely due to built-in constraints and safety barriers (see Supplementary Table 1).

Analyses

All analyses were conducted with R v 4.3.1.

Agreement. We assessed agreement between each LLM model output and the ground truth label by calculating a stratified bootstrap analysis of Cohen's kappa (κ), to address repeated measures within participants and imbalances in emotion categories via oversampling. For each of 1000 bootstrap iterations, participants were sampled with replacement, and within-participant emotion categories were balanced via oversampling. We report mean κ and 95% confidence intervals and interpreted agreement using standard thresholds (moderate: 0.4–0.6, substantial: 0.6–0.8, and almost perfect: ≥0.8[15,16]. We applied the same oversampling bootstrap method to calculate κ for emotion class, sex, and race categories separately. Finally, we benchmarked model performance against κ values reported in the NimStim dataset by comparing 95% confidence interval overlap[15].

Confusion matrix, accuracy, recall, precision, and F1. To evaluate the classification performance of each LLM, we computed confusion matrices and derived standard metrics including accuracy, precision, recall, and F1-score for each model across emotion categories. The matrix quantifies the performance of the classification model by showing the count of samples for each combination of actual and predicted emotions, as well as the corresponding row and column totals to reflect the total occurrences of each actual emotion across the dataset, the number of times each emotion was predicted by the model represented by the diagonal elements, and a grand total representing the overall number of samples in the analysis. Note that the per-class balanced accuracy was equivalent to recall, a common metric in multi-class classification. Metrics were calculated per class and overall, with 95% confidence intervals estimated. κ and accuracy were also stratified by sex and race.

Methods of model comparison to NimStim

We benchmarked the performance of LLM models against the κ reported for untrained human observers in the NimStim dataset. However, it is important to note that the original authors did not specify how they calculated κ. Tottenham et al. [15] presented κ for each emotion by mouth state of mouth open and closed. To obtain a single κ estimate per emotion category to allow for comparability to results in the current study, we aggregated κ from the two mouth-states. First, κ and their associated standard deviations (SD) were extracted separately for mouth open and closed. The mean κ for each emotion was computed as the arithmetic average of the κ values from mouth-states. To account for variability across mouth-state conditions, we calculated the pooled SD using the square root of the mean of squared SD

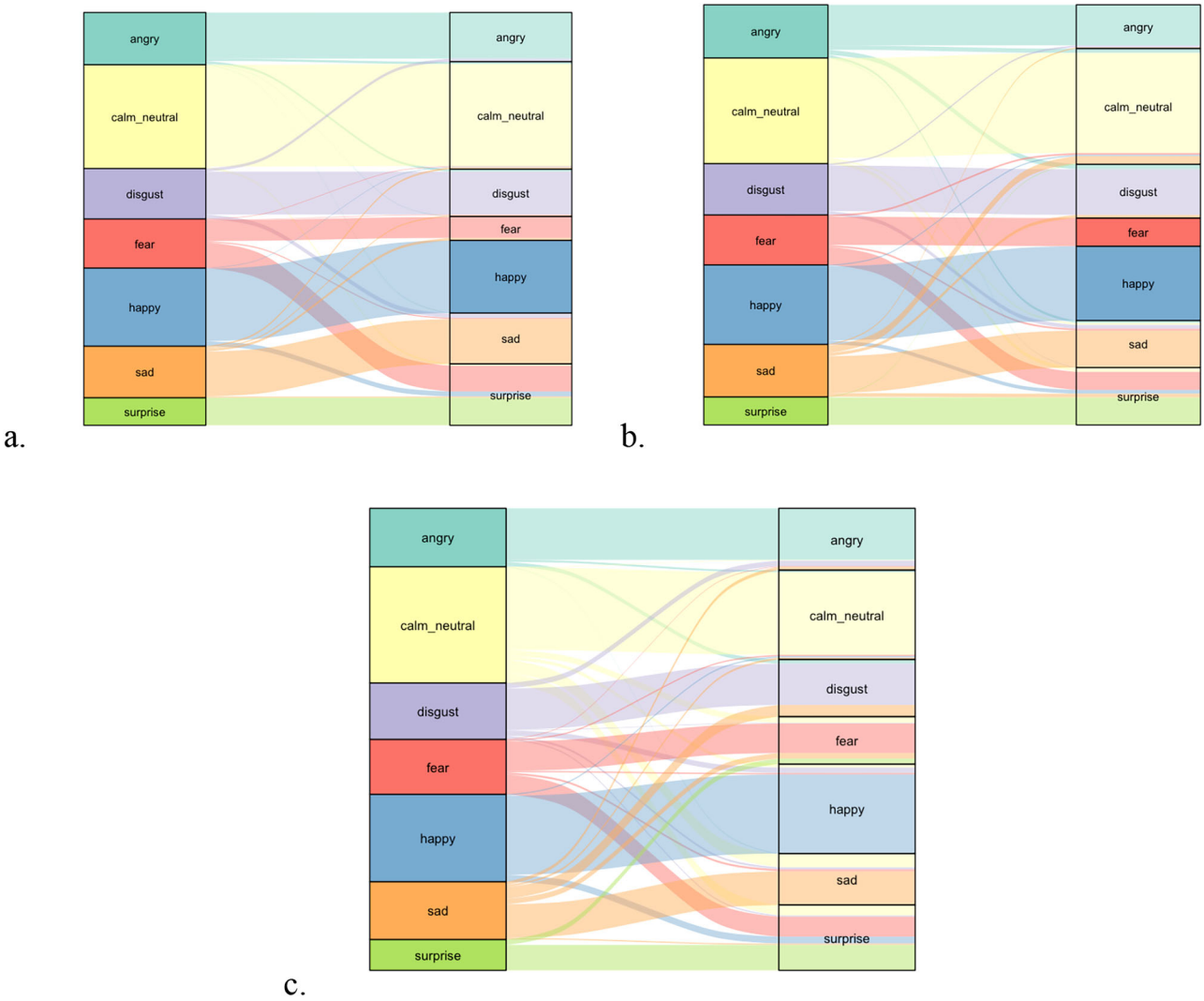

Fig. 3 | Alluvial plot. a ChatGPT 4o, **b** Gemini 2.0 Experimental, **c** Claude 3.5 Sonnet. Left column in ground truth and right column is model.

Table 2 | LLM agreement by sex and race

Model	Variable	Subgroup	Cohen's Kappa (95% CI)	Accuracy (95% CI)
ChatG-PT 4o	Sex	Male	0.82 (0.79–0.85)	0.86 (0.83–0.9)
		Female	0.83 (0.78–0.89)	0.87 (0.83–0.91)
	Race	European American	0.81 (0.77–0.85)	0.85 (0.81–0.88)
		African American	0.85 (0.79–0.90)	0.88 (0.83–0.93)
		Asian American	0.87 (0.82–0.92)	0.91 (0.86–0.97)
		Latino American	0.83 (0.67–0.96)	0.87 (0.75–0.99)
Gemini 2.0 Experi-mental	Sex	Male	0.79 (0.75–0.83)	0.83 (0.8–0.87)
		Female	0.82 (0.77–0.87)	0.85 (0.81–0.89)
	Race	European American	0.81 (0.77–0.86)	0.84 (0.8–0.87)
		African American	0.78 (0.73–0.84)	0.83 (0.77–0.89)
		Asian American	0.82 (0.71–0.92)	0.87 (0.8–0.94)
		Latino American	0.81 (0.58–0.96)	0.84 (0.71–0.97)
Claude 3.5 Sonnet	Sex	Male	0.72 (0.67–0.77)	0.77 (0.73–0.81)
		Female	0.67 (0.62–0.72)	0.71 (0.66–0.76)
	Race	European American	0.67 (0.62–0.72)	0.72 (0.67–0.76)
		African American	0.72 (0.65–0.80)	0.77 (0.7–0.83)
		Asian American	0.76 (0.65–0.85)	0.8 (0.72–0.88)
		Latino American	0.75 (0.62–0.88)	0.74 (0.59–0.9)

values, ensuring equal weighting across conditions. This approach provided a single, representative estimate of κ for each emotion while preserving the contributions from both facial configurations. Finally, to determine if the LLM models performed similarly, we assessed whether the 95% confidence intervals of these κ values overlap, indicating comparable (or different) levels of agreement.

Data availability

The NimStim data is available to researchers upon request at https://danlab.psychology.columbia.edu/content/nimstim-set-facial-expressions.

Code availability

All code is available on Open Science Framework at https://osf.io/dhkuy/.

Received: 13 May 2025; Accepted: 30 August 2025;
Published online: 16 October 2025

References

1. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl Acad. Sci. USA* **121**, e2405460121 (2024).
2. Strachan, J. W. A. et al. Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
3. Cohn, J. F. Foundations of human computing: facial expression and emotion. In *Proc. 8th International Conference on Multimodal Interfaces* 233–238 (ACM, 2006).
4. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at https://doi.org/10.48550/ARXIV.2302.08399 (2023).
5. Refoua, E. et al. The Next Frontier in Mindreading? Assessing Generative Artificial Intelligence (GAI)'s Social-Cognitive Capabilities using Dynamic Audiovisual Stimuli. *Comput. Hum. Behav. Rep.* 100702 (2025).
6. Elyoseph, Z. et al. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. *JMIR Ment. Health* **11**, e54369 (2024).
7. Feuerriegel, S. et al. Using natural language processing to analyse text data in behavioural science. *Nat. Rev. Psychol.* https://doi.org/10.1038/s44159-024-00392-z (2025).
8. Stade, E. C. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj Ment. Health Res.* **3**, 12 (2024).
9. Meskó, B. The impact of multimodal large language models on health care's future. *J. Med. Internet Res.* **25**, e52865 (2023).
10. Laksana, E., Baltrusaitis, T., Morency, L.-P. & Pestian, J. P. Investigating facial behavior indicators of suicidal ideation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* 770–777 (IEEE, 2017).
11. Girard, J. M. & Cohn, J. F. Automated audiovisual depression analysis. *Curr. Opin. Psychol.* **4**, 75–79 (2015).
12. Chen, C. et al. Cultural facial expressions dynamically convey emotion category and intensity information. *Curr. Biol.* **34**, 213–223.e5 (2024).
13. Durán, J. I. & Fernández-Dols, J.-M. Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. *Emotion* **21**, 1550–1569 (2021).
14. Ferrer, X., Nuenen, T. V., Such, J. M., Cote, M. & Criado, N. Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technol. Soc. Mag.* **40**, 72–80 (2021).
15. Tottenham, N. et al. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Res.* **168**, 242–249 (2009).
16. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
17. Sannasi, M. V., Kyritsis, M. & Gray, K. L. H. *What Does A Typical CNN "See" In An Emotional Facial Image?* https://doi.org/10.11159/mvml23.114 (2023).
18. Arsalidou, M., Morris, D. & Taylor, M. J. Converging evidence for the advantage of dynamic facial expressions. *Brain Topogr.* **24**, 149–163 (2011).
19. Tang, G., Xie, Y., Li, K., Liang, R. & Zhao, L. Multimodal emotion recognition from facial expression and speech based on feature fusion. *Multimed. Tools Appl.* **82**, 16359–16373 (2023).
20. Dawel, A., Miller, E. J., Horsburgh, A. & Ford, P. A systematic survey of face stimuli used in psychological research 2000–2020. *Behav. Res. Methods* **54**, 1889–1901 (2021).
21. Manelis, A. et al. Working memory updating in individuals with bipolar and unipolar depression: fMRI study. *Transl. Psychiatry* **12**, 441 (2022).
22. Fan, X. et al. Brain mechanisms underlying the emotion processing bias in treatment-resistant depression. *Nat. Ment. Health* **2**, 583–592 (2024).
23. Martens, M. A. G. et al. Acute neural effects of the mood stabiliser lamotrigine on emotional processing in healthy volunteers: a randomised control trial. *Transl. Psychiatry* **14**, 211 (2024).

Acknowledgements

No funding was granted for this study. We would like to thank Dr. Nim Tottenham for providing access to the NimStim dataset for research purposes.

Author contributions

B.W.N. conceptualized and drafted the manuscript. A.W. performed analyses. M.F. and S.S. edited the manuscript. N.A. and J.T. edited and reviewed the manuscript.

Competing interests

A.W., S.S., M.F. and J.T. declare no competing interests. B.W.N. is employed and has equity ownership in Verily Life Sciences. N.A. is employed and has equity ownership in Ksana Health.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01985-5.

Correspondence and requests for materials should be addressed to Benjamin W. Nelson.