Emotion Recognition Through GPT-4 Computer Vision Analysis of Facial Expressions

Greyson Shafiei

University of North Carolina at Charlotte

April 30, 2025

Abstract

Recent advancements in large language models (LLMs) have demonstrated strong capabilities in processing and reasoning about visual information; however, their effectiveness in recognizing facial expressions remains underexplored, particularly in areas such as orientation and group contexts. This study aims to investigate GPT-4's ability to interpret facial expressions and determine the emotional content under various conditions. Using the Facial Action Coding System (Ekman & Friesen, 1978), GPT-4 was tested on images of facial expressions presented in upright and inverted orientations, as well as individual and matrixed contexts. The dataset included individual images that were manually compiled into facial matrices to analyze ensemble perception. It was hypothesized that GPT-4 would perform better in upright, individually presented faces and less accurately with inverted or matrixed presentations. The findings revealed that GPT-4's emotion recognition is significantly impaired by face orientation and configuration, with upright individual images leading to the model's optimal performance. In contrast, inverted and matrixed images resulted in the model performing the poorest. In addition, providing explanatory prompts did not lead to any significant effect on accuracy, indicating there is limited alignment with human elaborative processing models. Machine learning analyses, including Random Forest Classification and Principal Component Analysis (PCA), verified a non-random structured pattern in GPT-4's responses. These results emphasize the model's sensitivity to visual layout and perceptual complexity, suggesting that GPT-4's factual emotion processing partially mimics human perceptual mechanisms, although there are notable limitations in configural and ensemble processing.

Keywords: large language models, emotion recognition, GPT-4

LLMs and Facial Expression Interpretation

In recent years, there has been an explosion of Large Language Models (LLMs). One that has taken a significant stake in LLM usage is ChatGPT. One of the recent models OpenAI released was GPT-4, which introduced the ability to scan images and extract data from the contents of the photo submitted. LLMs like GPT-4 have demonstrated impressive capabilities in computer vision, including object recognition, contextual reasoning, and visual data interpretation. According to a technical report completed by Achiam et al. (2024), when GPT-4 was fed 3 photos of an unusual charging cord and prompted, “What is funny about this image? Describe it panel by panel”, GPT-4 was able to decipher that the object in the first image as a VGA connector, an incorrect cable type, being plugged into an iPhone. In the second image, it was able to scan the manufacturer's package that initially contained the cable and read that the cable was labeled as a “Lighting Cable” adapter with a picture of a VGA connector on it. In the last photo, GPT-4 could then detect the object as a close-up of a VGA connector with a Lightning connector at the tip. GPT-4 then compiled the information from the three images together to detect the humorous element of the image due to using a large, outdated VGA connector in a modern smartphone that uses a smaller charging port. The example used in the technical report highlights GPT-4’s ability to use reasoning and abstract thinking to connect the data extracted from the different photos as recognizable objects with unusual details that were atypically connected, which could be perceived as humorous, properly satisfying all elements of the chat request.

LLM Processing

Chat-GPT, like many of these LLMs, identifies and processes elements based on statistical patterns learned from massive data sets. GPT-4 must first process the image into patterned features using vision transformers, which are a form of neural network (Lahat et al., 2024). These transformers break the image down into patches and convert them into embeddings, which are numerical representations of the data. These embeddings are then processed by the transformer architecture, where the data is sent through hidden layers of the neural network, and the activation of the different nodes is compared along with the learned weights to create an output based on these influences (Lahat et al., 2024). The process can be argued to be a form of cognition due to the complexity of pattern recognition and decision-making in deep learning algorithms used to compare data in the hidden layers to the output layer, which is then used to create a final decision. It is important to note that these machines do not possess the skills of cognition that humans have, and they are statistical systems that can perform complex tasks that people can. LLMs do not have consciousness or real cognitive abilities, and the use of cognition should be considered as a way to represent the sophisticated process these models use to break down input to create output.

Facial Expression Interpretation Into Basic Emotions

A landmark study of cognitive psychology was completed by Ekman & Friesen in 1978, where they established how humans express basic emotions through facial expressions and labeled their system the Facial Action Coding System. In this system, they established the facial movements that described the six basic emotions. Another significant study showed that understanding facial expressions becomes difficult when faces are flipped upside down

(McKelvie, 1995). The survey participants became significantly less accurate when asked to identify the emotions of inverted faces than right-side-up faces. McKelvie was able to decipher that there are two features that facial recognition relies on: Configural Processing and Componential Processing. Configurational processing refers to perceiving the overall layout and relationship between the different facial features. Componential processing is defined as the analysis of separate features, such as mouth shape or eyes, and how they are expressed (McKelvie, 1995). The study showed that people tend to rely on more componential processing when faces are inverted. A vital machine learning study by Marian et al. (1999) measured facial expressions by using computational image analysis. The study resulted in an automated facial expression recognition algorithm that achieved a high success rate at recognizing basic emotions. The computer was trained to recognize facial feature shapes for a decision-based output. This is similar to the componential processing that McKelvie referred to in their work. The computer analysis, however, struggled with minor changes in facial expressions where the expression became more complex or subtle. The study showed the incredible ability of machine learning to process facial expressions while still showing that such systems did not reach human levels of performance. Since the survey in 1999, machine learning techniques and processes have become more efficient and can complete more complex tasks.

Ensemble Representation in Facial Perception

Ensemble representation is a fundamental cognitive mechanism in human visual perception, allowing individuals to create a summary from a group of objects or faces instead of processing each of the several elements individually. Bayne & McClelland (2018) argue that ensemble properties, for instance, the mean of the emotion of a group of faces, are directly encoded in visual experience, which allows for rapid recognition of trends in large-scale visual stimuli. Research suggests that when humans view multiple faces simultaneously, they automatically form an impression of the average emotional expression rather than focusing on each individual face (Haberman et al., 2015). This ability is crucial for social perception and may have unique implications for machine vision models, where the model processes groups of facial expressions differently than humans.

The present study examines whether GPT-4 exhibits similar ensemble perception abilities by testing the performance in recognizing emotions from a matrix of faces. If GPT-4 has difficulties with detecting group emotions but performs well on individual emotions, it could be an indication that its vision processing system lacks ensemble coding mechanisms. This is relevant when comparing its accuracy across individual upright faces, inverted faces, and matrixed presentations, as ensemble perception theories predict that facial recognition should decline under conditions that interrupt typical configurational processing.

GPT-4 and Facial Recognition Capabilities

LLMs like GPT-4 have demonstrated new multimodal abilities, including facial image processing, but they remain fundamentally different from the current traditional biometric recognition systems. A study completed by DeAndres-Tame et al. (2024) evaluated GPT-4's capacity for face verification and soft biometrics estimation, such as age, gender, ethnicity, and explainability of results. They found that while GPT-4 can analyze facial features, it does not function as a true facial recognition system. Instead, GPT-4 relies on descriptive inference rather than biometric matching, which means that its performance on tasks like face inversion or matrixed faces may be influenced by contextual cues rather than by actual facial similarity computations.

Further technical testing by Iyer et al. (2024) explored the integration of real-time analysis using facial expressions into GPT-based conversational models, exemplifying that AI can use facial emotion recognition to adapt to the user by curating responses based on user emotions. This suggests that GPT-4 has the ability to process emotional cues from faces, but its capabilities are held back due to the limitation of a dedicated biometric framework. In the context of this study, it is key to determine if GPT-4's facial emotion recognition is affected by orientation and complexity or if the performance remains constant throughout the varying conditions.

This research expands on the prior findings by investigating whether GPT-4's multimodal vision capabilities align more with ensemble representation theories in human perception or if its performance is hindered by its reliance on text-based processing strategies. In the present study, GPT-4 exhibited significant variability under different conditions, including face orientation, emotion type, and presentation format. Inverted faces and matrixed faces resulted in a prominent decline in accuracy. Machine learning techniques, including Random Forest Classification and Principal Component Analysis (PCA), were utilized.

Present Study

GPT-4 demonstrates improved capabilities compared to the simplistic model used by Marian et al. (1999). The enormous database that GPT-4 was trained on provides the promise of being capable of processing facial expressions into emotions. The study by Marian et al. (1999) did not test the ability of the algorithm to process and accurately output emotions based on inverted facial expressions. Given the increased complexity of GPT-4 compared to the model of the previous study, there was no testing of the model's ability to average the mood of the faces shown, as well as the model's ability to decipher moods from a matrix of faces shown at one time.

The present study aims to test the ability of GPT-4 computer vision processing to interpret and translate facial expressions into emotions in different formats of visual data. This will expand upon the prior research in AI and psychology. In this study, GPT-4 was tested under different manipulated conditions: face orientation (upright vs. inverted), presentation style (individual vs. matrixed), and prompt style (with and without explanation). Classical statistical tests were leveraged to ascertain the accuracy of the model's predictions across the conditions, and then further explored using machine learning methods (Random Forest Regression and PCA), to determine if any specific patterns emerged from prompts led to predictive patterns in GPT-4's responses. These techniques revealed patterns in GPT-4's recognition behavior, suggesting that the outputs were structured and not likely to be random chance, but methodically influenced by the visual and contextual features. The findings reinforce the notion that GPT-4's visual reasoning capabilities are sensitive to perceptual complexity and layout, aligning comparatively with ensemble theories while exposing limits in configural processing.

Hypothesis 1: GPT-4's ability to correctly identify facial expressions as emotions will not differ significantly from the human baseline accuracy of the dataset, at 82.35% (Olszanowski et al., 2015).

Hypothesis 2: GPT-4 will be able to categorize facial expressions into emotions at a similar success rate of inverted faces as normal-oriented faces.

Hypothesis 3: GPT-4 will have reduced accuracy in interpreting faces into emotions in a matrix format for both image orientations.

Hypothesis 4: GPT-4 will have a low success rate in averaging facial expressions, regardless of orientation.

Hypothesis 5: There is no significant accuracy difference between prompts with and without explanations.

Methods

Materials

A sample of facial expression images was obtained from the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP) developed by Olszanowski et al. (2015). This validated dataset includes 210 photographs from 30 individuals (14 men and 16 women), each portraying one of the six basic emotions: happiness, surprise, fear, sadness, anger, and disgust, as well as neutral expressions. The images were accessed through the Open Science Framework (Olszanowski et al., 2015), and the set is widely used in emotion research due to its controlled visual properties and diverse representation of affect.

The experiment used GPT-4, an LLM with computer vision capabilities for processing images, provided by OpenAI. GPT-4's multimodal functionality enabled it to visually analyze the facial features and infer their emotional states using pre-trained knowledge and pattern recognition.

The complete set of individual facial expressions from the WSEFEP dataset was used to test GPT-4's ability to classify discrete emotional expressions. These images were presented in both upright and inverted orientations. The six basic emotions used in this test align with those identified initially by Ekman and Friesen (1978) in the Facial Action Coding System. This setup provided a baseline for evaluating GPT-4's accuracy on single-image emotion recognition. They were then compared to the overall average agreement of the facial expressions recorded by Olszanowski et al. (2015), which served as the average human baseline accuracy.

To assess ensemble perception, selected individual faces from the dataset were manually composed into matrices consisting of 16 individual facial expressions displayed simultaneously. Each matrix was designed to include varied emotional compositions; however, each matrix was comprised of a single majority expression (e.g., a majority of joyful faces with a mix of other expressions), enabling GPT-4 to either recognize each face individually or estimate the average emotion across the group.

All code used to preprocess images, interact with GPT-4, and conduct the statistical and machine learning analyses is available at GitHub.

Design

The study implemented a within-subjects experimental design with six primary conditions: (a) individual upright faces, (b) individual inverted faces, (c) matrixed upright faces, (d) matrixed inverted faces, (e) explanation of selection, and (f) no explanation of selection. Each condition tested GPT-4's performance under constraints of perception to evaluate the model's resilience in recognizing emotions. The use of the WSEFEP dataset ensured that emotional expressions were standardized across all conditions. The inclusion of inverted faces was instructed by McKelvie's (1995) work, which highlighted a decline in human facial recognition accuracy under inversion due to disrupted configural processing.

Procedure

1. **Image Preprocessing**: All images were formatted for compatibility with GPT-4's vision input requirements. For matrix conditions, the images were created using grid layouts to obtain realistic group-viewing contexts.
2. **Input to GPT-4**: The processed images were submitted to GPT-4 via Python scripts using the OpenAI API. Each submission included a structured prompt to identify the emotional state of the person or people in the image. The prompt for individual faces asked GPT-4 to identify the specific emotion the person in the picture is experiencing. For the matrix images, the prompt asked GPT-4 to identify the average emotion the group is experiencing.
3. **Orientation Testing**: Each image was tested in both upright and inverted formats, based on methods validated by McKelvie in 1995, to assess the true impact of facial rotation on GPT-4's recognition accuracy.
4. **Explanation Testing**: Each image in both orientations and image type (individual vs. matrixed) was tested using two different prompts. One prompt asked GPT-4 to provide only the emotion, while the other asked it to give the emotion and an explanation of why it selected that emotion.

Plan of analysis

To evaluate the multiple hypotheses, quantitative analyses were conducted to examine the different accuracy rates of GPT-4 in recognizing emotions from images of facial expressions across various orientations and configurations. The data were analyzed using Python and Scikit-learn, which includes machine learning and inferential statistics, to determine the association between these variables. Inferential statistics were utilized to test the hypotheses, encompassing t-tests and z-tests, to determine significant differences in GPT-4's performance across the six conditions.

Results

1. Hypothesis 1: A one-sample t-test comparing GPT-4's accuracy against the human benchmark of 82.35% showed a significant difference ($t = -14.026$, $p < .001$), indicating GPT-4 performed significantly worse than the human average overall.
2. Hypothesis 2: A paired t-test comparing upright and inverted faces found a significant difference ($t = 9.96$, $p < .001$), confirming that inverted orientation significantly impairs GPT-4's recognition ability. (Figure A3)
3. Hypothesis 3: An independent samples t-test comparing matrixed vs. individual faces also revealed a non-significant difference ($t = 1.91$, $p = .07$), which did not support the hypothesis that matrixed presentations decrease accuracy. But there is a notable difference between the averages.(Table 1)
4. Hypothesis 4:A one-proportion z-test on matrixed face performance versus chance (14.3%) revealed that GPT-4 performed significantly above chance ($z = 4.899$, $p < .001$), showing ensemble perception capability to a degree.
5. Hypothesis 5: A paired t-test between prompts with vs. without explanations yielded no significant difference ($t = -0.962$, $p = .3365$), suggesting explanation prompts do not meaningfully improve accuracy.

Comparison of Conditions

GPT-4's recognition accuracy varied significantly across orientations and presentation formats. A paired-samples t-test showed higher performance for upright faces ($M = 0.74$, 95% CI [0.69, 0.80]) than for inverted faces ($M = 0.45$, 95% CI [0.39, 0.51]), $t(227) = 9.96$, $p < .001$. (Table 1)

An independent-samples t-test also revealed that individually presented faces ($M = 0.61$, 95 % CI [0.56, 0.67]) outperformed matrixed faces ($M = 0.41$, 95 % CI [0.20, 0.62]); however, this difference did not reach conventional significance, Welch $t(21.74) = 1.91$, $p = .07$, $d = 0.51$, this is likely due to the small sample size for the matrix faces. (Table 1)

In contrast, a paired-sample t-test comparing prompts with explanations ($M = 0.59$, 95% CI [0.54, 0.64]) and without explanations ($M = 0.60$, 95% CI [0.55, 0.65]) showed no significant difference, $t(227) = -0.96$, $p = .34$. (Table 1)

Table 1

Summary of Recognition Accuracy by Experimental Condition

Comparison	Group 1 Mean [95% CI]	Group 2 Mean [95% CI]	t(df)	p-value	Cohen's *d*	Statistical Test
Upright vs. Inverted	0.74 [0.69, 0.80]	0.45 [0.39, 0.51]	$t(227) = 9.96$	< .001	—	Paired *t*-test
Individual vs. Matrixed	0.61 [0.56, 0.67]	0.41 [0.20, 0.62]	Welch $t(21.74) = 1.91$	.07	0.51	Independent *t*-test
Explanation vs. No Explanation	0.59 [0.54, 0.64]	0.60 [0.55, 0.65]	$t(227) = -0.96$	.34	—	Paired *t*-test

Exploratory Machine Learning Findings

A Random Forest Classifier trained on encoded features achieved an accuracy of 89% on a held-out test set (N = 274, Figure A9). Feature importance analysis indicated that variables such as orientation, face type, and emotion category were the strongest predictors of GPT-4's accuracy (Figure A11). A PCA visualization revealed separability between correct and incorrect predictions along two principal components, suggesting structured performance patterns rather than randomness. K-means clustering further supported this by identifying three distinct behavioral patterns in GPT-4’s emotion classification responses. These findings support the hypothesis that GPT-4's emotion recognition capabilities are modulated by cognitive-like processing conditions, paralleling human tendencies in interpreting facial expressions.

However, Hypothesis 5 was used to determine if GPT-4 mimicked the human effect observed in the Elaboration Likelihood Model (ELM), where people process thoughts through different routes. The central route refers to deep, analytical thinking, which occurs when a person is asked to elaborate on their reasoning and reflect on it. In contrast, the peripheral route refers to shallow, surface-level thinking, where less deliberate decision-making occurs (Petty & Cacioppo, 1986). The no-explanation prompt was used to model the peripheral route of thought, while the explanation prompt was used to model the central route of thought. Since the results of Hypothesis 5 suggest that prompting the explanation had no significant effect on accuracy, likely that GPT-4 does not successfully mirror human cognition in the manner of thinking more deeply about the subject it is analyzing by asking it to explain its decision-making.

Discussion

Summary

This study aimed to explore GPT-4’s capability to accurately recognize emotional expressions across various visual conditions (upright vs. inverted and individual vs. matrixed) and prompt contexts (with explanation vs. without explanation). The results showed significant variation in GPT-4’s performance depending on these submission factors. Hypotheses 3 through 5 were supported, while Hypotheses 1 and 2 were not.

A notable result of GPT-4's performance was that it performed significantly above chance (1/7 or 14.3%), but still fell below the human baseline accuracy reported by Olszanowski et al. (2015) on the dataset. The model demonstrated greater accuracy with upright and individually presented faces, which resembles the human-like impairments in configural processing under the inversion observed by McKelvie et al (1995). Uniquely, the presence vs. absence of explanation prompts did not lead to a significant overall impact on GPT-4's accuracy of labeling the emotion, which does not reflect the human-like chain of thought discussed in the ELM. This raises questions about the internal reasoning techniques the model utilizes to exhibit a response.

The additional use of machine learning tools, such as Random Forest Classification and Principal Component Analysis (PCA), revealed that the model exhibited structured behavior in its responses. This only upholds the notion that consistent visual and context cues heavily influence GPT-4's visual decisions.

Strengths and Limitations

One of the primary strengths of this study is its integration of psychology with advanced machine learning and artificial intelligence tools. Drawing from the well-studied Ekman's foundational work on emotions, which discusses how six basic human emotions are recognizable cross-culturally. The study also draws on McKelvie's inversion effect and modern ensemble perception theory, allowing for a multifaceted assessment of GPT-4's performance. To further strengthen the study's findings, the use of machine learning models to evaluate performance predictors added an ancillary layer of resilience, enabling analysis beyond standard statistical practices of comparison.

The study’s methodological design also ensured standardization and control over variables, particularly by using a validated image set (WSEFEP) and manipulating factors within a within-subjects framework.

However, despite the study's strength, several limitations must be considered. First, the analysis was restricted to static images, limiting its ability to be generalized to real-time facial expressions. Another noteworthy limitation is that although the dataset used was validated, the emotional expressions were posed rather than spontaneous, lacking the reflection of real-world complexity. A final consideration is that the explanation prompt may have been of an inadequate length to elicit a response from GPT-4, where the model used deeper reasoning to analyze the image. The model’s lack of significant performance change suggests that GPT-4 may exceedingly rely on visual token features, rather than eliciting a response through the use of language.

A final limitation to consider is that the study used a relatively small dataset for the matrixed faces. The study only contained 40 total matrix samples (upright and inverted); a larger sample would lead to higher validation of the results presented. The larger sample should contain varying facial expressions to ensure a wholesome evaluation of the model's accuracy.

Implications

These findings have considerable implications for the use of LLMs in affective computing. GPT-4 shows favorable potential as a tool for facial emotion analysis, with performance patterns resembling similar aspects of human visual processing, such as residing accuracy under inversion and matrix presentation—the model’s ability to perform low-level ensemble-style emotion classification is significant in the realm of using the capabilities for real-world applications like adaptive learning platforms, therapeutic tools, or emotion-aware AI agents.

With that, the failure of using explanation prompts to improve performance indicates GPT-4's reasoning may not reflect proper conceptual elaboration. This limitation is fundamental when considering the application of the model in areas where understanding the rationale is crucial.

Overall, GPT-4 may be better suited as an assistive tool in emotion recognition tasks when paired with dedicated biometric or feedback systems, which can provide higher accuracy in recognizing emotions and then utilize GPT-4 to offer a linguistic explanation for why the biometric system chose that emotion. Pairing these systems could counteract GPT-4's current limitations in configural processing and nuanced interpretation. OpenAI is continually improving its LLM models, which means a future model may be able to address the current weaknesses in these areas.

Future Research

While this study offers novel insights into GPT-4's capacity for facial emotion recognition, several avenues remain for further exploration of its capabilities. A significant step in the right direction would be to incorporate dynamic stimuli, such as providing the model with a video of a person changing their facial expressions throughout the video's duration. This would highlight the ability of GPT-4 to track emotional changes over time, which would resemble real-world scenarios. This would emphasize the model's utility in adaptive and interactive systems. Allowing us to assess the feasibility of using the model in areas where accuracy is essential, such as educational software, therapy bots, or customer service interfaces.

Further research should use a larger set of matrixed faces to further validate the results of the present study. The study used the lower limit of data size to be able to calculate an independent samples t-test. A larger set of data would lead to higher confidence in the evaluation of the model, as the confidence interval would be smaller and more accurate to the performance of the model.

Additionally, as OpenAI releases newer models, future studies could reproduce this study using a newer model and compare the results to determine if GPT-4's areas of weakness have been improved. Investigating how OpenAI improves its model could provide a deeper understanding of how visual transformers are utilized in operations surrounding image reasoning.

Lastly, further research could more heavily manipulate the prompt structure to more explicitly induce a process where the model uses deeper reasoning skills while analyzing the image. Future research could explore this by using varying levels of detail, emotional vocabulary, or task framing, allowing for a more rigorous determination of whether specific linguistic cues can lead to more accurate or human-like reasoning from the model. Improving the prompting for the model could enable optimal performance and refine how GPT-4 is utilized in practice.

Conclusion

The present study provides evidence that GPT-4 exhibits imperfect performance in recognizing facial emotions across different visual and contextual conditions. While the model surpassed the random chance level of accuracy and displayed human-like perceptual patterns, it could not match the human baseline for the dataset and showed limitations in inversion, requiring it to process multiple smaller images, which exemplified the model's struggles with increased complexity.

The machine learning analysis confirmed that GPT-4's emotion recognition is systematically influenced by visual features, indicating that it is able to recognize certain emotions with higher levels of accuracy, suggesting that its behavior is guided by patterns rather than randomness. These findings underscore the potential of GPT-4 in psychological and technological applications; however, further evaluation of the model is necessary before deploying it in emotionally sensitive contexts.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Bao, H., Bavarian, M., Belgium, J., Belgum, I., ... Zoph, B. (2024). *GPT-4 technical report*. arXiv. https://arxiv.org/abs/2303.08774

Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology, 36*(2), 253–263. https://doi.org/10.1017/S0048577299971763

Bayne, T., & McClelland, T. (2019). Ensemble representation and the contents of visual experience. *Philosophical Studies, 176*, 733–753. https://doi.org/10.1007/s11098-018-1225-0

DeAndres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., & Ortega-Garcia, J. (2024). How good is ChatGPT at face biometrics? A first look into recognition, soft biometrics, and explainability. *IEEE Access, 12*, 22548–22558. https://doi.org/10.1109/ACCESS.2024.3369833

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychologists Press.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–446. https://doi.org/10.1037/xge0000053

Iyer, A. A., Vojjala, S., & Andrew, J. (2024). Augmenting sentiments into Chat-GPT using facial emotion recognition. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 69–74). IEEE. https://doi.org/10.1109/ICACCS57996.2024.10314682

Lahat, A., Sharif, K., Zoabi, N., Shneor Patt, Y., Sharif, Y., Fisher, L., ... & Klang, E. (2024). Assessing generative pretrained transformers (GPT) in clinical decision-making: Comparative analysis of GPT-3.5 and GPT-4. *Journal of Medical Internet Research, 26*, e54571. https://doi.org/10.2196/54571

McKelvie, S. J. (1995). Emotional expression in upside-down faces: Evidence for configurational and componential processing. *British Journal of Social Psychology, 34*(3), 325–334. https://doi.org/10.1111/j.2044-8309.1995.tb01076.x

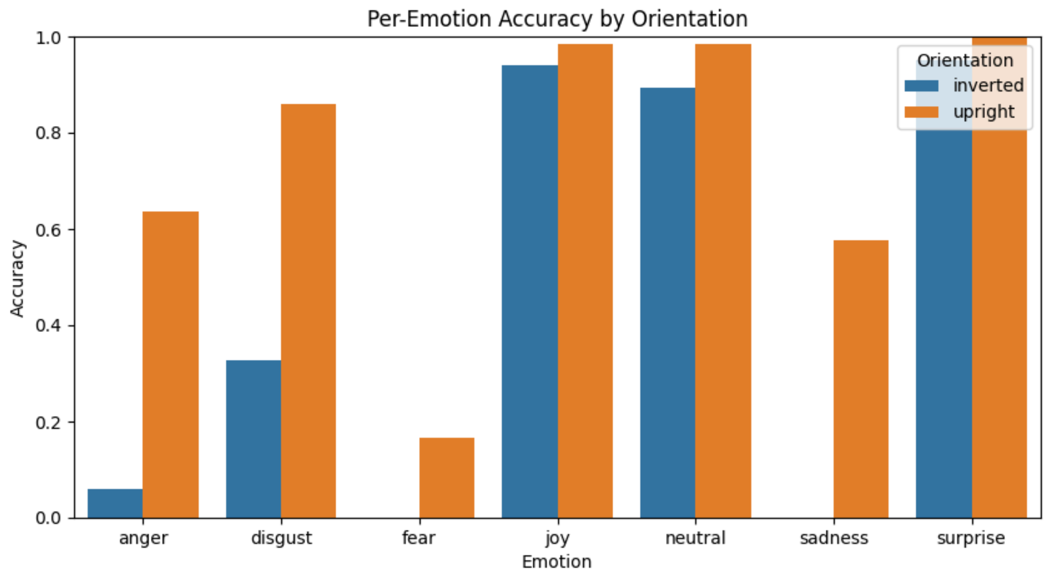
Olszanowski, M., Pochwatko, G., Kukliński, K., Ścibor-Rylski, M., Lewicki, P., & Ohme, R. (2015). Warsaw set of emotional facial expression pictures: A validation study of facial display photographs. *Frontiers in Psychology, 5*, 1516. https://doi.org/10.3389/fpsyg.2014.01516

Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* (pp. 1-24). Springer New York.

Appendix

Figure A1

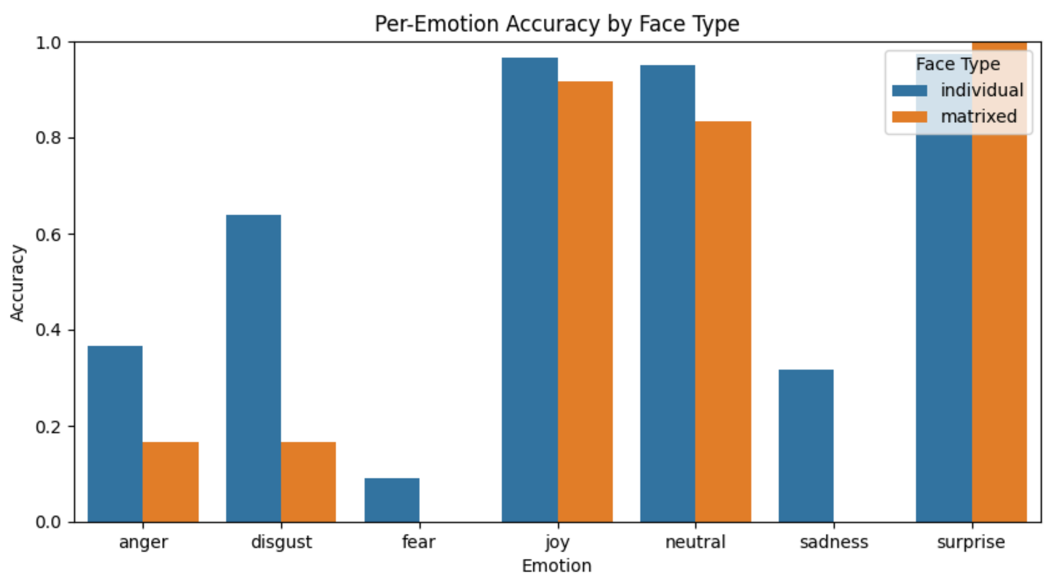
Bar chart depicting the difference in accuracy rates between orientations across expressions.



Note. Overall, the performance across all emotions was better when the images were upright; however, there was a significant difference in accuracy when the presented emotion was ‘anger’ or ‘disgust’.

Figure A2

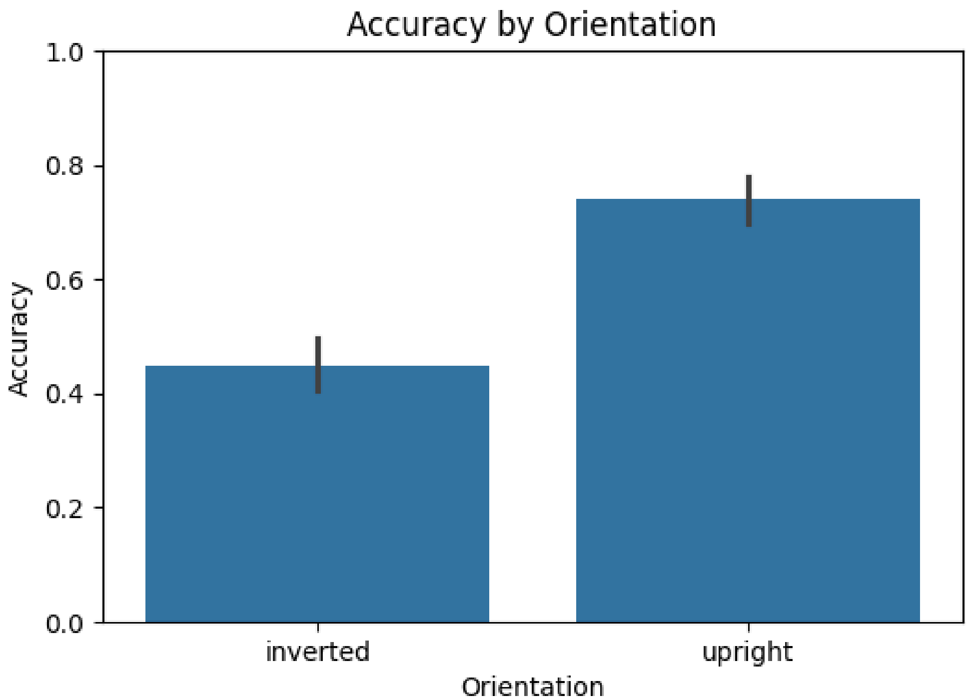
Bar chart depicting the difference in accuracy rates between face types across expressions.



Note. Overall, the performance across all emotions was better when the images were individual; however, there was a significant difference in accuracy when the presented emotion was ‘anger’ or ‘disgust’.

Figure A3

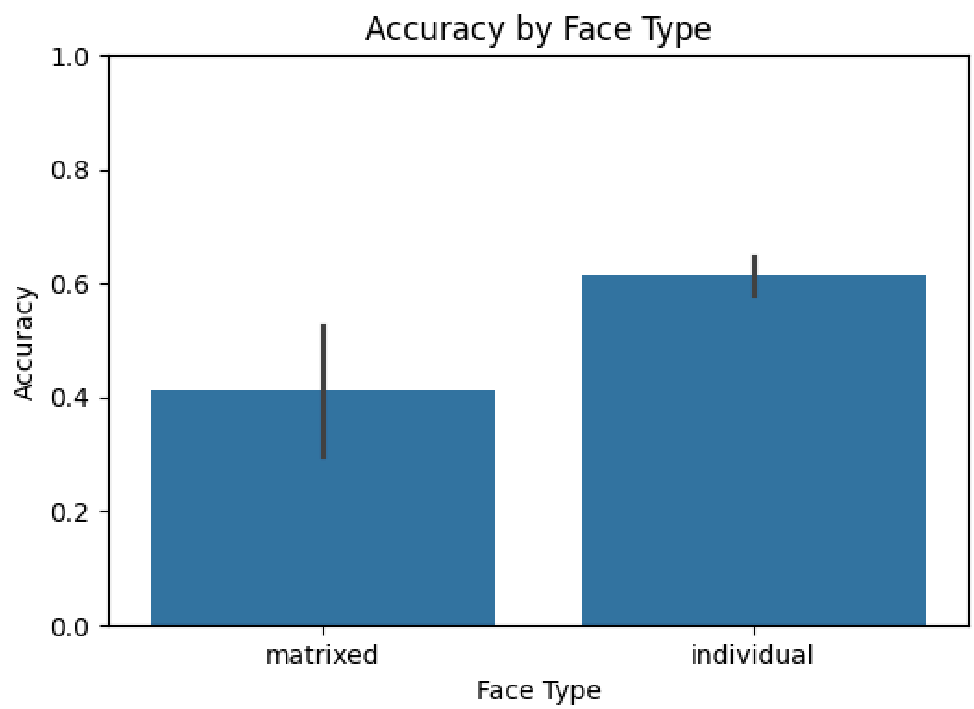
Bar chart depicting the overall accuracy rate between orientations.



Note. There is a significant drop-off in performance when the images are presented in inverted orientation.

Figure A4

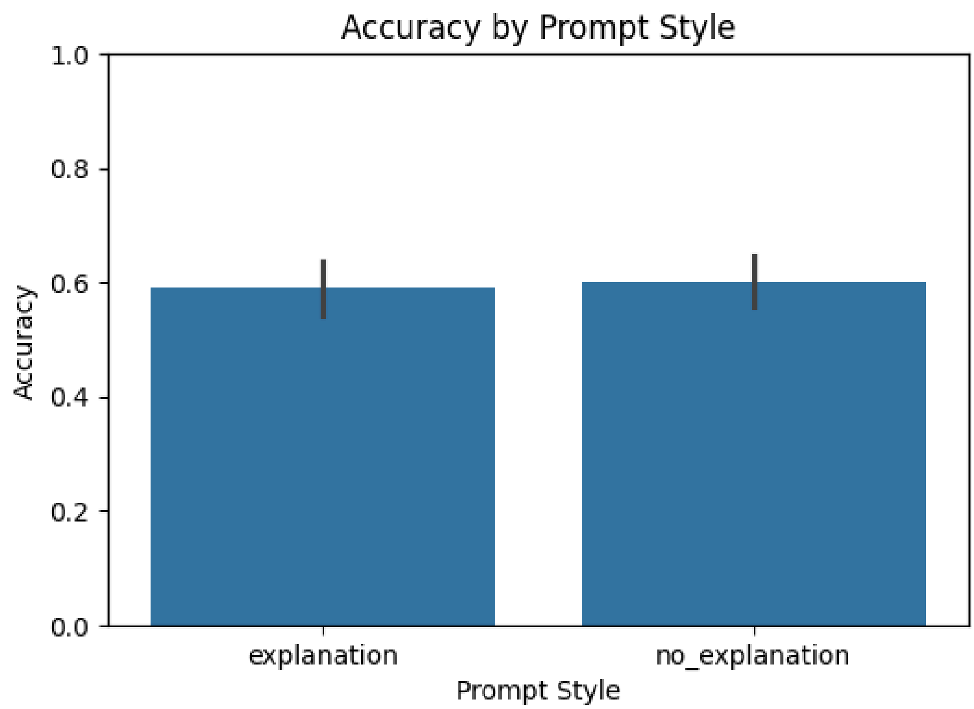
Bar chart depicting the overall accuracy rate between face types.



Note. The confidence band for the matrixed is much larger due to the small sample size of 40, while the individual sample size was 410.

Figure A5

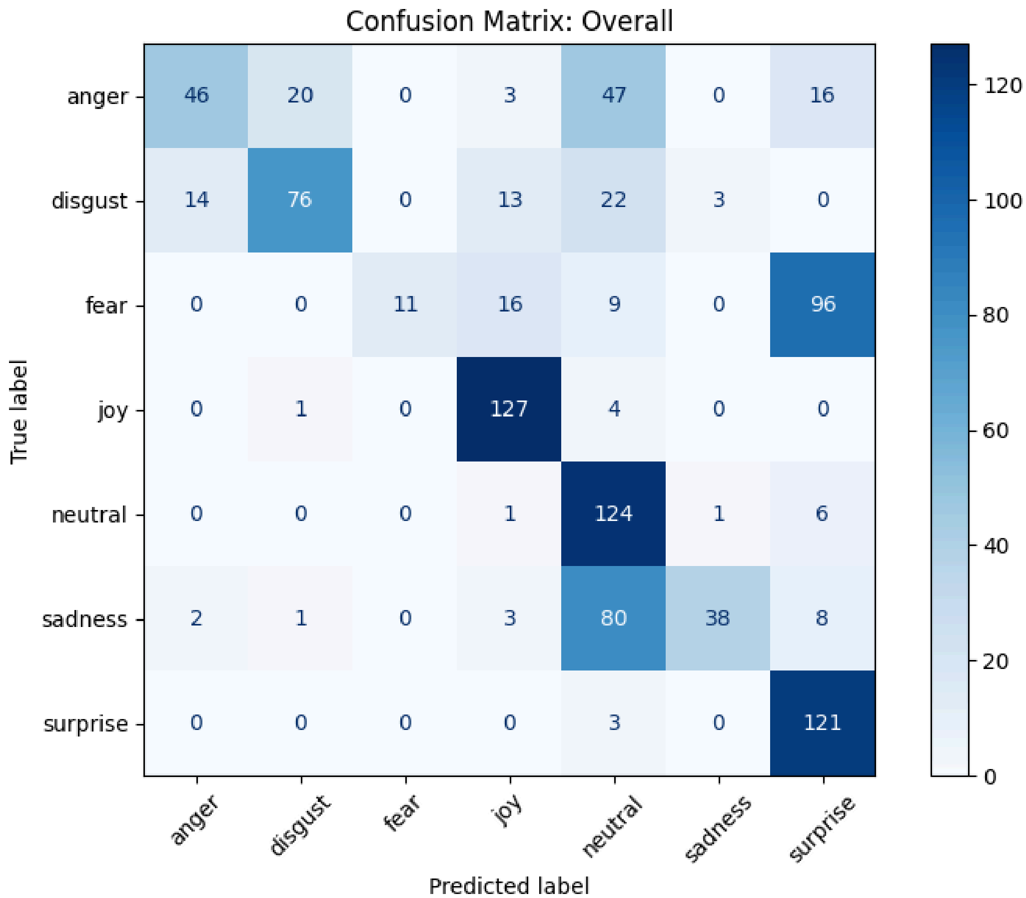
Bar chart depicting the accuracy rates between prompt styles.



Note. The confidence range of the accuracy was nearly the same for both prompts.

Figure A6

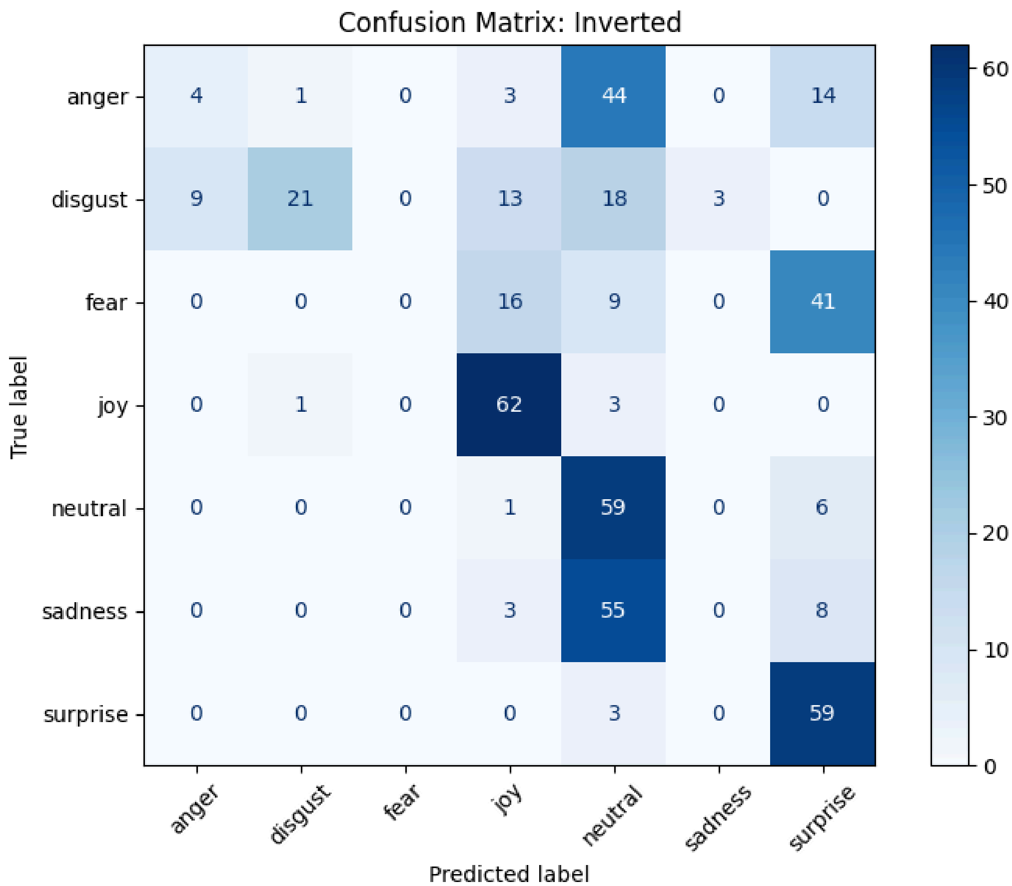
Confusion matrix of the overall responses, comparing them to the true facial expressions.



Note. Overall, the GPT-4 performed more accurately when the image's true label was 'joy', 'neutral', or 'surprise'.

Figure A7

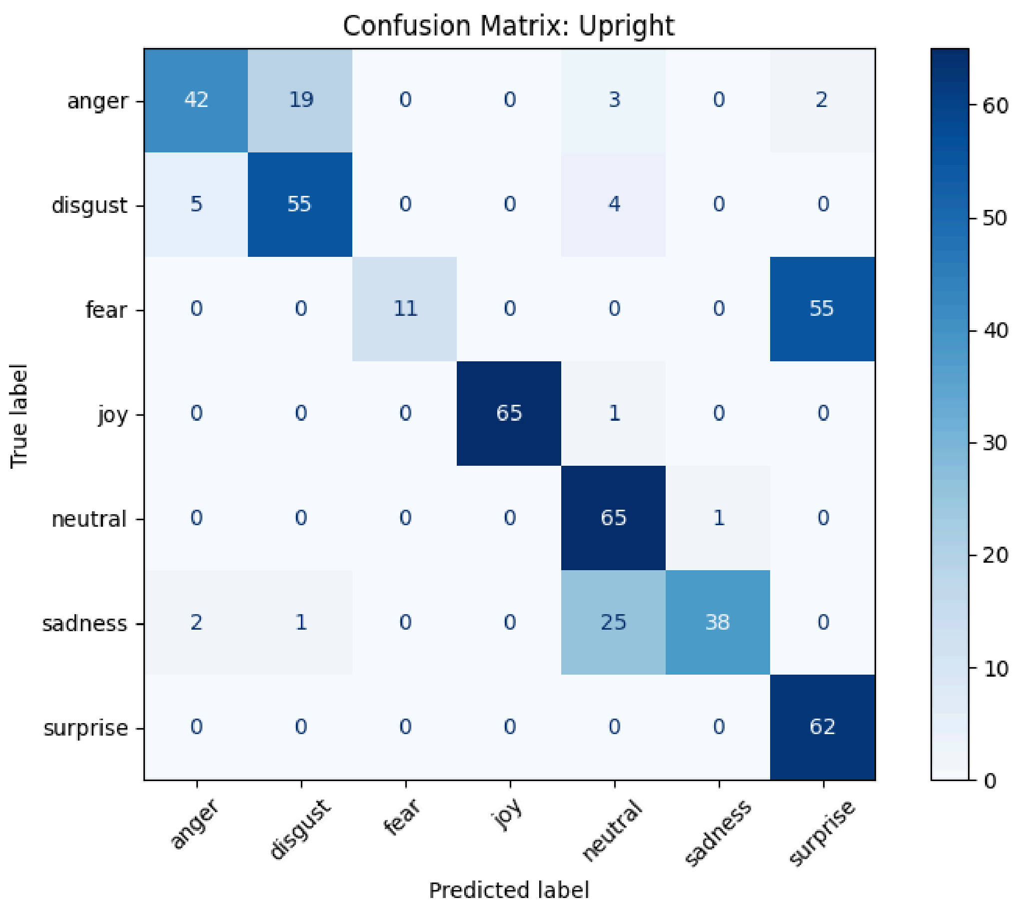
Confusion matrix of the inverted responses, comparing them to the true facial expressions.



Note. The inverted responses resulted in more sporadic predicted responses, leading to fewer matches with the true label.

Figure A8

Confusion matrix of the upright responses, comparing them to the true facial expressions.



Note. The upright responses resulted in more predicted responses matching the true label.

Figure A9

The resulting accuracy of the Random Forest Classification model.

Classification Report

```
Accuracy: 0.89

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       103
           1       0.92      0.89      0.91       171

    accuracy                           0.89       274
   macro avg       0.88      0.88      0.88       274
weighted avg       0.89      0.89      0.89       274
```

Note. The model produced an accuracy of 89%.

Figure A10

Confusion matrix of the Random Forest Classification model.

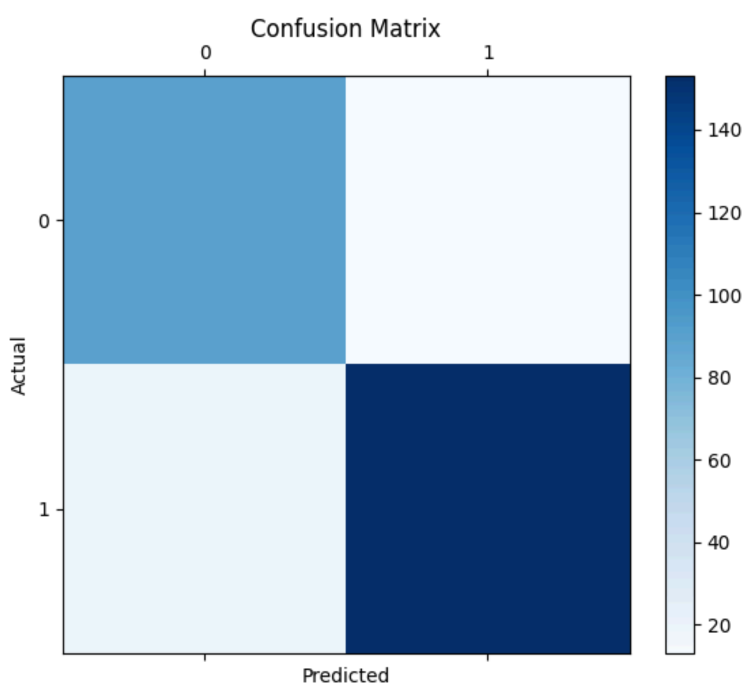
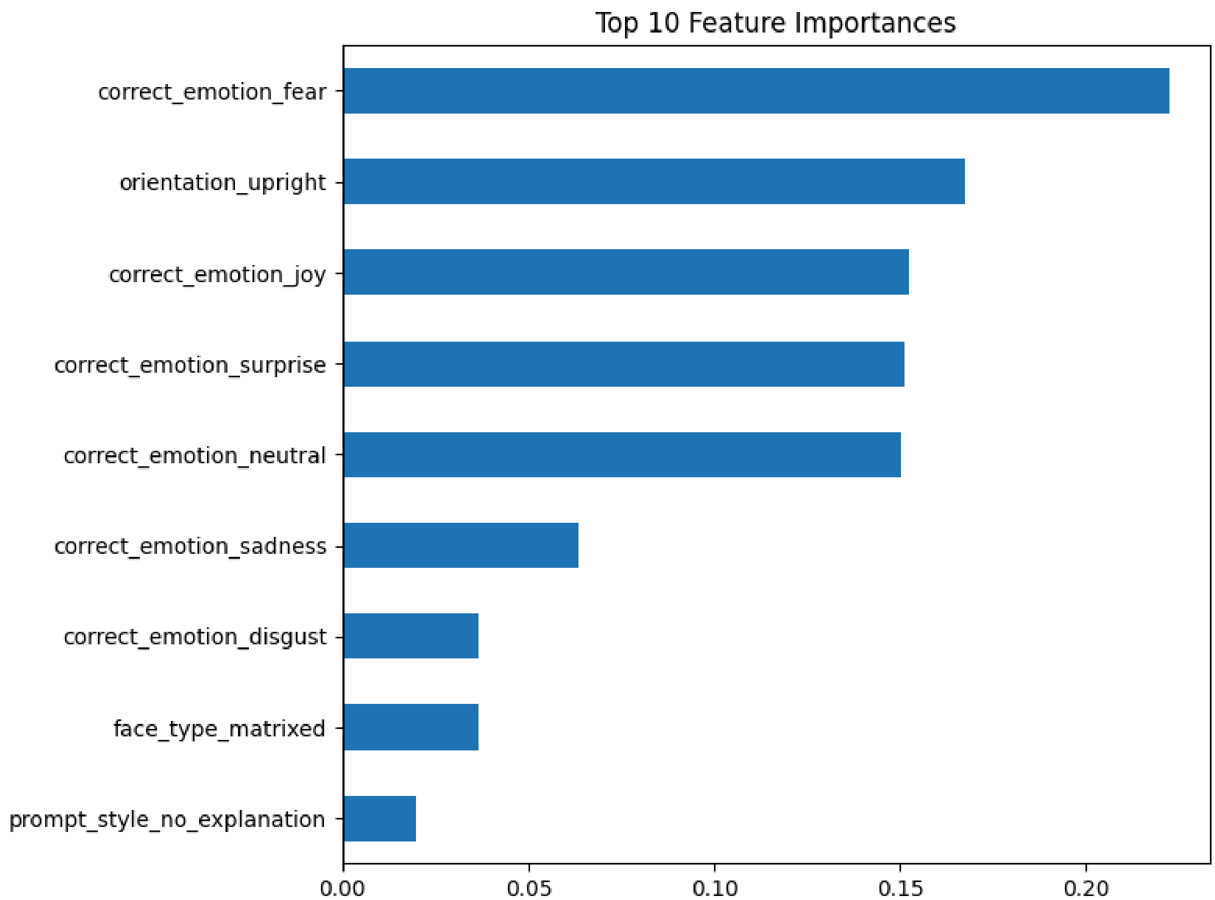


Figure A11

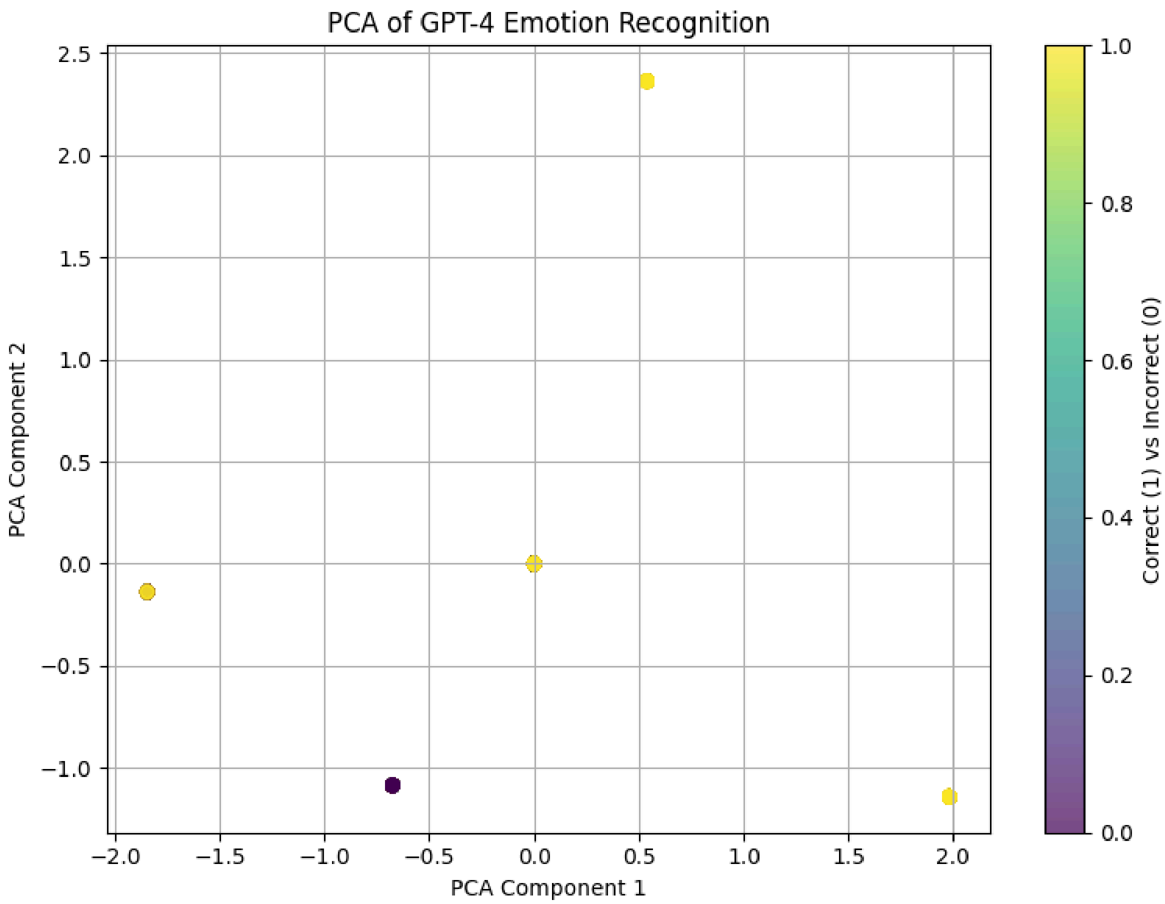
Top 10 Feature Importances in the Random Forest Classifier Predicting GPT-4's Emotion Recognition Accuracy.



Note. "Correct emotion: fear" and "orientation: upright" were the strongest predictors.

Figure A12

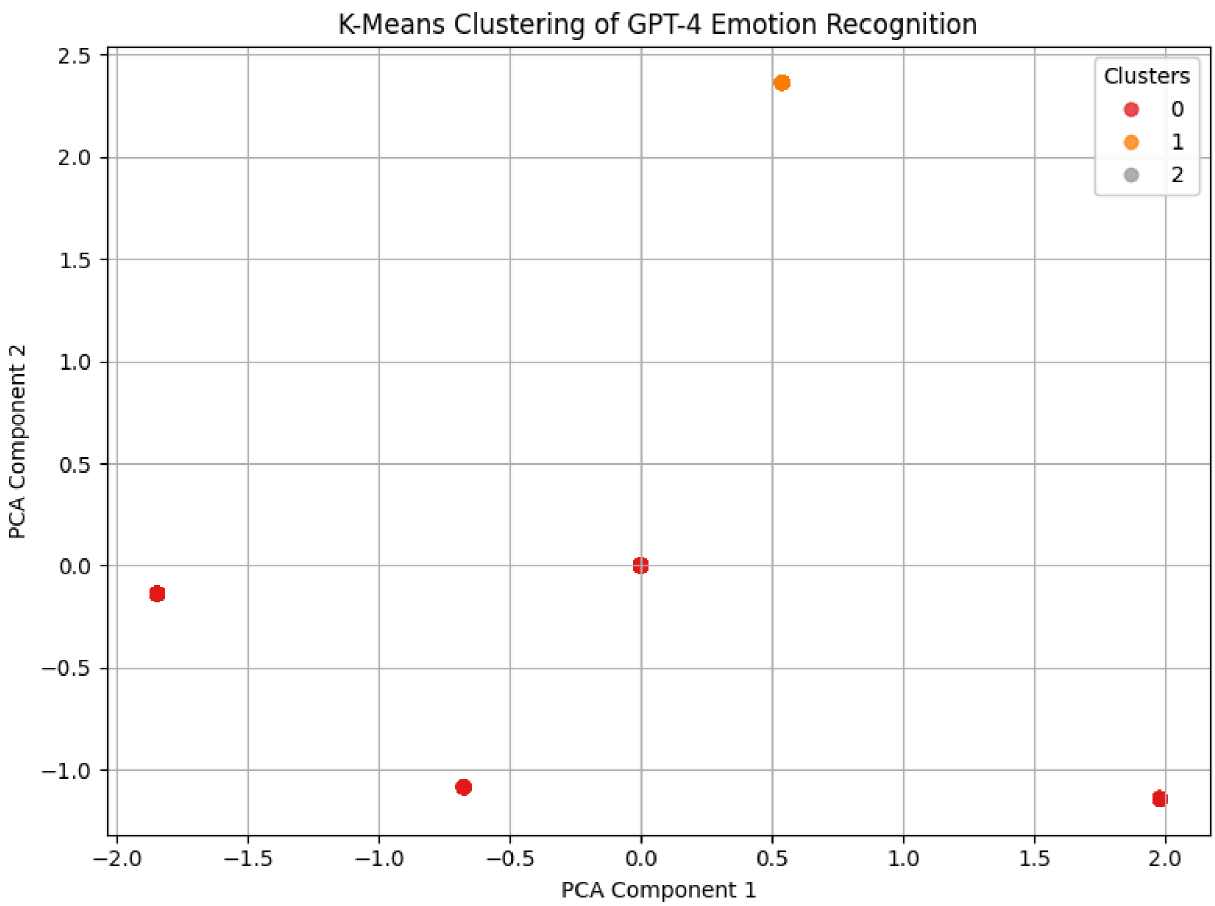
PCA scatterplot of GPT-4's emotion recognition outcomes, color-coded by accuracy (1 = correct, 0 = incorrect).



Note. The separation along principal components suggests the presence of structured patterns in recognition behavior.

Figure A13

K-means clustering of GPT-4's emotion classification responses based on PCA-transformed features.



Note. Three distinct clusters indicate potential processing modes under different input conditions.