GPT-4 accurately predicts human emotions at phenomenological and neural levels

Neuroimaging

Master of Science in Technology Thesis

Master's Degree Programme in Biomedical Engineering and Health Technology

Department of Computing, Faculty of Technology

Author:

Lauri Suominen

Supervisors:

Prof. Lauri Nummenmaa (Turku PET Centre)

PhD. Severi Santavirta (Turku PET Centre)

PhD. Jari Björne (University of Turku)

May 2026

Turku

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Neuroimaging, Master's Degree Programme in Biomedical Engineering and Health Technology
Author: Lauri Suominen
Title: GPT-4 accurately predicts human emotions at phenomenological and neural levels
Supervisors: Prof. Lauri Nummenmaa (Turku PET Centre), PhD. Severi Santavirta (Turku PET Centre), and PhD. Jari Björne (University of Turku)
Number of pages: 50 pages
Date: May 2026

Emotions are evoked by both internal and external events related to survival challenges. Recent advances in multimodal large language models ((M)LLMs), such as GPT-4, enable them to accurately analyze and describe complex visual scenes, raising the question whether LLMs can also predict human emotional experiences evoked by similar scenes. Here we asked GPT-4 and humans (N = 519) to provide self-reports of 48 unipolar emotions and affective dimensions for emotionally evocative videos and images. We evaluated GPT-4's emotion ratings using three natural socio-emotional stimulus datasets: two video datasets (234 and 120 videos) and one image dataset (300 images). We found that GPT-4 can predict emotions of human observers with high accuracy. The multivariate emotion structure (correlation matrices of emotions' ratings) converged between GPT-4 and humans and across datasets indicating that GPT-4 ratings for different emotions follow similar structural representations as the human evaluations. Finally, we modeled the brain's hemodynamic responses for emotions elicited by videos or images in two fMRI datasets (N = 97) with GPT-4 or human-based emotional evaluations to highlight the usefulness of GPT-4 in neuroscientific research. The results showed that the brain's emotion circuits can be mapped with high accuracy using GPT-4 emotion ratings as the stimulation model. In conclusion, GPT-4 can predict human emotion ratings to the extent that GPT-4 ratings can also model the associated neural responses. Our results indicate that LLMs provide novel and scalable tools that have broad potential in emotion research, cognitive and affective neuroscience, and that it can also have practical applications.

Keywords: Emotions, large language models, LLM, GPT-4, functional magnetic resonance imaging, fMRI.

Table of Contents

List of Figures

List of Tables

1 Introduction

This master's thesis has led to the publication of a scientific article, which was completed simultaneously with this thesis and therefore contains significant overlaps in content. The scientific article has been published as a preprint in the bioRxiv repository (Santavirta, Suominen, et al. 2025) and is conditionally accepted for publication *IEEE Transactions on Affective Computing* journal after the peer review process is complete.

The research for the master's thesis and the scientific article has been carried out in parallel and is based on the same research material and analysis methods. The aim of the scientific article is to present the main research results and observations in a concise and informative form. Instead, this master's thesis covers the topic more broadly, including a more detailed background on the research methods used, neuroimaging, large language models (LLM), key theories, and deepens the conclusions drawn from the research.

1.1 Using AI tools

AI tools (ChatGPT, Copilot, Grammarly or DeepL) have only been used to assist with text editing and grammar checking. AI has also been used occasionally to assist with some areas of code scripts in the analyses. However, AI has not been responsible for designing or performing the analysis method in any way. AI tools have also not been used to generate any text in this thesis or article related.

1.2 Research Questions

The purpose of this master's thesis is to investigate how multimodal large language models (MLLMs) such as OpenAI's GPT-4 model can estimate human-specific emotions from images and videos that have been shown to evoke emotions. The research section of the thesis aims to answer the following research questions:

- RQ1: Can multimodal large language models (MLLMs) estimate human-specific emotions from images and videos, producing human-like evaluations?
- RQ2: How similar are the structural relationships of the emotion evaluations produced by GPT-4 to human evaluations?
- RQ3: Can the emotion evaluations produced by GPT-4 be reliably utilized in modeling the brain's emotional networks in fMRI data?

In addition, the master's thesis aims to answer the research methods, concepts, and techniques used in the research section, the mastery of which is of paramount importance for understanding this research.

1.3 Thesis Content Summary

Emotions support homeostasis by preparing organisms for action based on the relationship of their current environment, bodily states, and goals (Nummenmaa 2022). Subjective emotional feelings reflect the emotion-dependent central and peripheral changes, providing an interface between the conscious executive functions and the autonomic control (Damasio and Carvalho 2013; Sander 2025; Scherer and Moors 2019; Frijda 2009). Researchers have tried mapping the subjective emotion space with self-reports through different theoretical emotion models such as two-dimensional valence and arousal framework (Russell et al. 1989), appraisal dimensions (Yeo and Ong 2024), based on discrete emotions with different survival functions (Ekman 1992), and more recently using data-driven approaches for mapping high-dimensional representations for emotional experiences (Cowen and Keltner 2017; Keltner et al. 2019).

Collecting self-reports for large datasets necessitated by data-driven approaches limits the scalability of emotion research both from practical and financial perspectives. For example, in our previous data-driven study into human socioemotional perception it was necessary to recruit 2000+ participants to annotate 100+ social perceptual features from naturalistic stimuli (Santavirta et al. 2024). This required over $10 000 in participant compensations and over 1 100 hours of their time. An automated method for predicting human emotional experiences would tackle this bottleneck (Ziems et al. 2024) because it would allow mapping emotional representations based on larger and more diverse sets of stimuli than before. This would provide means for comparative studies across the widely used stimulus sets whose normative data however vary widely from low-dimensional ratings (Bradley and Lang 2007) to basic emotion evaluations (Riegel et al. 2016) and high-dimensional emotional assessments (Cowen and Keltner 2017).

Due to the cost of data collection and standardization, researchers have mainly focused on such standard datasets, although those are limited by their contents or annotated features. As pioneered by the field of affective computing, reliable method for automated annotations would liberate researchers from using the standardized datasets only and lower significantly the need for human subjects. This would benefit particularly neuroimaging studies, as previously collected costly imaging datasets could be re-analyzed in depth through comprehensive

automated remapping of the stimulus space, allowing delineating the brain's emotion circuits with unprecedented precision. Finally, automated emotion recognition would allow the research and development of important, concrete, real-life applications in areas such as mental health monitoring, healthcare, marketing, customer service improvement, facilitating human-robot interaction, and security. These applications could be used in homes, hospitals or workplaces (Khare et al. 2024; Guo et al. 2024). For example, patients' wellbeing could be closely monitored automatically or customers' reactions to advertisements or visual products could be estimated in advance.

Artificial intelligence (AI) and large language models (LLMs) might solve this scalability problem. Recently, LLMs have shown improving capabilities across disciplines. It is already evident that LLMs possess advanced knowledge indicated by high performance in standardized knowledge and intelligence tests (Katz et al. 2024; King 2023; Kung et al. 2023). However, the LLMs' abilities extend beyond factual knowledge. A growing body of evidence suggests that LLMs can offer more abstract insights into human behavior and psychology (Demszky et al. 2023; Ke et al. 2025) and LLMs are increasingly studied as proxies of human participants (Dillion et al. 2023; Horton 2023; Sohail and Zhang 2025). For example, LLMs can predict the opinions of people belonging to different sociodemographic groups (Argyle et al. 2023), make economic choices similar to humans (Horton 2023), predict different personalities (Y. Wang et al. 2025) and perform human-like mental state inferences (Strachan, Albergo, et al. 2024). These investigations are however solely based on textual input for LLMs, and in real life, a large bulk of human socioemotional processing is sensory rather than linguistic.

These findings have sparked research of LLMs and emotions. LLMs are increasingly capable of recognizing others' emotions from textual standard emotion recognition tasks (Sabour et al. 2024; J. Huang et al. 2024; Tak and Gratch 2024; X. Wang et al. 2023; Aher et al. 2023; Tak et al. 2025). However, real-life emotions are predominantly elicited by complex audiovisual information from the surrounding environment. Thus, studies that attempt to evoke emotional experiences typically use image or video stimulation in the laboratory (Marchewka et al. 2014; Bradley and Lang 2007; Lettieri et al. 2019; Hudson et al. 2020; Karjalainen et al. 2017; H. Saarimäki et al. 2023). Currently, many LLMs have multimodal reasoning capabilities. For example, we have previously demonstrated that GPT-4 (ChatGPT) can analyze complex social information from image and video stimuli (Santavirta, Wu, et al. 2025), but it is currently not established whether LLMs can predict how humans rate their own emotions for naturalistic stimuli.

The development of LLMs is rapid, and new models are released weekly. At the time of writing, popular large-language models include OpenAI's GPT (ChatGPT), Google's Gemini, Anthropic's Claude, DeepSeek's R1, X's Grok, and Meta's Llama. GPT-4 is the currently most studied model, as it has already been available for some time, and has been the most popular for users. Previous research has shown that GPT-4 outperforms many previous models in its diverse capabilities in psychological research (Ke et al. 2025). Recent studies have also showed that GPT-4 can recognize emotions from standard close-up images of human eyes (Elyoseph et al. 2024; Strachan, Pansardi, et al. 2024). Thus, rather than exhaustively testing every available model, we use GPT-4 as a representative case to demonstrate the core approach.

1.4 The Current Study

The purpose of the current study was to investigate whether GPT-4 can predict subjective emotions evoked by videos and images in humans, and whether these predictions extend to concordant hemodynamic brain responses in the brain for human versus GPT generated evaluations. We prompted GPT-4 to evaluate a wide range of video clips and images known to evoke emotional experiences in humans. GPT-4 was tasked to predict the human ratings for 48 unipolar emotions and affective dimensions when viewing the stimuli. Similar ratings were obtained from human participants for all the stimuli with exactly the same wording (i.e. prompt for GPT; question for humans), to allow estimating the convergence between GPT-4 and human emotion ratings. GPT-4 and humans were explicitly asked "To what extent this makes you feel [emotion]?". Further, we modeled neural responses obtained from functional magnetic resonance imaging (fMRI) data with human and GPT-4 emotion ratings to investigate the similarity of the neural representations. This was done to evaluate the feasibility of using GPT-4 derived emotion ratings for large-scale annotation of stimuli in brain imaging studies, where high-dimensional stimulus spaces occur commonly (such as in videos and images) but they are difficult to characterize accurately due to the high demands imposed on human observers. The results show that GPT-4's emotion ratings were highly similar to those given by humans, and that, accordingly, the neural emotion circuits identified based on GPT-4 or human emotion ratings converged well. These results indicate that GPT-4 can predict human ratings for emotional experiences and consequently, GPT-4 ratings can be used to reliably model the brain responses of audiovisual stimuli.

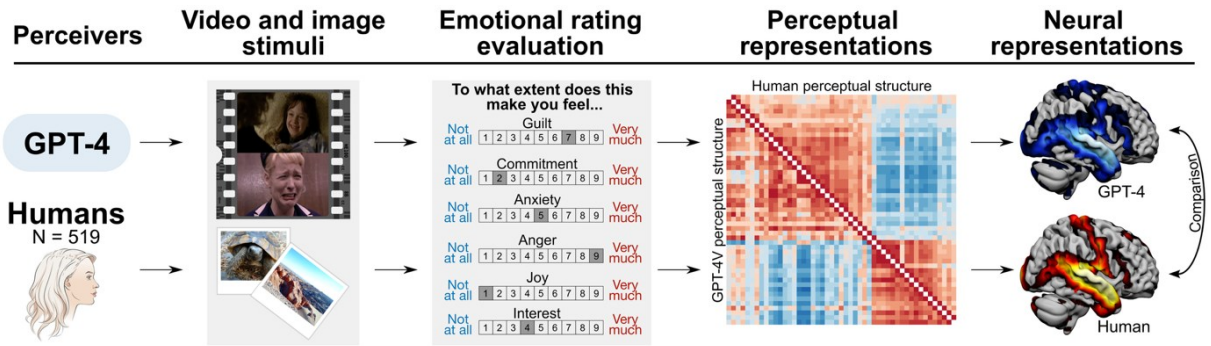

Figure 1. The analytical workflow of the study. First, GPT-4 and human observers provided ratings for 48 emotions for a set of videos and 8 emotions for images. We then compared the similarity of the emotion ratings with real human responses to the same stimuli. The emotion ratings were subsequently used to create stimulation models for mapping neural representations of emotions in large fMRI datasets (96 subjects, video and image stimuli) to compare the resulting neural representations between human and GPT-4-derived models.

2 Background

This master's thesis examines the capacity of multimodal language models (MLLMs) to measure human emotions from both phenomenological and neural viewpoints. Incorporating psychological and neuroscientific perspectives with advanced artificial intelligence models, the research investigates whether MLLM can serve as human-like raters for emotion in multisensory video and image data. This study uses functional magnetic resonance imaging (fMRI) to compare neural activation between human and artificial intelligence (GPT-4V) emotion assessments. Therefore, it is essential to understand the key theoretical and technological foundations for the research which are presented in this chapter.

The four central themes of the chapter are: 1) theories of emotion in psychology and neuroscience, 2) fMRI imaging and the neural correlations of emotions, 3) eliciting emotions through external stimuli, and 4) the structure and operating principles of large language models (LLMs) and multimodal large language models (MLLMs).

2.1 Emotion Theory in Psychology and Neuroscience

Emotions are intricate psychological and biophysical processes associated with adaptation, decision-making, social interaction, and personal well-being. Although emotions are a popular research topic in both psychology and neuroscience, there is still no consensus on how to precisely define an emotion. Emotions are, in general, considered a multistage process that begins with a stimulus-triggered cognitive appraisal and physiological response, which lead to an emotional experience and behavioral reaction, resulting ultimately the full-blown emotional state. (Celeghin et al. 2017)

Paul Ekman's theory of basic emotions is one of the most well-known theories of emotion. According to this theory, certain emotions are universal and have clear, recognizable expressions, neural correlates, and evolutionary significance. These basic emotions typically include joy, sadness, anger, fear, disgust, and surprise (Ekman 1992). Robert Plutchik also presented a similar theory in which eight basic emotions are arranged in a circular pattern as opposites of each other and can be combined to form more complex emotional experiences (Plutchik 2001). These theories propose that emotions are biological reactions that evolved to promote survival. The assumption is based on the idea that emotions are automatic, fast reactions to important stimuli in the environment.

An alternative theory to basic emotion theories is the dimensional models, which treat emotions as continuums with two or more dimensions. The most commonly used model is the "Circumplex model of affect" developed by James A. Russell, which presents emotions in terms of two main dimensions: valence (unpleasant to pleasant) and arousal (deactivation to activation) (Russell 1980). In this model, valence describes the pleasantness of an emotion. Happy and content describe high valence, while sad and upset describe low valence. Arousal, in turn, describes the level of physiological activation. Tense and alert describe high arousal, while bored and calm describe low arousal. The model is key to the systematic study of emotional processes in disciplines such as neuroimaging and machine learning because it allows emotions to be measured quantitatively. (Posner et al. 2005)

The literature often distinguishes between the concepts of emotion and feeling. An emotion is a complex, multi-component psychophysiological response that involves changes in states such as our body, perception, thought, acting, or behavior. A feeling, on the other hand, is the subjective experience of an emotion (Barrett et al. 2007). These subjective experiences are also at the core of the present study.

Emotions can be divided into positive and negative, based on their valence. This evolutionarily property guides approach and avoidance behavior. Positive emotions, such as joy and love, encourage approaching, while negative emotions, such as fear and disgust, encourage moving away from a dangerous situation (Cacioppo et al. 1999). Emotions can be further divided to primary and secondary emotions. Primary emotions are fast, automatic, and shared across species, while secondary emotions, on the other hand, develop from an individual's experiences and social environment (TenHouten 2017).

As these theories show, interpreting emotions is complex, both in theory and practice. Another challenge in studying emotions is that people see and express them differently. These differences come from personal experiences and cultural backgrounds. Consequently, similar stimuli can trigger different responses based on an individual's background and experiences. This situation creates a challenge for artificial intelligence models and psychological tests designed to identify or predict emotional states based on external signals. Therefore, it is crucial to understand and address these challenges when assessing the ability of artificial intelligence models to evaluate human emotions and emotional processes.

Theory-driven vs. data-driven approaches in emotion research

Traditional emotion research has often been very theory driven. Such research typically begins with specific hypotheses drawn from psychological or neuroscientific theory models. These studies aim to test how well experimental data fit established theories, like basic emotion theory or dimensional methods (Angkasirisan 2024; Adolphs et al. 2016). However, this approach can be limited by the theory's scope and fixed emotion categories. In recent years, technological advances have made large datasets and strong computing power more accessible. As a result, data-driven and statistically based methods have become more common in emotion research. Unlike theory-driven models, data-driven approaches do not start with a theory-based hypothesis. Instead, they look for patterns in large data sets to show hidden emotional structures and their relationships (Adolphs et al. 2016). These methods are especially common in machine learning, where models are trained to find emotional patterns in behavioral, neural, or multimodal signals.

A notable example of a data-driven approach is "Semantic Space Theory", created by Cowen and Keltner in 2021. Using large-scale statistical modeling and machine learning techniques, these researchers identified over 25 distinct emotion categories based on participants' self-reported reactions. These categories came from the data itself, rather than being predefined, and provided a nuanced view of human emotions. Cowen and Keltner's study showed that emotional experiences are much more complex and sensitive to context than traditional theories suggest. (Cowen and Keltner 2021)

This master's investigates the relationship between GPT-4V's emotion ratings, human subjective assessments, and neural responses using a data-driven method. The study uses the 34 emotion categories introduced by Cowen and Keltner in 2017, based on the work of well-known theorists. It also incorporates the 14 affective dimensions commonly used to measure subjective experience, including valence, arousal, control, and safety. (Cowen and Keltner 2017) These 48 emotional features comprise a comprehensive, fine-grained emotional feature space that enables large-scale comparisons between GPT-4V's ratings and human ratings and neural responses.

2.2 Neural Correlates of Emotions

Non-invasive imaging of the living brain

Several non-invasive imaging methods have been developed to study and measure brain function in living brain tissue. Functional imaging studies brain activity during different events, such as cognitive, sensory, or emotional processes. The most commonly used functional imaging methods are electroencephalography (EEG), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). Each method has its strengths and limitations, especially regarding temporal and spatial resolution.

Electroencephalography (EEG) involves placing electrodes on the scalp to record electrical activity from brain neurons. While EEG offers high temporal resolution, it has low spatial accuracy due to sparsely spaced sensors on the scalp. The clear advantages of EEG are that it is radiation-free, portable, and inexpensive in terms of equipment and measurements. However, EEG is sensitive to artifacts such as muscle contractions, eye movements, breathing, and external electrical devices. In addition, EEG cannot measure activity in deep brain structures because it mainly registers electrical potentials in the cerebral cortex. (Kirschstein and Köhling 2009)

Magnetoencephalography (MEG) is closely related to electroencephalography (EEG). Unlike EEG, MEG measures the magnetic fields generated by electrical activity in the brain. Like EEG, MEG has excellent temporal resolution, and in certain cases, such as when the area of interest is well-defined, it can provide better spatial resolution than EEG. The advantages of MEG include no radiation exposure and no need for electrodes or conductive gel during measurements. However, MEG also has several disadvantages. It involves more complicated and expensive instrumentation than EEG. Also, using MEG requires a magnetically shielded room. (Ahlfors and Mody 2019; Gross 2019)

Functional near-infrared spectroscopy (fNIRS) is an optical imaging method that employs near-infrared light to measure blood flow activity in the cerebral cortex. The most common fNIRS technique assesses the ratio of oxyhemoglobin (Hb) to blood volume. Near-infrared light (700–900 nm) is strongly absorbed by oxyhemoglobin (Hb) and deoxyhemoglobin (dHb), but not by water. This property allows the light to penetrate tissues effectively. Light detectors can sense scattered photons to estimate changes in oxyhemoglobin (Hb) and deoxyhemoglobin (dHb)

concentrations near the cortical surface. In terms of temporal resolution, fNIRS outperforms fMRI but falls short compared to EEG and MEG. Regarding spatial resolution, fNIRS is more precise than EEG and MEG, but less precise than fMRI and PET. Advantages of this method include being radiation-free and portable. However, its limitations include shallow measurement depth and a lack of standardization. (Bunce et al. 2006; Wilcox and Biondi 2015)

Positron emission tomography (PET) is an imaging technique that uses radiotracers to measure tissue metabolism or blood flow accurately. One commonly used radiotracer in cognitive and emotional studies is oxygen-15, which can be used to image cerebral blood flow (Ter-Pogossian and Herscovitch 1985). Radioactive tracers decay and release gamma rays, also known as annihilation photons, which travel in opposite directions. A PET scanner detects photons that reach opposite detectors simultaneously and creates an image map of the examined area. This method offers very high spatial resolution. Additionally, PET can be paired with other imaging techniques like CT or MRI. However, PET imaging has several disadvantages, including radiation exposure, high costs, and the short half-life of the tracers, which requires careful timing and a nearby radiochemistry lab. (Phelps 2000; Shukla and Kumar 2006)

Although all the functional imaging methods described above have advanced the study of emotions and cognition, functional magnetic resonance imaging (fMRI) has become one of the most widely used techniques. This is due to its combination of high spatial resolution, moderately good temporal resolution, and no radiation exposure. Its disadvantages include high costs and a noisy measuring environment. This master's thesis will also use fMRI because its features fit the research setting well. The following subsections provide more detailed information about fMRI operations, data processing, and analysis.

Hemodynamic responses and fMRI

Functional magnetic resonance imaging (fMRI) is one of the most important non-invasive methods in neuroscience and the study of human brain function. It enables a wide range of studies of motor, sensory, and cognitive processes in the human brain. The most common application of fMRI is based on BOLD (Blood Oxygenation Level Dependent) imaging, which measures changes in blood oxygenation. The BOLD signal indirectly reflects neuronal activity and is based on the relative amounts of oxyhemoglobin (Hb) and deoxyhemoglobin (dHb) in the blood. (Huettel 2004)

Oxyhemoglobin (Hb) is a diamagnetic molecule, whereas deoxyhemoglobin (dHb) is paramagnetic. The magnetic susceptibility, or the ability to be magnetized in an external magnetic field, changes with oxygen concentration, so that as an active brain area receives more oxygen, the amount of oxyhemoglobin (Hb) increases and the amount of deoxyhemoglobin (dHb) decreases. These changes lead to local magnetic field fluctuations that can be measured by fMRI equipment as a BOLD signal. (Huettel 2004)

A linear relationship is assumed between brain activity and the BOLD signal. This means that as activity increases, the signal increases proportionally (scaling), and the response to a single stimulus is independent of the other present stimuli (superposition) (Huettel 2004). These linearity assumptions allow the signals to be modeled mathematically. For instance, a generalized linear model (GLM) can analyze GPT-4V's emotional judgments about the BOLD signal, as discussed in subsequent sections.

fMRI implementation and preprocessing

In a typical fMRI experiment, images of the brain are taken every few seconds. Hundreds of volumes can be collected during the scan to form a four-dimensional (x, y, z, and time) dataset. Statistical analysis of this raw data requires preprocessing to reduce noise and standardize the signal for subsequent analysis. (Smith 2004) The most common preprocessing steps are slice-timing correction, motion correction, spatial filtering, intensity normalization, and time series filtering. Slice-timing correction aligns the times of the slice recordings with the times of the simultaneous measurements (Parker and Razlighi 2019). Motion correction aligns the volumes with each other to compensate for distortions caused by head movements. Spatial filtering blurs the volumes by averaging signal in adjacent voxels to improve the signal-to-noise ratio and increase sensitivity to signal differences. Intensity normalization aims to correct global variations between volumes (Sun et al. 2015). Finally, time series filtering uses linear and nonlinear filters to remove low- and high-frequency artifacts, such as respiratory sounds and heart rate. (Cohen et al. 2017)

Statistical analysis

The preprocessed data is analyzed statistically at the voxel level, meaning each small, three-dimensional part of the brain image is analyzed separately. The most used method is the generalized linear model (GLM) analysis, which was previously mentioned. GLM allows for

the modeling of responses caused by different stimuli. The purpose of GLM is to build a predictive model of how stimuli affect the BOLD signal over time.

Model fitting involves convolving the stimulus function with the hemodynamic response function (HRF) over time. This mathematical process mimics the delay and dispersion of the BOLD signal caused by neurophysiological processes. The result is a statistical map showing the brain regions whose activation is statistically significantly related to a particular stimulus or regressor. (Smith 2004) Since the analysis involves thousands of simultaneous statistical tests (one per voxel), corrections for multiple testing are necessary to prevent an uncontrollable increase in false positives.

The most common method of correcting statistical tests is family-wise error rate (FWE) correction. This method controls the likelihood that even one null hypothesis is incorrectly rejected in the entire voxel space. FWE corrections can be implemented using random field theory or permutation analysis, for example. Another option is the false discovery rate (FDR) correction. This method allows for a controlled proportion of false discoveries, making it somewhat more sensitive. (Nichols and Hayasaka 2003)

Limitations of fMRI

Although fMRI is widely used in different studies, it has several technical, physiological, and practical limitations. First, fMRI does not directly measure neuronal electrical activity. Instead, it measures brain activity indirectly using the BOLD signal, as discussed in the chapter "Hemodynamic Responses and fMRI". This creates a delay in the hemodynamic response, which typically lasts 4 to 10 seconds from neural activation. This limitation affects fMRI's temporal resolution compared to methods like EEG and MEG (Uludağ 2023). The BOLD signal is also sensitive to various physiological and environmental factors. These include changes in respiratory and heart rates, as well as head movements. Even minor movements can create significant artifacts in the signal, which can distort the analysis results. Additionally, variations in vascular structure, hemoglobin concentration, and vascular activity can influence signal strength and interpretation (Liu 2016; Constable 2023).

The third limitation is spatial resolution, which is limited by the size and density of neurons. Since a single imaging point, or voxel, contains millions of neurons, the resulting signal reflects the average activity of large cell populations. Consequently, fine-grained, local activation

patterns may be obscured. This is why fMRI has lower spatial resolution than PET imaging. (Goense et al. 2016)

From a technical perspective, the limitations of fMRI arise from the physical properties of MRI sequences. For instance, T2 and T2* relaxation result in signal attenuation during the echo. At very high resolutions, this can blur the image, causing different tissue types to appear at different resolutions depending on their characteristic T2 value. Additionally, the diffusion rate of water molecules limits the accuracy of fMRI because their movement during data acquisition causes blurring and weakens signal localization. However, this is a secondary limitation because, in practice, fMRI measurements do not reach this theoretical limit. (Constable 2023)

In terms of practical limitations, fMRI measurements are expensive and require large, fixed equipment, magnetically shielded facilities, and specialized supplies. The fMRI tube is cramped and noisy, which can complicate measurements if the subject feels uncomfortable during the procedure. For instance, 2.3% of patients cannot undergo imaging due to claustrophobia, and some need sedation to complete the session (Enders et al. 2011). However, sedation involves medical risks and is also not possible for studies that require active participation. The diameter and length of the imaging tube, as well as the weight limit of the patient table, set physical limits on the size of the person who can undergo imaging, which is why studies often have BMI limits (Carucci 2013). It is also important to consider whether the subject has metal implants, piercings, a pacemaker, or anything else that could pose a risk or limit participation. Nowadays, pacemakers and other cardiac implantable electronic devices (CIEDs) do not directly prevent patients from undergoing clinical imaging (Bhuva et al. 2024; Russo et al. 2017). However, these devices most likely restrict patients from participating in research studies. For fMRI and MRI studies, it is crucial to check the latest guidelines and restrictions to determine a subject's suitability. (Shellock and Spinazzi 2008)

Despite its limitations, fMRI offers high spatial resolution without exposing patients to radiation. Although its temporal resolution is lower than that of EEG and MEG, fMRI can be combined with simultaneous EEG recordings. These measurements provide insight into the electrical activity of the cerebral cortex and the brain's hemodynamic responses, including those in the inner brain structures. However, this combined method is complex and challenging, increasing the number of artifacts in fMRI data and complicating analysis and interpretation. (Warbrick 2022) Nonetheless, fMRI's strengths, such as its ability to explore neural processes and the neural correlates of emotions, make it a key tool.

Brain regions in emotional processing

The imaging and analysis methods mentioned can help locate the specific brain regions activated during the processing of emotional stimuli. Identifying these regions is crucial for understanding how emotions manifest at the neurological level. In this study, pinpointing these regions allows us to look at the similarity between the emotional assessments produced by GPT-4V and those created by humans at the neural level.

Different brain regions are involved in arousing and regulating various emotional states. Numerous fMRI studies and combinations of fMRI and EEG measurements have shown that the dorsolateral prefrontal cortex (DLPFC), ventrolateral prefrontal cortex (VLPFC), anterior cingulate cortex (ACC), and posterior parietal cortex (PPC) support the perception and regulation of emotions (Fang et al. 2024). Additionally, the amygdala and hypothalamus play a key role in the emergence, appearance, and experience of negative emotions, such as shame, sadness, fear, and anger. (Banks et al. 2007; Parvizi et al. 2022)

2.3 Studying Emotions Through Stimulus-evoked Responses

For a long time, psychological and neuroscientific research has aimed to provoke different emotional states in lab settings using controlled and repeatable stimuli. One of the most widely used methods is presenting photographs to test subjects, because photographs are an easily standardized and repeatable stimulus format. Commonly used sets of pictures include the Affective Picture System (IAPS) and the Nencki Affective Picture System (NAPS), which provide a wide range of visual stimuli whose emotional properties, such as valence (positivity or negativity), arousal (intensity of arousal), and sense of control, have been evaluated by a large group of participants. This lets researchers choose emotionally neutral, positive, or negative stimuli for their experiments. (Uhrig et al. 2016)

Along with images, film clips have been widely used in emotional research because they can evoke emotions through their dynamic nature and multiple channels by combining visuals, sound, music, and narration. One of film's biggest strengths is its ability to reflect real-life situations more closely than static images, leading to stronger and more nuanced emotional responses. Furthermore, the technical and artistic elements of films, like editing, music, and camera angles, can enhance emotional impact beyond what individual images can achieve. Several studies have shown that film clips are effective and reliable at evoking emotions in experimental research. (Uhrig et al. 2016)

Recently, research has highlighted the need for more natural and multimodal stimuli that better reflect everyday experiences and the emotions they bring about. Unlike strictly controlled, single images or words limited to lab conditions, natural stimuli can be dynamic, multisensory, and rich with context. For instance, they can include videos with dialogue or situations in which visual and auditory information are combined. (Heini Saarimäki 2021)

Multimodal stimuli allow a more diverse and realistic evaluation of emotional responses because emotions in real life are rarely generated by a single stimulus. Instead, they are generated by the combined effects of many simultaneous sensory factors. For example, facial expressions, tone of voice, background music, and the visual environment all contribute to the formation of emotional states. (Gerdes et al. 2014) This is particularly crucial when evaluating the ability of multimodal large language models (MLLMs) to interpret emotions in a multimodal context, including text, images, audio, and video.

The research section of this master's thesis uses three natural emotional stimulus datasets that have been proven to evoke emotions effectively. These datasets include the Nenchki Affective Picture System (NAPS) (Marchewka et al. 2014), the Tettamanti video set (Tettamanti et al. 2012), and the Megaperception video set, which has been used in our previous studies (Santavirta et al. 2024; Santavirta, Wu, et al. 2025). Section 3 "Materials and Methods" will discuss these datasets in more detail. These stimulus sets allow us to examine the effects of both single images and dynamic, multimodal stimuli on emotion processing. When choosing stimuli, we focused on the diversity and naturalness of the content to best showcase the emotional reasoning abilities of the GPT model in various contexts.

From the perspective of artificial intelligence, especially regarding MLLMs, these stimulus formats offer a unique opportunity to assess a model's ability to interpret emotional content in different forms. This includes visual media, such as images and videos, as well as textual content, such as video transcripts.

2.4 Artificial Intelligence and Large Language Models

Basic structure and function of a neural network

Artificial neural networks (ANNs) are machine learning models inspired by biological neural networks. The similarities to biological neural networks are mostly metaphorical. The basic structure of an ANN is based on the idea that the network consists of several interconnected

units called artificial neurons. Warren McCulloch and Walter Pitts successfully developed the first artificial neural network in 1943 (McCulloch and Pitts 1943).

In an artificial neural network, each artificial neuron receives inputs, calculates a weighted sum, and produces the result using an activation function. Activation functions are central to an ANN, as they determine the accuracy, efficiency, and performance of the network. Many activation functions have been developed for different applications. The most common ones include the sigmoid, hyperbolic (tanh), and rectified linear unit (ReLU) functions, which are non-linear. These functions allow neural networks to learn complex, non-linear relationships from data. (Chaudhary et al. 2025; Mostafanejad 2024)

An artificial neural network usually has three parts: an input layer, a hidden layer, and an output layer. The input layer receives the original raw data, such as numerical representations of text or pixel values of an image and feeds the weighted sum to the first hidden layer. The hidden layers perform various intermediate calculations, combining and modifying the values produced by the previous layer. The depth of an artificial neural network is determined by the number of hidden layers. An ANN is called a deep neural network if it has at least two hidden layers. The deeper the ANN, the more complex features it can recognize. After the last hidden layer, the weighted sums go to the output layer, which produces the final prediction or classification. (Pires et al. 2025; Chaudhary et al. 2025)

Learning in artificial neural networks is based on adjusting the weights of the model to improve predictions. Many training methods exist, but they generally fall into three main learning types: supervised learning, unsupervised learning, and reinforcement learning. One effective method is backpropagation, which occurs in two stages: the forward pass and the backward pass. In the forward pass, the input moves through the network, and the final prediction is calculated in the output layer. During the backward pass, a separate loss function measures the prediction error by comparing the correct answer to the predicted result. Then, the weights of the artificial neural network are updated using optimization algorithms. This process repeats several times until the ANN learns to detect complex structures in the data, and the loss function value is minimized. (Schmidgall et al. 2024; Mienye and Swart 2024)

Over time, the development of artificial neural networks has produced different architectures, each with unique characteristics. For instance, convolutional neural networks (CNNs) are effective in image recognition, while recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are designed to handle temporal data, making them suitable for

applications like speech recognition. The latest application of artificial neural networks is the Transformer architecture, which has revolutionized natural language processing and enabled the creation of large language models (LLMs). (Chaudhary et al. 2025; Mienye and Swart 2024)

Introduction to LLMs

Large language models (LLMs) have quickly become a central topic in technology and society. Following the recent release of artificial intelligence applications by OpenAI, Google, and Microsoft in recent years, LLMs have entered the public consciousness. However, the creation of language models goes back to statistical language models (SLMs) developed in the 1990s. In the 2000s, neural network-based pretrained language models (PLMs) appeared. Today's LLMs are improved versions of PLMs with many more parameters and larger training datasets. They can also generalize their knowledge more broadly across different contexts. (Zhao et al. 2025)

Current LLM models rely on deep learning, specifically the Transformer architecture, first introduced by Vaswani et al. in 2017 (Vaswani et al. 2023). Typically, LLMs consist of hundreds of millions or even billions of parameters and are trained on vast amounts of text data.

The Transformer architecture consists of several stacked layers of self-attention, which allow tokens in the input text to be processed simultaneously and in context. First, the input text is processed through tokenization, which breaks the text into smaller units (tokens) that can be modeled for their meaning. However, the self-attention mechanism does not include information about the tokens' positions in the text. Therefore, positional encoding is used to provide this information to the model. (Naveed et al. 2024; Zhao et al. 2025; Raffel et al. 2023)

The attention mechanism calculates a weighted value for each token pair, considering how one token relates to another. This helps maintain context, even in long and complex text. The representation values generated in the different layers are then combined and fed to subsequent layers, and then to the decision-making component. After these steps, the model produces a response to the original input text. (Tutek and Šnajder 2022)

Transformers-based LLM models, such as the Generative Pretrained Transformer (GPT) suite developed by OpenAI, are made to handle a wide variety of language tasks without needing extra fine-tuning. These models can answer questions, summarize text, generate text, and analyze visual input, which is especially notable in multimodal models like GPT-4. (OpenAI et al. 2024)

Multimodal reasoning in emotion recognition

Traditional large language models (LLMs) have their limitations. For instance, they cannot process non-linguistic information, including images, audio, and video. LLMs operate in a purely textual space and excel at natural language processing (NLP) tasks. Similarly, large visual models (LVMs) can recognize visual features, but they usually have weaker abstract reasoning skills compared to LLMs. The combination of these two types of models has led to the development of new types of multimodal large language models (MLLMs).

Multimodal large language models (MLLMs) can handle and combine different types of information, such as text, images, audio, and video, to provide more meaningful insights. This makes them helpful for tasks like evaluating emotional responses based on multiple sensory inputs.

A typical MLLM structure has three main parts: a modality encoder, a large language model (LLM), and a modality interface. The modality encoder roughly speaking acts like the human sense by receiving and preprocessing raw data, such as images, audio, or video. It converts the raw data into a vector representation that the LLMs can process. The LLM is in charge of reasoning and understanding and thus roughly speaking acts like the "brain" of the model. However, the LLM's reasoning is based on probabilities and guesses that it has formed using statistical principles from the material it has learned. The modality interface connects the other two components, enabling interoperability between different modalities at the semantic level. (Yin et al. 2024; D. Huang et al. 2024)

A modality encoder transforms input modalities into a more compact form that can be processed in feature space. These encoders are trained on large sets of image-text pairs, enabling the resulting information to be associated with a linguistic representation. Since LLMs can only process text, a multimodal system must connect the visual or auditory information with the LLM's understanding capabilities. To this end, a learnable adapter often acts as a link between the visual encoder and the language model. (Yin et al. 2024; D. Huang et al. 2024) Another method is to use separate expert models to convert visual data into text. The interpreted content is then provided to the LLM. While this approach avoids direct matching between different types of information, it may be less adaptable, especially for complex reasoning tasks. Additionally, some MLLM models can generate new content in various formats, such as images, audio, or video. This requires an additional generator component. (Yin et al. 2024; D. Huang et al. 2024)

The ability of multimodal models to handle emotional reasoning relies on understanding information from multiple senses. This understanding combines visual cues like facial expressions and body language with text comprehension, situational context, and environmental factors. This allows for interpretations that go beyond simple categories and capture the complexity of emotional states. Therefore, MLLMs, such as GPT-4V, present a valuable opportunity to compare the emotional reasoning abilities of humans and artificial intelligence. This master's thesis explores how well MLLM assessments align with human emotional experiences and their neural representations.

3 Materials and Methods

Stimuli

To assess GPT-4's ability to predict human emotional responses to images and videos, we tested its performance on a wide range of emotionally evocative images and video material used in previous studies. For the video material, we collected new, previously unpublished emotion ratings from humans to make sure that GPT-4 did not have the access to the original human ratings.

As video stimuli, we selected two independent sets of short videos, mainly derived from mainstream Hollywood movies (video dataset 1 (VD1): 234 clips and video dataset 2 (VD2): 120 clips). VD1 has been previously validated to map brain basis of socioemotional perception (Nummenmaa et al. 2023; Santavirta et al. 2023; Putkinen et al. 2023; Karjalainen et al. 2017; Lahnakoski et al. 2012) and VD2 has been curated to elicit basic emotions to map their brain representations (Tettamanti et al. 2012; Heini Saarimäki et al. 2016). The videos of VD1 were 10.5 seconds long on average (range: 4.1-27.9 seconds), and all videos of VD2 were cut to 9.6 seconds. The total duration of all video material was 60 minutes. See **Table SI-1** for descriptions of the movie clips in VD1, and **Table SI-1** in the original publication (Tettamanti et al. 2012) for the clip descriptions of VD2.

As image stimulus, we selected the standardized Nencki Affective Picture System (NAPS BE) (Riegel et al. 2016). NAPS BE contains a total of 510 realistic, high-quality images divided into five categories: people, faces, animals, objects and landscapes. This dataset has been widely used in previous emotion research and found to elicit consistent emotional responses in observers (Putkinen et al. 2023; Riegel et al. 2016; Horvat et al. 2022; Riegel et al. 2017). To limit the stimulation time in the fMRI experiment we selected 300 images as the final stimulus set. See **Table SI-2** for selected NAPS images and associated basic emotions for the current image dataset (ID).

Evaluated emotions and affective dimensions

34 unipolar emotions and 14 bipolar affective dimensions (referred simply as 48 emotions) were annotated from the naturalistic video stimuli. This set of emotions was selected from a previous study that aimed to map data-driven neural representations for emotions in naturalistic video stimuli (Cowen and Keltner 2017). This set covers a wide range of emotions and affective dimensions derived from existing emotion theories, including basic emotions, valence and

arousal, and more complex emotions. We collected human annotations for these 48 emotions for the video stimuli. NAPS BE images have previously been annotated for valence, arousal and six basic emotions, and only these emotions were selected for GPT-4 annotation for the images. See **Table SI-3** for the full list of rated emotions and affective dimensions.

Human reference evaluations for video stimuli

We collected 10 independent emotion ratings from each video clip for each of the 48 emotions. Our previous stability analyses indicate that 10 independent ratings are sufficient for estimating the population average for socio-emotional perception in videos (Santavirta, Wu, et al. 2025). Participants were instructed to rate how strongly they felt the given emotion after watching a video clip. The ratings were collected on a Likert scale from 1 to 9, with 1 being "not at all" and 9 being "very much". To minimize cognitive load and ensure the attention of the participants, each participant rated only 9 – 10 emotions across a subset of the stimuli (30 – 39 videos), which took approximately 30 minutes to complete. The order of presentation of the videos was randomized for each participant to ensure that the order of the stimulus does not bias the population-level results.

The human rating data for VD1 included 315 participants from 31 nationalities. See **Table SI-4** for the full list of participants' nationalities. 49.5 % of the participants were females, and the average age was 33.1 years (range: 18-67 years). Participants' self-reported ethnicities were: Black (61.9%), White (26.7%), Mixed (7.0%), Asian (3.8%), and Other (0.6 %). 13 participants were excluded from the analyses based on visual data quality control. The human rating data for VD2 included 204 participants from 34 nationalities. 48.0% of the participants were females, and the average age was 30.7 years (range: 18-73 years). Participants' self-reported ethnicities were: White (47.1%), Black (40.7%), Mixed (4.9%), Asian (4.4%), and other (1.5%). 14 participants were excluded from the analyses based on visual data quality control.

Human reference evaluations for image stimuli

The NAPS images that were used in this study have been previously annotated for valence, arousal and basic emotions (Riegel et al. 2016). In that study, 124 healthy volunteers annotated the images for the basic emotions, valence and arousal. 54.0% of the participants were female, and the average age was 23.0 (range: 19-37 years). These annotations were used as human reference instead of collecting new annotations.

GPT-4 emotion prediction protocol

We aimed to design GPT-4's input prompt to mimic the instructions given to humans as closely as possible while considering the limitations of GPT-4. Both humans and GPT-4 were simply asked to rate "To what extent this makes you feel [emotion]?" when provided with a video or an image. Where a single human only rated a subset of the emotions from one video or image, GPT-4 was instructed to provide a numerical rating between 1 and 9 for each of 48 emotions in one response. GPT-4 ratings were collected using the GPT-4 application programming interface (API), with each video clip and image presented as a separate query. The order of queries is irrelevant for GPT-4 API as it does not store consecutive requests by default, unlike ChatGPT, which remembers previous conversations that could bias the following responses (https://community.openai.com/t/does-the-open-ai-engine-with-gpt-4-model-remember-the-previous-prompt-tokens-and-respond-using-them-again-in-subsequent-requests/578148). This allowed us to ensure that each request is independent of previous requests. GPT-4 refused to provide ratings for some images and videos (primarily with sexual content), presumably due to content moderation (response: "I'm sorry, but I can't help with this request"). However, some ratings could still be obtained when the same items are presented several times. If GPT-4 still refused the items after several repetitions, the data also from humans were excluded from analyses.

GPT-4 is a stochastic model and doesn't give the same output when prompted with the same prompt repeatedly. In our previous study on social perception of GPT-4, the results indicated that the accuracy of the GPT-4 ratings increase when the ratings are collected repeatedly and then averaged (Santavirta, Wu, et al. 2025). Therefore, we asked for responses to each video clip and image ten times. Emotion evaluations were also more consistent with the human ratings for emotions when collecting several rounds of data, but a ceiling was reached after a few rounds of data collection (see **Figure 2**). Ten rounds of ratings were ultimately collected, and the averages over all rounds were used to compare the GPT-4's emotion ratings with the human participants' ratings. The GPT-4 rating data was collected in May 2025 using the GPT-4 "gpt-4.1-2025-04-14" model (https://platform.openai.com/docs/models/gpt-4.1).

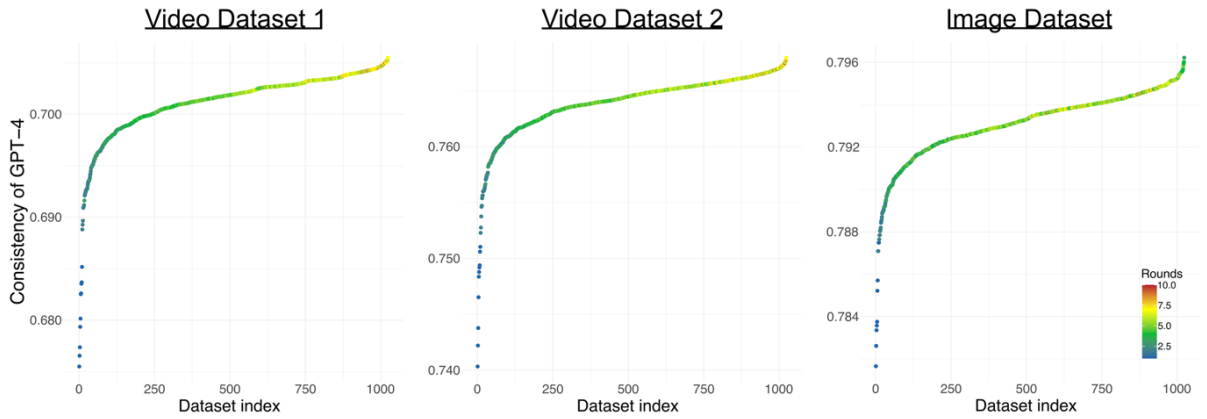

Figure 2. Collecting multiple independent evaluations for the same stimuli increased the accuracy of GPT-4 ratings. We calculated how the average emotion evaluations of GPT-4 correlate with the human average rating (consistency of GPT-4) when the GPT-4 emotion ratings are calculated from single evaluations or as an average of multiple independent GPT-4 evaluations (from 2 to 10). The y-axis shows the correlation between GPT-4 and human ratings, while the x-axis shows the dataset index (we calculated the average GPT-4 evaluations for all possible combinations drawn from 10 independent GPT-4 Evaluations). The color gradient shows that increasing the number of independent GPT-4 evaluations when calculating the average also increased the rating consistency with humans. Blue colors indicate that the GPT-4 data for average calculation only included 1-2 independent evaluations, while yellow/brown colors indicate close to 10 evaluations.

GPT-4 video annotation experiment

At the time of data collection, GPT models could not natively process videos. Thus, we extracted eight frames from each video clip evenly based on the video duration which results in minor differences in frame rates for videos with different durations. However, most of the videos were approximately 10-second-long so the extracted frame rate was ~ 0.8 frames/s. Most of the movie clips in VD1 included human speech (the videos in dataset 2 were silent). The VD1 clips were fed to "whisper-1" model (https://platform.openai.com/docs/models/whisper-1) to convert their language content to text. The transcripts were checked and corrected when necessary. For few videos without any speech, GPT-4 provided unreliable transcripts (e.g. "Thanks for watching) which were manually discarded. The extracted video frames and transcript, along with the rating instructions, were sent to the GPT-4 API as a single input prompt. See the section “Prompt for video annotation” in the Supplementary Materials for the specific prompt.

GPT-4 image annotation experiment

In the image experiment, each image was sent to GPT-4 API individually with the rating instructions. See the section “Prompt for image annotation” in the Supplementary Materials for the exact input prompt. GPT-4 was unable to provide ratings for four images despite repeated

requests. These images contained blood and potentially disgusting contents, such as purulent surgical wounds and removed facial skin.

Emotion rating consistency between GPT-4 and humans

First, we analyzed how similarly GPT-4, and humans rated the evoked emotions. Similar analytical methods were used for both video and image-based emotion data. As some of the original human ratings were collected using a scale from 1 to 7 and some from 1 to 9 int the image dataset, we first normalized all ratings into the same scale from 0 to 10 before statistical analyses. The overall similarity of the emotion ratings was assessed by calculating the Pearson correlation between GPT-4 and human ratings and visually from density plots. To investigate how similarly GPT-4 and humans rated the stimuli for each emotion, we calculated the raw rating distance (in the normalized scale) between GPT-4 and human ratings.

We also calculated the "consistency of GPT-4" as the Pearson correlation with the human average ratings for each emotion. The consistency was then compared with the rating consistency across different human observers. Since humans do not experience or report emotions with 100 % agreement with each other, we calculated "intersubject consistency" and "group-level consistency" as similarity metrics for agreement between different human individuals or groups. These were then used as benchmarks against the consistency of GPT-4. Intersubject consistency was calculated by leaving a single subject out and calculating the correlation between the average ratings of the rest.

Since the human datasets contained ten independent human ratings for each item, the group-level consistency was calculated as the correlation between the ratings of two randomly selected groups of five individuals. All possible combinations were calculated in both calculations, and the average overall combinations were selected as the final metric for intersubject consistency and group-level consistency. This analysis was conducted with the video datasets only since we did not have individual-level data for the image evaluations.

Consistency of the emotional structure between GPT-4 and humans

Next, we investigated how similar the structural representations of the 48 emotions are between GPT-4 and humans to reveal how well GPT-4 can predict the dependencies between all individual emotions. For this, we calculated the Pearson correlation matrices from the emotion ratings for all features to identify the rating associations between each individual emotion. Correlation matrices were calculated separately for each dataset for GPT-4 and human based

ratings. Comparing these matrices enables investigating how structurally similar emotion ratings were between GPT-4 and humans but also investigating the stability of the emotion structure across datasets. Matrix similarity was estimated with Pearson correlation and statistical significance of the similarity of two correlation matrices was tested using a non-parametric Mantel test with 1 000 000 permutations (Mantel 1967). To investigate whether the structural consistency is similar for unipolar emotions and affective dimensions, this analysis was conducted also separately for emotions and dimensions.

Consistency of the evaluation between GPT-4 and Grok-4

We also tested our method with another LLM model (Grok, grok-4-1-fast-reasoning) to ensure the general functionality of LLM models for this type of task. We use the same prompt as with GPT-4 and calculated both structural correlation and correlation of overall similarity between Grok-4 and human evaluations and Grok-4 and GPT-4 evaluations. This test was performed on all stimulus data sets, and all evaluations were collected 10 times, and the evaluations were averaged in the same way as in the main study.

The impact of prompts on GPT-4 evaluations

Because we know that LLM models can be sensitive to the way things are asked, we tested how GPT-4's estimates vary across different prompts. We used a total of three different prompts (original prompt (OP) + two alternative prompts (AP1, AP2)) in the testing. The effect of the prompt on evaluations was tested only on the image dataset. We collected ratings from 10 times for each alternative prompt and averaged the evaluations in the same way as in the main study with the original prompt. We assessed the effect of the prompt by calculating correlations between the averaged estimates of different prompts (OP and AP1, OP and AP2, and AP1 and AP2) and correlations between alternative prompts (AP1 and AP2) and humans.

Neuroimaging experiments

For VD1 and ID, we had previously obtained fMRI data from healthy volunteers. This allowed us to quantify how well GPT-4 based ratings can serve as models for functional neural responses during emotional experiences. We built stimulation models for fMRI data separately from gold-standard human ratings and from GPT-4 ratings and then compared how similar neural response patterns these stimulation models produce.

In the fMRI experiment with videos, we used a validated socioemotional "localizer" paradigm, which provides a reliable method for localizing social and emotional functions (Karjalainen et al. 2017; Nummenmaa et al. 2021; Karjalainen et al. 2019; Nummenmaa et al. 2023; Lahnakoski et al. 2012; Santavirta et al. 2023). The experimental setup and stimulus selection are described in detail in the original study that used the same setup (Lahnakoski et al. 2012). Participants viewed 96 movie clips with a median duration of 11.2 seconds (range: 5.3-28.2 seconds) without breaks and the total duration of the experiment was 19 min 44 seconds. 87 of these movie clips were included in the GPT-4 emotion feature judgment stimulus set. In the fMRI image experiment, the same participants watched the 300 images extracted from NAPS BE during fMRI scanning. Images were displayed in random order for 1.5 seconds each, followed by a black screen with fixation cross for 2-3 seconds before the next image. The total study duration was 22 minutes. Details for the experiment have been reported previously (Putkinen et al. 2023).

Neuroimaging participants

Both fMRI studies were part of a previous multi-session fMRI project run with the same protocol. Study exclusion criteria included a history of neurological or psychiatric disorders, alcohol or substance abuse, body mass index less than 20 or over 30, current use of medications affecting the central nervous system, and standard MRI exclusion criteria. Altogether 104 participants were scanned. Two participants were excluded from further analyses due to anatomical abnormalities on structural MRI, two due to gradient coil malfunction and three due to visible motion artifacts on preprocessed functional neuroimaging data. The final sample was 97 participants, including 50 females, and the average age of participants was 31 years (range: 20-57 years).

Neuroimaging data acquisition and preprocessing

MR imaging was conducted at Turku PET Centre. The MRI data were acquired using a Phillips Ingenuity TF PET/MR 3-T whole-body scanner. High-resolution structural images were obtained with a T1-weighted (T1w) sequence (1 mm3 resolution, TR 9.8 ms, TE 4.6 ms, flip angle 7°, 250 mm FOV, 256 × 256 reconstruction matrix). A total of 467 (video fMRI experiment) and 511 (image fMRI experiment) functional volumes were acquired for the experiment with a T2∗-weighted echo-planar imaging sequence sensitive to the blood-oxygen-level-dependent (BOLD) signal contrast (TR 2600 ms, TE 30 ms, 75° flip angle, 240 mm FOV,

80 × 80 reconstruction matrix, 62.5 kHz bandwidth, 3.0 mm slice thickness, 45 interleaved axial slices acquired in ascending order without gaps).

MRI data were preprocessed using fMRIPrep 1.3.0.2 (Esteban et al. 2019). The following preprocessing was performed on the anatomical T1-weighted (T1w) reference image: correction for intensity non-uniformity, skull-stripping, brain surface reconstruction, and spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al. 2009) using nonlinear registration with antsRegistration (ANTs 2.2.0) and brain tissue segmentation. The following preprocessing was performed on the functional data: coregistration to the T1w reference, slice-time correction, spatial smoothing with a 6-mm Gaussian kernel, non-aggressive automatic removal of motion artifacts using ICA-AROMA (Pruim et al. 2015), and resampling of the MNI152NLin2009cAsym standard space.

Modeling the similarity of the neural representations for emotional processing

To test if the GPT-4-derived stimulus models produce similar neural representations compared to those based on human ratings, we first modeled the BOLD responses measured by fMRI separately with GPT-4- and human-derived emotion regressors and then compared the similarity of the results. We performed simple regressions separately for all rated emotions using SPM12 (Wellcome Trust Center for Imaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm). Emotion ratings were convolved with a canonical double-gamma hemodynamic response function and the resulting regressors were then fitted to the fMRI data (first-level analysis, massive univariate approach). In the image fMRI analysis, a convolved regressor identifying time points with the black fixation screen between the images was added to the model with the emotion regressor. Emotion-specific results were then identified as a contrast between the emotion main effect and the control condition (subtraction: emotion – control). The resulting subject-level β-coefficient maps were analyzed at the group level to identify population-level associations between emotion ratings and hemodynamic responses. One-sample t-tests on the voxel level were used to statistically threshold the population-level results.

Similarity of emotion-specific neural representations between GPT-4 and human-derived stimulation models was investigated in three ways. First, to test the overall spatial similarity of the result distributions, we calculated the Pearson correlation between unthresholded β-coefficient maps for each emotion. Second, we compared the positive and negative predictive values of thresholded GPT-4 results (PPV = true positive / all positive, NPV = true negative /

all negative) to evaluate how reliable the thresholded GPT-4 response maps were for different emotions. PPVs and NPVs were calculated for the positive main effect of emotion ratings, considering the human-derived results as the ground truth. Results are reported for both conservative (voxel-level FWE-corrected, $p < 0.05$) and more lenient ($p < 0.001$, uncorrected) statistical thresholds.

Mapping the cumulative neural responses to all studied emotions reveals the overall neural circuit associated with emotions. This approach involves identifying how many emotions associate with neural responses in each brain area. This analysis allows summarizing the results and distinguishing brain areas that are broadly tuned by emotions (associate with many emotions) from those with narrower response profiles (associate with few emotions). The cumulative result maps were formed by binarizing the statistically thresholded positive contrasts ($p < 0.001$, uncorrected) for each emotion and then calculating the sum of these brain maps. Cumulative results were calculated separately for GPT-4 and human-derived results to enable comparison between them.

4 Results

Similarity of the emotion ratings for video stimuli

The overall correlation calculated over 48 emotions between the GPT-4 ratings, and the human average was 0.71 for VD1 and 0.77 for VD2 (**Figure 3**), indicating robust similarity in the ratings.

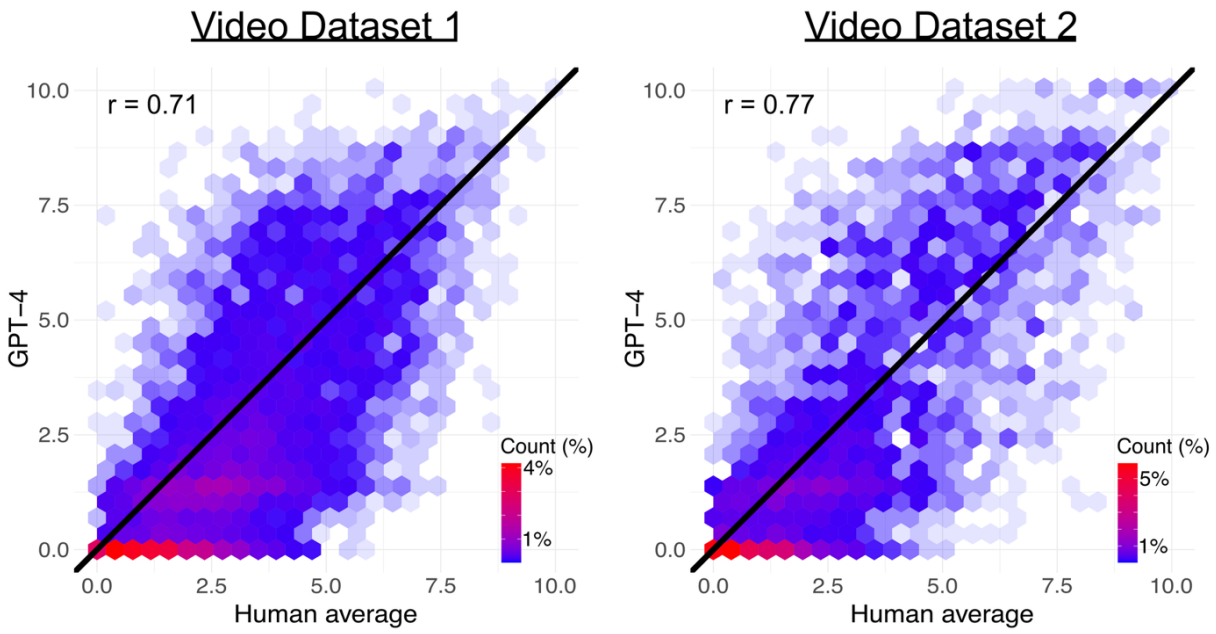

Figure 3. The overall similarity of the emotion ratings between GPT-4 and humans for video stimuli. The x-axis shows the average human ratings for each item, while the y-axis shows the GPT-4 ratings. The color gradient shows the density of the individual data points, with red representing the densest area and the lightest blue (transparent) representing the sparsest area.

Figure 4 shows the distance between the GPT-4 rating and humans (calculated as scaled GPT-4 ratings – scaled human ratings) for the video stimuli. A distance of 1 indicates 10% rating difference in the original rating scale. In VD1, the median distance was -0.75 and range was from -3.00 [Identity] to 1.02 [Certain]. The median distance was between -1 and 1 (indicating under 10% difference) for 34 out of 48 emotions. In VD2, the median distance was -0.38 and range was from -2.25 [Boredom] to 2.25 [Interest]. The median distance was between -1 and 1 (indicating under 10% difference) for 39 out of 48 emotions. Overall, the distances were small in both datasets, but generally GPT-4 slightly underestimated the human ratings (median distance under 0 for 42 emotions in VD1 and 32 emotions in VD2).

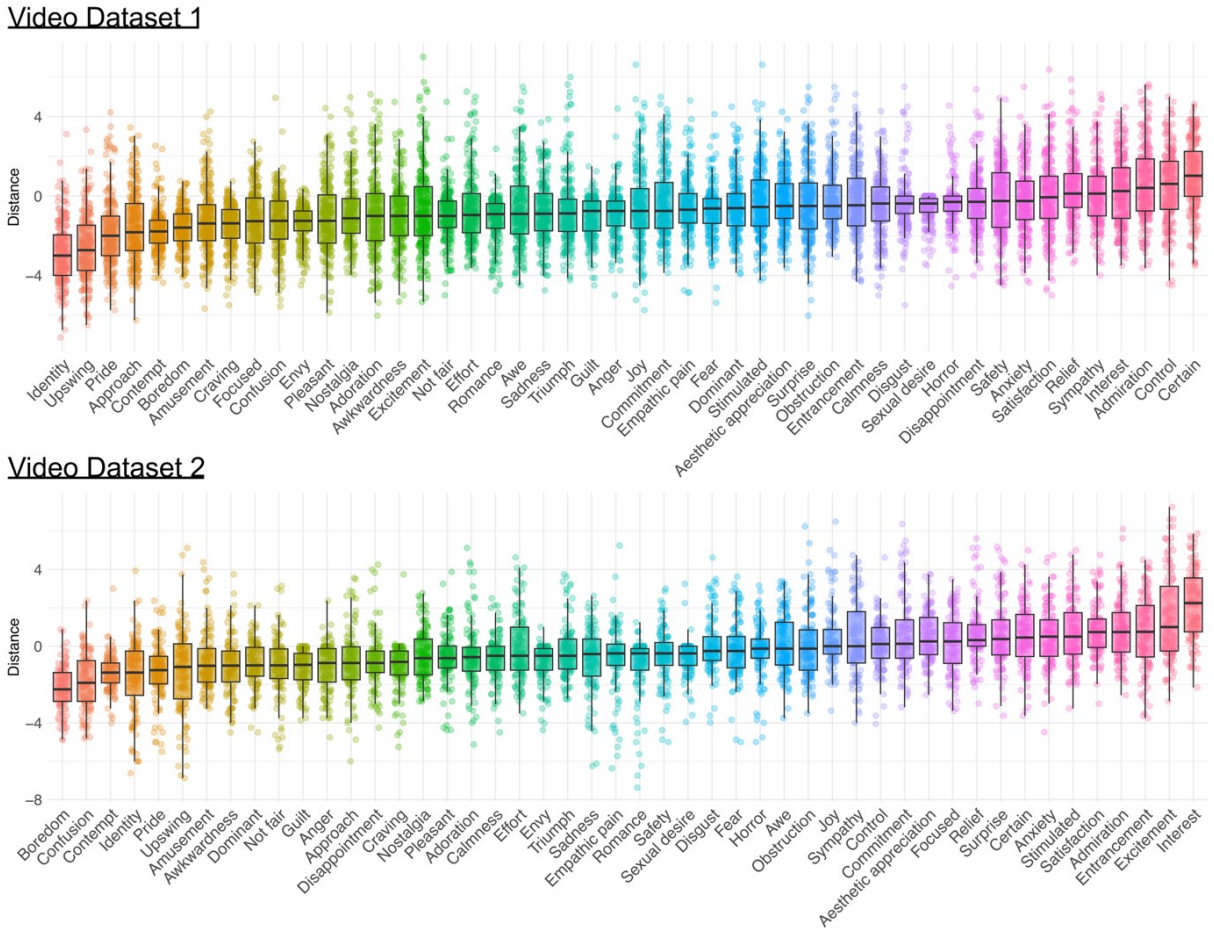

Figure 4. Boxplots of emotion-specific distances between human and GPT-4 ratings. The distances are calculated on the normalized scale (min_{dist} = 0, max_{dist} = 10). Negative distances indicate that GPT-4 underestimates the ratings compared to humans, and positive distances the other way around. Individual points indicate the rating of distances for single videos. A distance of 1 indicates a 10% difference (of the total scale length) in the original ratings and 10 indicates total disagreement.

Consistency of GPT-4 compared to the rating consistency between humans for videos

Figure 5 shows the correlation between GPT-4 ratings (consistency of GPT-4) against the mean correlation between two groups of five human participants (group-level consistency). In VD1, the median consistency of GPT-4 was 0.65 (range: 0.20 [Entrancement] – 0.88 [Empathic pain]). The median human group-level consistency was 0.62 (range: 0.15 [Obstruction] – 0.87 [Empathic pain]) and median intersubject consistency was 0.43 (range: 0.10 [Obstruction] – 0.72 [Empathic pain]). The consistency of GPT-4 was higher than the group-level consistency for 73% of emotions and higher than the intersubject consistency for 96 % of emotions.

In VD2, the median consistency of GPT-4 was 0.79 (range: 0.25 [Focused] – 0.95 [Disgust]). The median human group-level consistency was 0.79 (range: 0.15 [Contempt] – 0.94 [Romance]) and median intersubject consistency was 0.61 (range: 0.10 [Contempt] – 0.87

[Romance]). The consistency of GPT-4 was higher than the group-level consistency for 46% of emotions and higher than the intersubject consistency for 92 % of emotions.

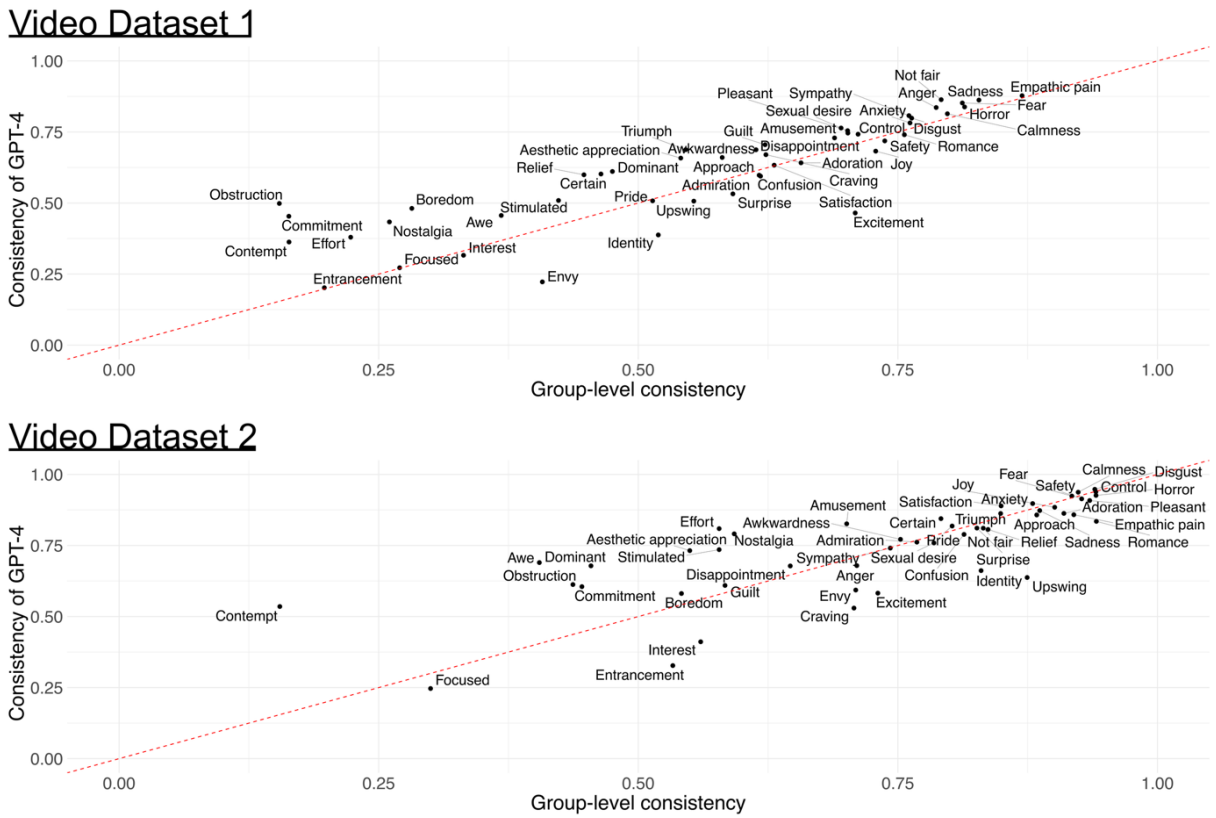

Figure 5. Similarity of emotion-specific ratings between GPT-4 and humans in VD1 (top) and VD2 (bottom). The x-axis shows the group-level agreement among human observers calculated as the mean of Pearson correlations between all passible groups of five independent human annotators. The y-axis shows the correlation between the GPT-4 and human average ratings. Points above the red line (y=x) indicate that the consistency of GPT-4 was higher than the human group-level consistency.

Similarity of the emotion ratings for images

For images, the overall correlation calculated over valence, arousal, and six basic emotions between the GPT-4 ratings, and the human average was 0.80 (**Figure 6**, left panel), indicating robust similarity in the ratings. The feature-specific rating distances on the normalized scale between GPT-4 and humans are shown in the right panel of **Figure 6**. GPT-4 slightly overestimated the human ratings, but overall, the median distances were small in the ID as well. (Valence: -1.34, Anger: 0.00, Disgust: 0.00, Happiness: 0.13, Fear: 0.31, Sadness: 0.68, Surprise: 0.89, and Arousal: 1.61). The median distance was between -1 and 1 (indicating under 10 % difference) for all basic emotions, while distances for valence and arousal where slightly larger.

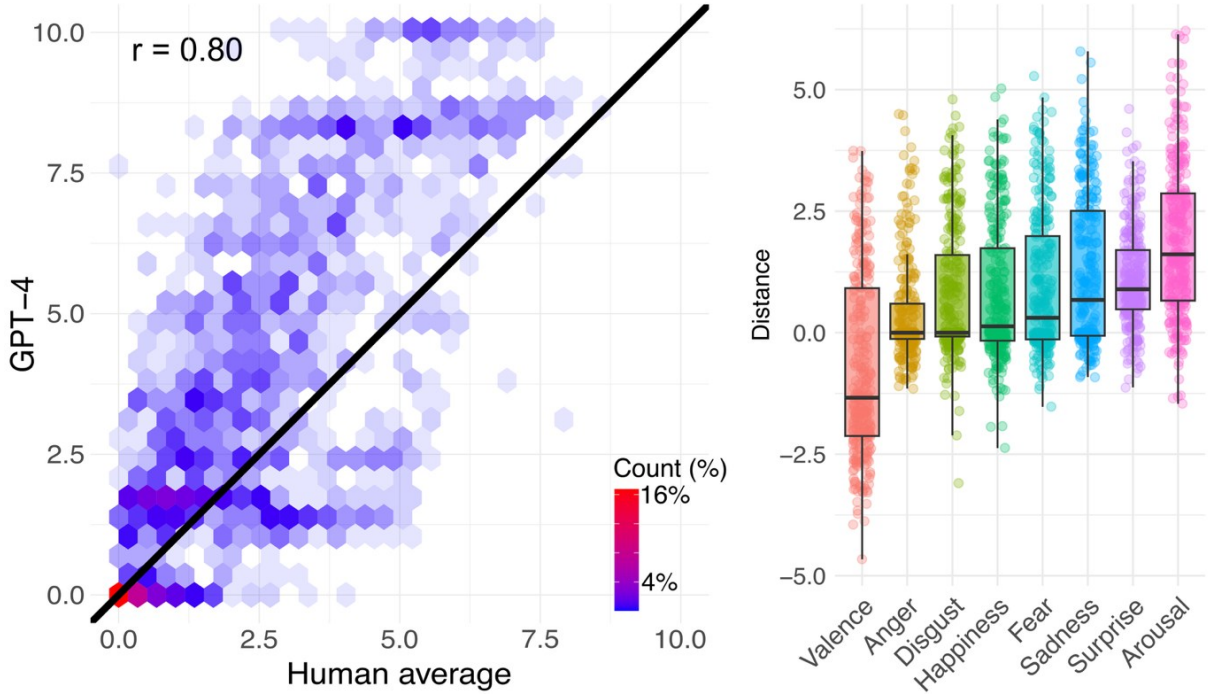

Figure 6. The similarity of the emotion ratings between GPT-4 and humans for images. The scatterplot in the left panel shows the overall correlation of the emotion ratings between GPT-4 and humans (similar to **Figure 3** for videos). Boxplots on the right panel show emotion-specific distances between human and GPT-4 ratings on the normalized scale ($min_{dist} = 0$, $max_{dist} = 10$, similar to **Figure 4** for videos).

The convergence of emotional structure between GPT-4 and human evaluations

The structural representation of emotion ratings was consistent between GPT-4 and humans (**Figure 7**). The correlation matrices calculated from the feature-specific ratings were structurally similar between the GPT-4 and human ratings in all three datasets ($r_{VD1} = 0.84$, $p < 9.9*10^{-7}$, $r_{VD2} = 0.88$, $p < 9.9*10^{-7}$, $r_{ID} = 0.97$, $p < 2.4*10^{-5}$, **Figure 7**). The emotion structures were also consistent across the two video datasets. Correlation between $GPT\text{-}4_{VD1}$ and $GPT\text{-}4_{VD2}$ was 0.93, and between $Human_{VD1}$ and $Human_{VD2}$ also 0.93. Correlation between $Human_{VD1}$ and $GPT\text{-}4_{VD2}$ was 0.84, and between $GPT\text{-}4_{VD1}$ and $Human_{VD2}$ was 0.83. Overall, the emotion ratings formed mostly a two-cluster solution around pleasant and unpleasant emotions. The structural convergence was also similar when calculated separately for unipolar emotions and affective emotions (**Figure 8**).

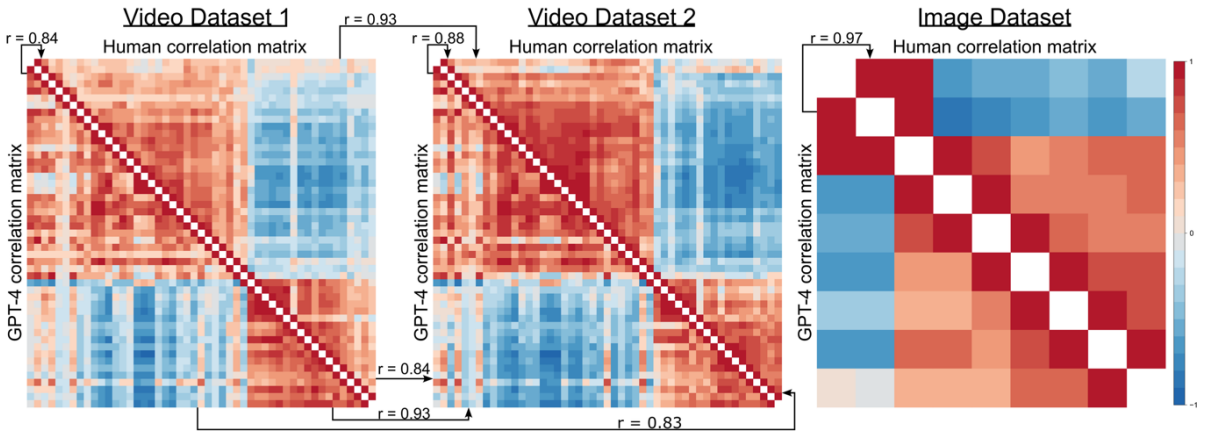

Figure 7. Similarity of the emotion rating structures for each dataset. The correlation matrices for video datasets are generated from emotion ratings of 48 different emotions, while the matrix for the ID is generated from the ratings of 8 emotions (valence, arousal and six basic emotions). Upper triangles show the emotion structure based on human ratings, while the lower triangles show the emotion structure based on GPT-4 ratings. Correlation matrices for video datasets are sorted based on hierarchical clustering of the VD2 human data to enable visual comparison across datasets. See **Figure 8** for separate correlation matrices for unipolar emotions and affective dimensions.

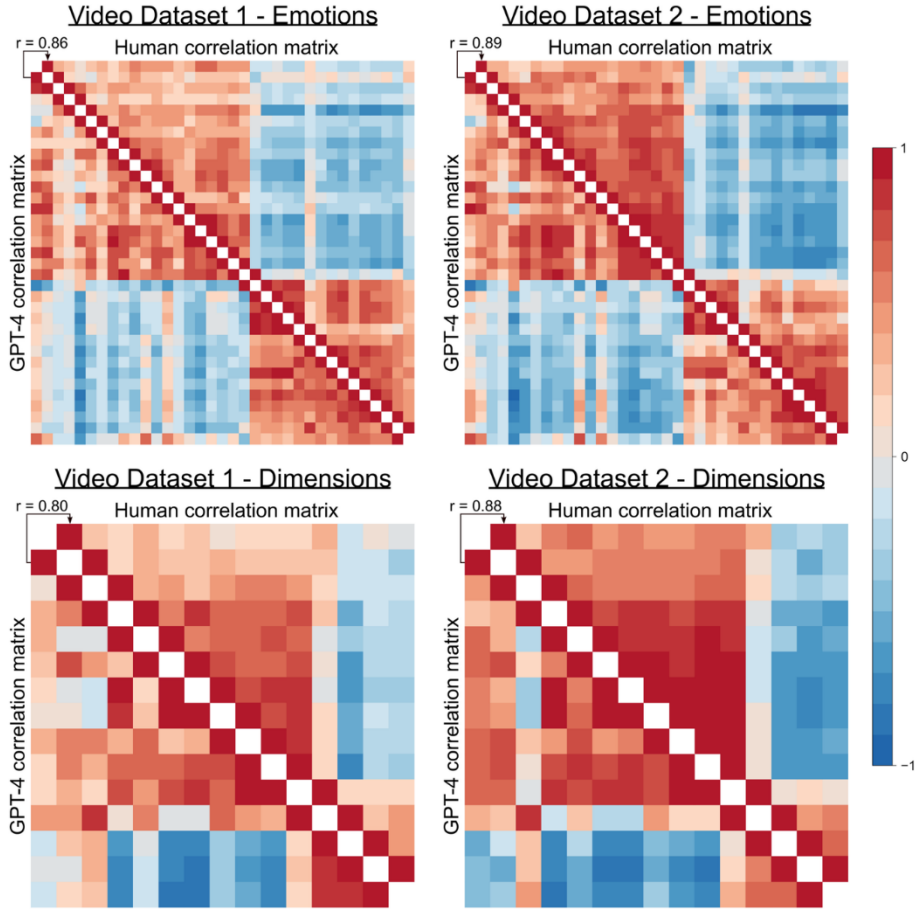

Figure 8. The similarity of the emotion rating structures for each dataset separately for unipolar emotions and affective dimensions. Upper matrices show the rating structure for unipolar emotions, while lower matrices show structures for affective dimensions. Correlation matrices for emotions and dimensions are sorted into the same order based on hierarchical clustering of the VD2 human data to enable visual comparison across datasets. Convergence between GPT-4 and human rating structures were similar for unipolar emotions and affective dimensions.

Similarity of emotion ratings for different LLM models

In the image dataset, the structural similarity between GPT-4 and Grok-4 was 0.88 ($p < 2.2*10^{-5}$) and between Grok-4 and humans was 0.75 ($p < 2.2*10^{-5}$). In the video dataset 1, the structural similarity between GPT-4 and Grok-4 was 0.90 ($p < 9.9*10^{-7}$) and between Grok-4 and humans was 0.80 ($p < 9.9*10^{-7}$). Correspondingly, in video dataset 2, the structural similarity correlations were 0.93 ($p < 9.9*10^{-7}$) and 0.83 ($p < 9.9*10^{-7}$).

When we examined the overall similarity between GPT-4 and Grok-4 and between Grok-4 and humans, we obtained correlations of 0.84 and 0.80, respectively, for the image dataset. The corresponding correlations for video dataset 1 were 0.73 and 0.65, while for video dataset 2, the correlations were 0.84 and 0.68. The results are also summarized in **Table 1**.

Table 1. Comparison between GPT-4.1 and Grok 4.1. The main results are reported with the GPT 4.1 model, but we tested how another LLM model performs on this kind of task.

	GPT 4.1			**Grok 4.1**		
	VD1	**VD2**	**ID**	**VD1**	**VD2**	**ID**
Model	gpt-4.1-2025-04-14	gpt-4.1-2025-04-14	gpt-4.1-2025-04-14	grok-4-1-fast-reasoning	grok-4-1-fast-reasoning	grok-4-1-fast-reasoning
Failed to respond	2.56%	0.83%	1.33%	0%	0%	0%
Data collected	5/2025	5/2025	5/2025	1/2026	1/2026	1/2026
Ratings: Overall correlation with humans	0.71	0.77	0.80	0.65	0.68	0.80
Ratings: Overall correlation with GPT 4.1	-	-	-	0.73	0.84	0.84
Ratings: Similarity of correlation matrices with humans (r)	0.84	0.88	0.97	0.80	0.83	0.75
Ratings: Similarity of correlation matrices with GPT 4.1 (r)	-	-	-	0.90	0.93	0.88

Similarity of emotion ratings across different prompts using GPT-4

The correlations measuring the similarity of emotion ratings between different prompts in the image dataset were 0.999 between OP and AP1, 0.998 between OP and AP2, and 0.999 between AP1 and AP2. When we compared the estimates obtained with alternative prompts to the values given by humans, we obtained correlations of 0.974 between AP1 and humans, and 0.969 between AP2 and humans.

The similarity of the neural representations for emotions between GPT-4 and humans in the video fMRI experiment

Finally, we modelled the hemodynamic responses for emotions based on GPT-4 and human ratings and compared the results. **Figure 9** (top bar plots) shows the similarity of neural response patterns calculated as the spatial correlation between the unthresholded population-level beta coefficient maps that were obtained with GPT-4 and the human-based stimulus models for each emotion. The average spatial correlation of the response patterns was 0.91 (range: 0.67 [Awe] – 0.99 [Calmness]), with a correlation over 0.90 for 65% of the analyzed emotions. We then calculated the consistency of statistically thresholded GPT-4 results by calculating the positive predictive values (PPV) and negative predictive values (NPV) for GPT-4 results, considering the thresholded human-based results as the ground truth. With the conservative threshold (voxel-level FWE-corrected, $p < 0.05$), the mean PPV was 0.80 (range: 0.45 [Anxiety] – 1.00 [Romance]), and with the more lenient threshold ($p < 0.001$, uncorrected), the mean PPV was 0.80 as well (range: 0.43 [Anxiety] – 0.99 [Romance]). The mean NPV for the conservative threshold was 0.97 (range: 0.93 [Nostalgia] – 1.00 [Sexual Desire]) and 0.96 for the more lenient threshold (range: 0.88 [Craving] – 1.00 [Sexual Desire]). The lower bar plot in **Figure 9** shows the emotion-specific PPVs for both statistical thresholds.

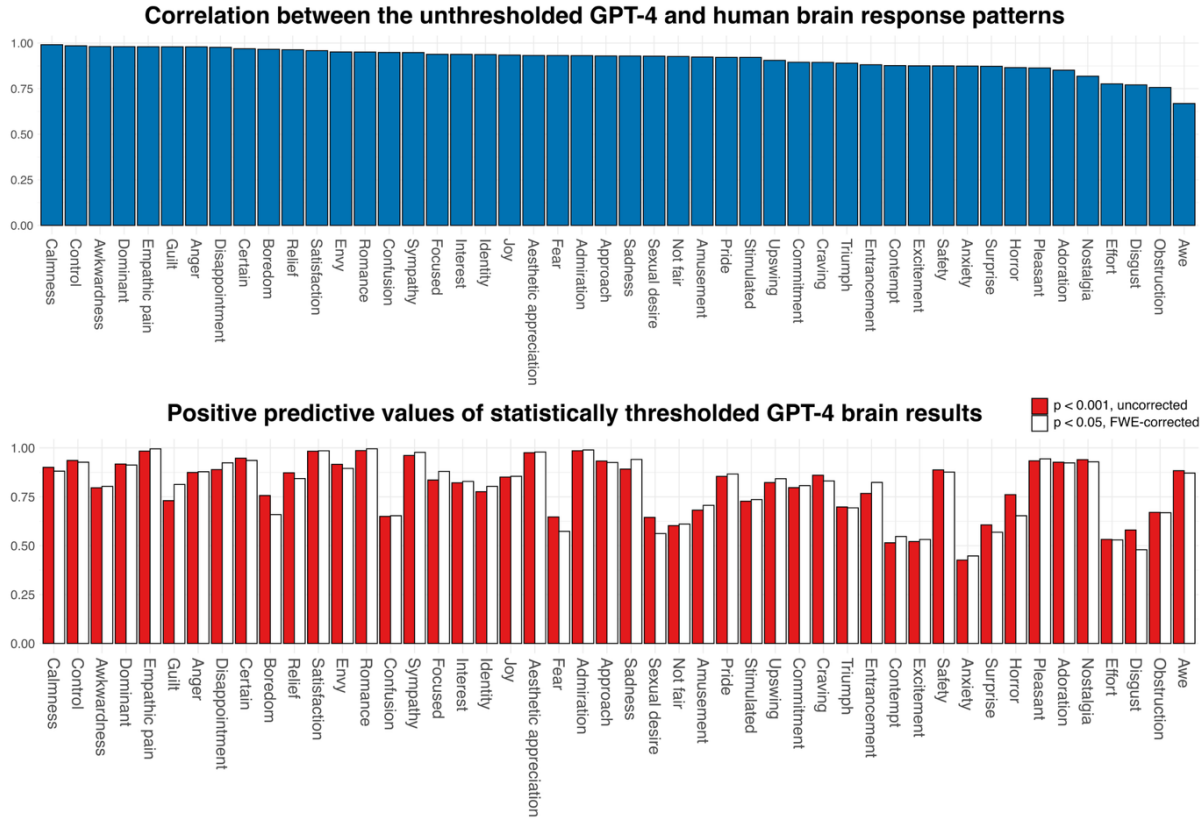

Figure 9. Similarity of the neural response patterns for the video fMRI experiment. The top bar plots show the correlation of the whole-brain response patterns for each emotion (correlations between the population-level unthresholded beta coefficients) between GPT-4 and human-based analyses. The lower graph displays the statistically thresholded positive predictive values (for positive association between BOLD signal and emotion) of the GPT-4 results. PPVs were calculated at two different statistical thresholds: a conservative threshold ($p < 0.05$, voxel-level FWE-corrected) and a lenient threshold ($p < 0.001$, uncorrected).

<u>The similarity of neural representations for emotions between GPT-4 and humans in the image fMRI experiment</u>

In the image fMRI experiment, the average spatial correlation of the unthresholded beta coefficient maps was 0.97 (range: 0.90 [Valence] – 1.00 [Sadness]), with correlation being above 0.98 for 63% of the analyzed emotions (**Figure 10**, left panel). With the lenient threshold ($p < 0.001$, uncorrected), the average PPV of the GPT-4 results was 0.72 (range: 0.48 [Fear] – 0.89 [Valence]). The average NPV, with the more lenient threshold, was 1.00 (range: 0.99 [Valence] – 1.00 [Happiness]). **Figure 10** (right panel) shows the emotion-specific PPVs. Most findings in the image experiment did not survive the conservative thresholding ($p < 0.05$, voxel-level FWE-corrected) either for human- or GPT-based results making the PPVs and NPVs for the conservative thresholding meaningless.

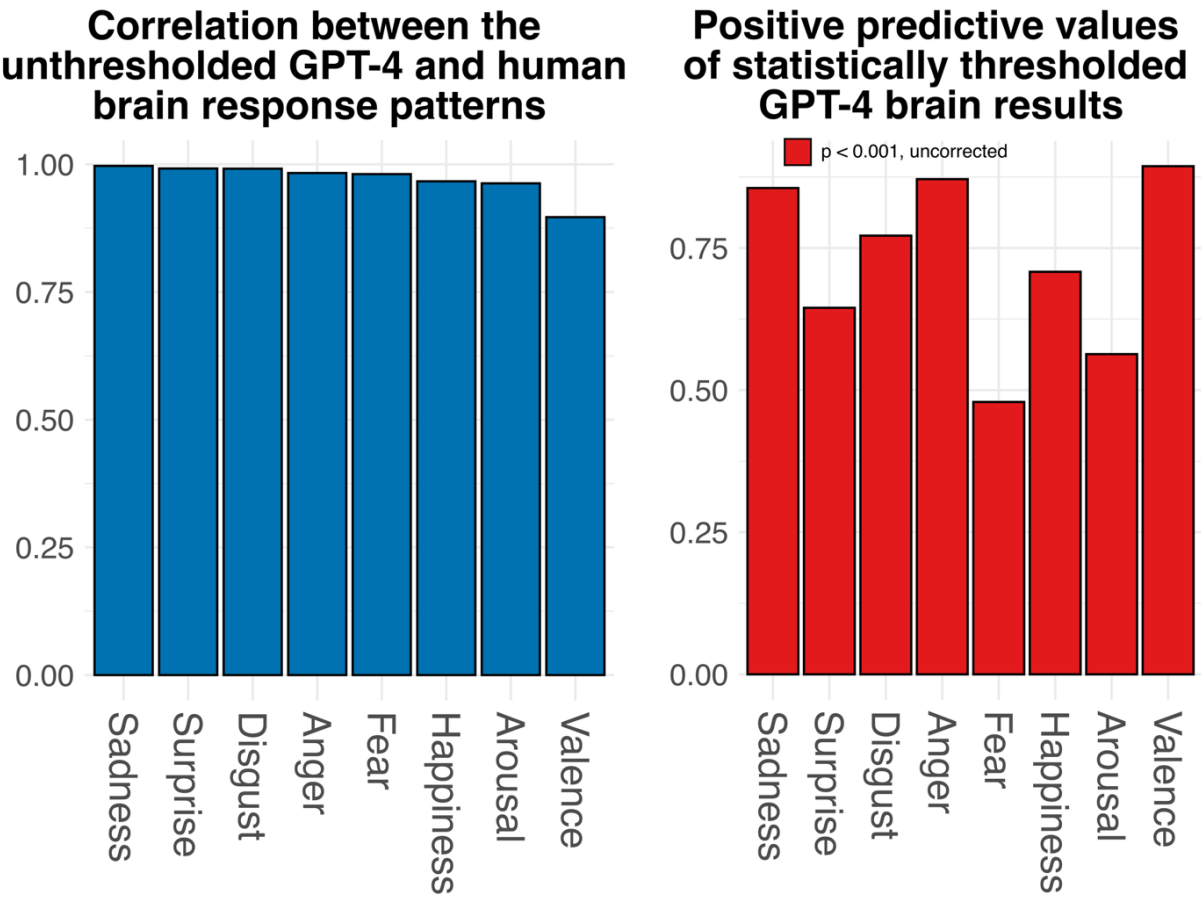

Figure 10. Similarity of the neural response patterns for the image fMRI experiment. The left graph shows the spatial correlation of the whole-brain response maps between GPT-4 and human results. The right graph displays the statistically thresholded positive predictive values (for positive association between BOLD signal and emotion) of the GPT-4 results.

Neural organization of emotional processing based on GPT-4 evaluations

Figure 11 shows cumulative brain activation patterns calculated as the sum over statistically thresholded response patterns for specific emotions ($p < 0.001$, uncorrected) to highlight the areas associated with emotional experiences. Consequently, the maps indicate how many emotions the BOLD signal in each brain region is associated with, indicating the brevity of tuning for different emotions. The cumulative maps for emotions based on GPT-4 ratings demonstrated a high degree of similarity to the cumulative map derived from human ratings in both video and image fMRI experiments ($r_{VD1} = 0.95$, $r_{ID} = 0.85$).

Cumulative maps of emotions in the video fMRI experiment highlight the commonly identified emotion and social perception networks, including *superior temporal sulcus* (STS), *superior temporal gyrus* (STG), *lateral occipitotemporal cortex* (LOTC), *temporoparietal junction* (TPJ), *Fusiform gyrus* (FFG), *inferior frontal gyrus* (IFG), *caudate nucleus* (Cau), *thalamus* (Tha), *Amygdala* (Amy), *Parahippocampal gyrus* (PhG), *Hippocampus* (HC) and *medial*

superior frontal gyrus (SFG). Cumulative maps for the static image experiment highlighted the *orbitofrontal cortex* (OFC), *Fusiform gyrus* (FFG), *Precentral gyrus* (PreCG), and *Postcenral gyrus* (PostCG).

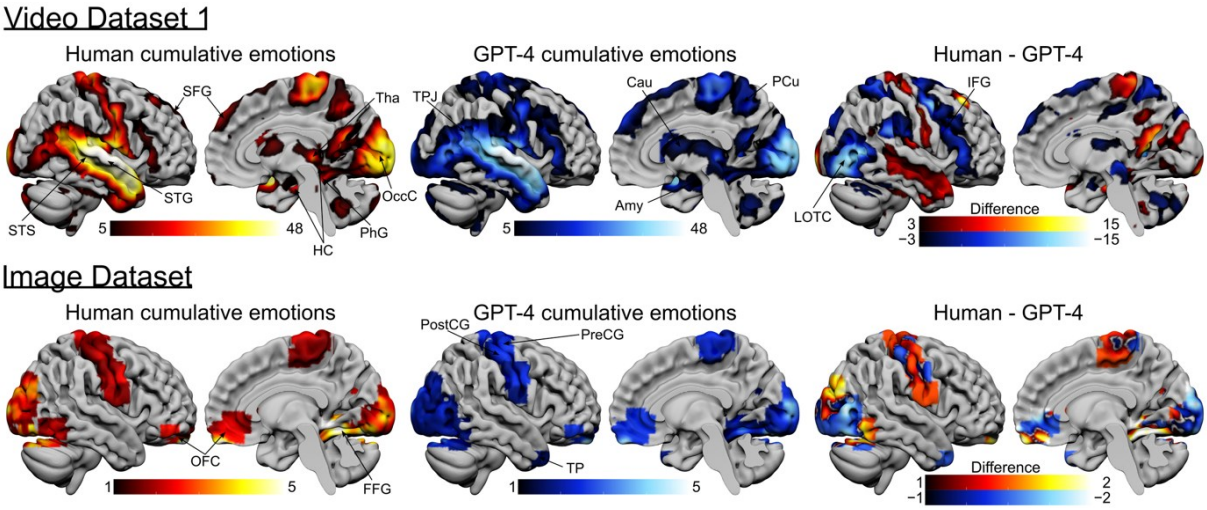

Figure 11. Organization of emotional circuits based on human and GPT-4 emotion evaluations. The surface maps show the number of emotions (out of all 48 emotions for videos and out of 8 for images) that were positively associated ($p < 0.001$, uncorrected) with the BOLD response. The top row shows the cumulative maps derived from the video fMRI experiment, while the bottom row shows results for the image fMRI experiment. The left column shows the cumulative maps based on human stimulus models, and the middle column shows the cumulative results for GPT-4 stimulus models. The right column shows the difference between the two cumulative maps, so that hot colors indicate areas where the human cumulative map shows more associations with emotions compared to the GPT-4 cumulative map. Amy: *Amygdala*, Cau: *Caudate nucleus*, FFG: *Fusiform gyrus*, HC: *Hippocampus*, IFG: *Inferior Frontal gyrus*, LOTC: *Lateral occipitotemporal cortex*, OccC: *Occipital cortex*, OFC: *Orbitofrontal cortex*, PhG: *Parahippocampal gyrus*, PCu: *Precuneus*, PostCG: *Postcentral gyrus*, PreCG: *Precentral gyrus*, SFG: *Superior frontal gyrus*, STG: *Superior temporal gyrus*, STS: *Superior temporal sulcus*, TP: *Temporal pole*, TPJ: *Temporoparietal junction*, Tha: *Thalamus*.

5 Discussion

The main finding of this thesis was that GPT-4 can accurately predict the emotions that humans experience when viewing dynamic video clips or static images. These responses formed a two-cluster structure around pleasant and unpleasant emotions, and the emotion structures accord between humans and GPT-4 as well as across independent datasets. Finally, functional brain circuits associated with emotional processing can be mapped with GPT-4-derived emotion annotations, yielding results that are comparable to those obtained with traditional stimulation models using human evaluations. These results are based on actual visual input, not just textual descriptions of situations, and they were replicated using datasets whose emotion ratings have never been published, preventing "leakage" of the knowledge of the material to GPT-4. These results highlight the GPT-4's capabilities to predict how humans feel when presented with a large variety of visual stimuli, paving the way to broad application potential in emotion research, cognitive neuroscience, and practical solutions.

This research extends our previous work, where we showed that GPT-4 can evaluate the presence of complex social information in dynamic social situations (Santavirta, Wu, et al. 2025). The previous study established that GPT-4 can extract complex social information from visual stimuli, but it did not investigate whether GPT-4 can predict human emotional ratings to such stimuli. Altogether these results show that GPT-4, and most likely other flagship LLMs, can describe the social contents of situations and predict human's emotional experiences for visual stimulation surprisingly well. The overall correlation of emotion ratings between GPT-4 and human was between 0.71 and 0.8 across datasets. Previous studies have already established that LLMs are able to recognize emotions from text descriptions. For example, LLMs can solve standard emotional intelligence tasks (Schlegel et al. 2025) and they exhibit elements of cognitive empathy (Sorin et al. 2024). Especially GPT-4 possesses significant capabilities in understanding situational emotions (Sabour et al. 2024; Tak and Gratch 2024; X. Wang et al. 2023; Tak et al. 2025). Our results go significantly beyond these data because in real-life emotions are mainly evoked by dynamic audiovisual information where emotional and social cues are often complex and context dependent. Such situations are difficult to describe accurately in text, and the present results extend these findings to more life-like conditions and natural emotional perception of visual scenes.

Consistent ratings for experienced emotions between GPT-4 and humans

The GPT-4 emotion evaluations were computed as the average of ten independent evaluation rounds for the same stimuli to ensure the stability of the evaluations. Taking the average of repeated GPT-4 evaluations also increased the consistency between GPT-4 and human evaluations (**Figure 2**) similarly as we observed in the previous study for social information annotation (Santavirta, Wu, et al. 2025). The increase in accuracy is most likely due to filtering out the randomness in the individual responses, or in other words, finding the most likely estimates produced by the internal probability distributions of the model.

With this approach the GPT-4 average ratings had high consistency with the human average annotations for 48 emotions (including 34 unipolar emotions and 14 affective dimensions) in two video datasets and for six basic emotions, valence and arousal for images ($r_{VD1} = 0.71$, $r_{VD2} = 0.77$, and $r_{ID} = 0.80$, **Figure 3** & **Figure 6**). On the level of individual emotions, the median distance between the human and GPT-4 ratings were mostly small (median distance ≤ 1 indicating under 10% difference for 34 emotions in VD1, for 39 emotions in VD2, and for all six basic emotions in ID). Although the overall consistency with humans was high, GPT often gave lower ratings (typically 0 or 1 normalized units lower) than humans. This was most prominent for the cases when humans had evaluated their emotional responses to be mild (normalized rating ≤ 5, **Figure 3**).

This difference was more clearly visible in VD1, which contained videos with dialogue and background music (VD2 only included silent videos). GPT-4 was fed with eight snapshots from a video clip along with the transcript, and therefore it did not have access to all visual and auditory information such as the tone of human voice or background music that may alter the emotional context of the situations (Cowen et al. 2019). Hence, humans had access to the most subtle cues that might have not been available for GPT-4, which could have resulted in GPT-4 responding that the emotional response was completely absent when humans reported slightly higher emotional responses. However, at higher values, this bias vanished (**Figure 3**), indicating that GPT-4 was more accurate in capturing strong emotional experiences. We anticipate that attempts to send even more information (especially auditory information) to GPT-4 might increase its accuracy even further.

Although the overall consistency with human was high for image evaluation as well, GPT-4 seemed to slightly overestimate rather than underestimate emotion ratings compared to humans (**Figure 6**, left graph). Although the NAPS BE dataset was originally curated to evoke basic

emotions, the original human evaluations for this data indicated that the participants did not feel very strong emotions when watching these images. In contrast, our human participants reported also strong emotional responses (normalized rating ≥ 8) for the video stimuli suggesting that videos evoke strong emotions more effectively than images. The reason for GPT's overestimation for images can thus be due to the lack of dynamic content needed to effectively evoke emotions in humans. However, this bias can also be an artefact stemming from differences in data collection and differences in the human rating samples across video and image datasets. While GPT-4 and humans had exactly same instructions to evaluate ratings on a scale 9-point Likert scale from "not at all" and 9 being "very much", the previously collected ratings for images used a visual analog scale (VAS) which could influence human evaluations differently than the Likert scale. Third possibility is that when humans watch images in quick succession, the emotional responses do not change very quickly, while GPT-4 as an artificial system is not temporally constrained in its responses.

Consistency of GPT-4 emotion ratings compared to consistency between humans

Emotional experiences are subjective and variable across individuals despite being broadly consistent. Hence, we benchmarked the consistency of GPT-4 emotion ratings (correlation with the human average rating) by calculating intersubject consistencies (single participants rating correlation with others) and group-level consistencies (rating correlation between two groups of five participants) for the human sample. This analysis indicated that consistency of GPT-4 was, on average, higher than either of these (**Figure 5**). The consistency of GPT-4 exceeded the intersubject consistency for 96% (VD1) and 92% (VD2) of emotions and even exceeded the group-level consistency for 73% (VD1) and 46% (VD2) of emotions. The median consistency of GPT-4 was 0.65 (VD1) and 0.79 (VD2) while the human group-level consistencies were 0.62 and 0.79, respectively. These results indicate that GPT-4 can predict the population average emotional responses for video stimulation with higher accuracy than single human observers and the consistency is even higher than between two groups of five human observers.

Consistent structural representation of emotion ratings between GPT-4 and humans

The structural similarity of emotion ratings between GPT-4 and human observers was high for all three datasets. The structural representation of emotion ratings was identified by calculating the correlation matrices of emotions from the original ratings. Correlation matrices were calculated separately for GPT-4 and human data and also for each dataset to allow cross-

comparison (**Figure 7**). Within datasets the structural representations of emotions were similar between GPT-4 and human ratings (r_{VD1} = 0.84, r_{VD2} = 0.88, and r_{ID} = 0.97). Additionally, the correlation matrices were consistent across VD1 and VD2 indicating that the emotion representations form mainly a two-cluster solution that distinguishes pleasant emotions from unpleasant ones. All cross-correlations between the correlation matrices of the two datasets ($VD1_{human}$ vs. $VD2_{human}$, $VD1_{GPT}$ vs. $VD2_{GPT}$, $VD1_{human}$ vs. $VD2_{GPT}$, $VD1_{GPT}$ vs. $VD2_{human}$) were over 0.8. These results indicate that the emotional space across 48 emotions is stable with different video stimuli and between human and GPT-4 annotations.

Neural circuits of emotional processing modeled with GPT-4 emotion predictions

To demonstrate the utility and reliability of GPT-4 emotion ratings in cognitive neuroscience, we modeled the neural representation of emotion circuits using retrospective fMRI datasets where emotions were elicited to 97 healthy participants with VD1 videos and ID images. The hemodynamic responses were modeled separately with GPT-4 and human ratings to allow comparative analysis of the results. Previous analyses based on ratings revealed high levels of consistency for self-reported feelings, and as expected, this consistency extended to the neural level.

Cumulative activation maps highlighted a broad, human-typical socioemotional processing network in the video stimulation dataset (**Figure 11**, top panel). The cumulative maps were remarkably similar when the data were modeled using GPT-4 versus human emotion ratings (r_{VD1} = 0.95 and r_{ID} = 0.85). This analysis highlighted *superior temporal sulcus* (STS), *superior temporal gyrus* (STG), *lateral occipitotemporal cortex* (LOTC), *temporoparietal junction* (TPJ), *Fusiform gyrus* (FFG), *inferior frontal gyrus* (IFG), *caudate nucleus* (Cau), *thalamus* (Tha), *Amygdala* (Amy), *Parahippocampal gyrus* (PhG), *Hippocampus* (HC) and *medial superior frontal gyrus* (SFG) as the main hubs for socioemotional processing for the VD1 social video stimulation. Cumulative maps for the static image experiment highlighted the orbitofrontal cortex (OFC), Fusiform gyrus (FFG), Precentral gyrus (PreCG), and Postcentral gyrus (PostCG). According to previous studies, these regions are central to emotion elicitation and response, emotion recognition, and especially processing of socioemotional cues, such as facial expressions, speech prosody, and body language (Lettieri et al. 2019; Heini Saarimäki et al. 2016; Koide-Majima et al. 2020; Heini Saarimäki et al. 2025).

When investigating the convergence at the level of individual emotions, the results were also highly consistent between humans and GPT-4. The average correlation of unthresholded beta-

maps for each emotion was 0.91 for VD1 and 0.97 for ID. However, in VD1, slightly lower consistency ($r < 0.80$) was observed for emotion "awe", "obstruction", "disgust", and "effort" which may be due to the ambiguity or their context-related nature. With respect to the ID, the individual correlations for the dimensions of valence and arousal, and for the six emotions, were all above 0.80. The average positive predictive value (PPV) of the GPT-4-based results for VD1 was 0.80 for both conservative (FWE-corrected, $p < 0.05$) and lenient (uncorrected, $p < 0.001$) statistical thresholds and the average negative predictive value (NPV) was 0.97 (FWE-corrected) and 0.96 (uncorrected), respectively. These metrics were highly similar in the ID. This suggests that the GPT-4-based models can predict neural emotion circuits associated with specific emotions with high similarity to the gold-standard human-annotation based models.

The effect of the chosen GPT model and content moderation

The reported GPT-4 annotations were collected using OpenAI's GPT-4.1 (gpt-4.1-2025-04-14) model but we initially collected the evaluations with GPT-4 Turbo model (gpt-4-turbo-2024-04-09). A comparison of the results between these two models is presented in **Table SI-5**. The performance of GPT-4.1 was minimally better compared to the older model. The most significant difference between the models was not their accuracy in making assessments, but rather the increased content moderation in the newer model. While GPT-4 Turbo provided emotional assessments for nearly all stimuli, GPT-4.1 refused to evaluate a few videos (1-2 %), mainly including sexual content, as well as a few images (1%) containing blood, wounds, or nudity. This suggests that the GPT-4.1 model has tightened internal content filters deliberately limiting its use cases with certain material. Content that raises or addresses conflicting moral opinions is also likely excluded from the training data. According to the GPT-4 model documentation, the content filters include, for example, "sexual," "hate," "violence," and "harassment" filters (https://platform.openai.com/docs/api-reference/moderations/object). While these filters may promote the safe and child-friendly use of LLM models, they also limit their usability in studies examining basic human behavior such as sexuality and aggression. This restricts the range of material that can be evaluated and may introduce bias if certain emotional domains (e.g. sexuality) are overlooked systematically.

With closed source-models, such as GPT-4, it is currently not possible to turn off content moderation. Open-source models could be useful tools for research when commercial content moderation significantly limits the use of closed source models. However, currently open-source models are inferior compared to GPT-4 in multiple benchmarks, and they require

significant local computational resources. One major benefit for closed-source models is how easily they can be accessed without strong computational background or resources, making them available for wide audiences. For these reasons closed-source models like GPT-4 may strike a practical balance between performance and accessibility.

Cost-efficiency

The potential economic impact of automated approach to emotion annotation is considerable. For example, in the current study, the total cost of collecting emotion ratings from human observers exceeded 3000 dollars in participant fees, requiring a total of 260 hours of the participants' time (~29 min/participant). In contrast, the collection of GPT-4 estimates for all datasets using the API was fast and costed roughly 100 dollars (around $0.02 per video query and < $0.01 per image query, **Table 2**), or just 3% of the cost of human data. Automated annotations are not only highly cost-efficient, but they also overcome issues such as subject compliance or vigilance. In the future, laborious and expensive human experiments could be more efficiently targeted to test the most promising hypotheses, which could be first generated in pilot studies or in preliminary analyses with LLMs, such as GPT-4, before being fully and specifically tested with human participants.

Table 2. Comparison between GPT-4 models. Preliminary analyses were conducted using the GPT 4 Turbo model, and the main results are reported with the GPT 4.1 model.

	GPT 4.1			**GPT 4 Turbo**		
	VD1	**VD2**	**ID**	**VD1**	**VD2**	**ID**
Model	gpt-4.1-2025-04-14	gpt-4.1-2025-04-14	gpt-4.1-2025-04-14	gpt-4-turbo-2024-04-09	gpt-4-turbo-2024-04-09	gpt-4-turbo-2024-04-09
GPT input	8 frames + transcripts	8 frames	image	8 frames + transcripts	8 frames	image
Total cost per image/video (ten collection rounds)	$0.023 x 10 = $0.227	$0.017 x 10 = $0.172	$0.003 x 10 = $0.027	$0.096 x 10 = $0.963	$0.083 x 10 = $0.830	$0.013 x 10 = $0.130
Failed to respond	2.56%	0.83%	1.33%	0%	0%	0%
Data collected	5/2025	5/2025	5/2025	1/2025	12/2024	3/2025
Ratings: Overall correlation with humans	0.71	0.77	0.80	0.68	0.73	0.81

Ratings: Feature-specific correlations	Mu: 0.65 (0.20 – 0.88)	Mu: 0.79 (0.25 – 0.95)	-	Mu: 0.58 (0.19 – 0.87)	Mu: 0.70 (0.16 – 0.88)	-
Ratings: For how many features, GPT ratings were more reliable population-level estimates than a single human's ratings / the average of five humans?	96% / 73%	92% / 46%	-	96% / 38%	75% / 25%	-
Ratings: Similarity of correlation matrices with humans (r)	0.84	0.88	0.97	0.88	0.92	0.90
FMRI: Feature-specific spatial correlations	Mu: 0.91	-	Mu: 0.97	Mu: 0.89	-	Mu: 0.97
FMRI: Positive predictive values ($p < 0.001$, uncorrected)	Mu: 0.80 (0.43 – 1.00)	-	Mu: 0.72 (0.48 – 0.89)	Mu: 0.74 (0.34 – 0.97)	-	Mu: 0.77 (0.52 – 0.98)
FMRI: Negative predictive values ($p < 0.001$, uncorrected)	Mu: 0.96 (0.88 – 1.00)	-	Mu: 1.00 (0.99 – 1.00)	Mu: 0.90 (0.66 – 1.00)	-	Mu: 0.99 (0.99 – 1.00)
FMRI: Cumulative map similarity (r)	0.95	-	0.85	0.94	-	0.87

Future directions and applications

The results provide a strong foundation for utilizing GPT-4 and other large language models for studying emotions and their neural bases. The neural results demonstrate the potential to automatically annotate complex psychological phenomena for dynamic stimuli and then reliably model functional neuroimaging data with these annotations. This opens new avenues

for automated approaches in large-scale neuroimaging experiments. In multidimensional and laborious stimulus mapping experiments, human annotations could be, at least partially, replaced by GPT-4 estimates (Dillion et al. 2023). This approach would accelerate the mapping of representation spaces, increase statistical power, and significantly reduce costs compared to previous multidimensional projects (Huth et al. 2016; Koide-Majima et al. 2020; Tarhan and Konkle 2020; Santavirta et al. 2023). For example, automated solution would allow re-annotating retrospective large-scale movie fMRI datasets in unprecedented detail and temporal scale easily and cost-efficiently for advanced analyses. Although the focus of this paper was to utilize LLMs for research purposes, automatic and real-time evaluations of complex socioemotional information from videos would have significant application potential in diverse real-life applications in surveillance and security systems, patient monitoring, customer experience analysis, and social robotics development, for example.

Currently, GPT-4 cannot simultaneously process high-frame-rate video with its full audio stream. Future research should attempt to increase the given information, especially auditory input, when prompting video evaluations from LLMs. However, we anticipate that full multimodality for native video input to the LLMs is nearby. While the present study used GPT-4, several other LLMs are available, and new versions of GPT as well as new generations of LLMs will very certainly soon emerge. Our general approach and specific results may provide a rationale for testing both existing and upcoming models: a new interdisciplinary subfield - at the crossroads of psychology, neuroscience and artificial intelligence – could propose standard procedures to evaluate specific LLMs with respect to their capacity to predict the outcomes of human psychological processes.

Limitations

Overall, the results supported the hypothesis that GPT-4 can indeed predict human emotional evaluations for video and image stimulation with accuracy that is higher than agreement between small groups of people. However, our human data consisted of ratings from ten participants for each stimulus. Collecting larger human sample might yield higher confidence in the overall population average and would also allow estimating the consistency between larger subgroups of humans, such as testing LLMs' performance in predicting emotions in different cultures. However, data collection costs would also increase significantly as described above.

The responses produced by LLMs are also sensitive to how the prompts are formulated. Previous studies have demonstrated significant impact of prompt wording on LLM model results (L. Wang et al. 2024), while we have previously found that minor changes in the prompts do not significantly change how GPT-4 evaluate social information from videos (Santavirta, Wu, et al. 2025). However, the same problem applies also to humans, who can interpret even simple instructions or adjectives such as “happy” in vastly different ways. Nevertheless, we stress that we used a consistent prompt that closely mimicked the instructions given to human observers while ensuring that GPT-4 produced its evaluations in a structured format.

Like other LLMs, GPT-4 is trained on large and non-disclosed datasets most likely collected from the internet, books, and other sources. This exposes the models to underlying cultural and societal biases. Consequently, LLMs’ responses could be biased towards perspectives of a specific populations, ignoring the experiences and interpretations of other groups (Demszky et al. 2023). However, research on these biases is inconclusive (Santurkar et al. 2023; Park et al. 2024; Almeida et al. 2024), and the rapid evolution of models makes it difficult to keep up with investigating the properties of any specific version. While the precise impact of these biases on our estimates is unknown, our study focused on predicting population average emotional responses rather than ideological attitudes. Hence, it is reasonable to assume that the impact of potential biases on emotion assessments is minimal. Additionally, the human ratings included participants from over 30 nationalities ensuring diversity in the reference data. The human data for video datasets was collected for this study, and they have not been made public before publications, which ensures that they have not being included in the GPT-4 training.

6 Conclusions

This master's thesis investigated whether multimodal large language models (MLLMs; mainly GPT-4), can predict human emotional responses to natural visual stimuli. In addition, the aim was to determine how well the estimates produced by MLLMs align with human subjective estimates and emotional neural responses measured with functional magnetic resonance imaging (fMRI). By combining multimodal stimulus material, fine-grained emotion assessments, and neuroimaging, this master's thesis succeeded in unifying artificial intelligence, emotion research, and affective neuroscience into a single empirical framework.

RQ1: Can multimodal large language models (MLLMs) estimate human-specific emotions from images and videos, producing human-like evaluations?

Our results provide strong evidence that multimodal large language models (MLLMs) can estimate human-specific emotions from both images and videos, producing ratings similar to those of human raters. GPT-4 was able to predict the intensity of 48 different emotions quite accurately for video stimuli in both video datasets. GPT-4 also managed to predict 8 different emotions well for static images. For both stimulus formats, GPT-4's predictions were found to be very consistent with average human ratings. Significantly, GPT-4 was able to produce higher agreement with averaged human estimates than even the average estimates of a group of five people were consistent with averaged human estimates, **Figure 5**.

Our findings show that GPT-4's emotional state estimates are not limited to simple affective dimensions such as valence or arousal. Instead, MLLM models, at least GPT-4, are able to handle a wide range of different emotional states and predict their relative intensities in a way that produces results that are very close to human self-assessments. This indicates that MLLM models are able to act as human-like emotional state assessors, at least in situations where emotions are represented by a fine-grained, multidimensional feature space.

RQ2: How similar are the structural relationships of the emotion evaluations produced by GPT-4 to human evaluations?

In addition to the similarity of individual emotion ratings, we examined how GPT-4 describes the structural entirety of people's emotional experience. Our analyses showed that the relational structure of emotions derived from the ratings produced by GPT-4 closely reflects the corresponding structure of human ratings. The structural representations of emotions

(correlation matrices) were highly consistent between GPT-4 and human ratings. Furthermore, these were highly stable across two independent video datasets.

The structural similarity we observed suggests that GPT-4 has learned well how emotions co-occur, contrast, and form groups in ways that resemble human emotion representations. These results are consistent with data-driven and dimensional theories of emotion, which view emotions as continuous and interconnected rather than as distinct categories. The presence of similar emotional structures across stimulus datasets suggests that GPT-4's emotion representations are not tied to specific images or videos but rather reflect the more general and stimulus-independent way that people structure emotions in relation to each other.

RQ3: Can the emotion evaluations produced by GPT-4 be reliably utilized in modeling the brain's emotional networks in fMRI data?

Another key contribution of this master's thesis is the demonstration that the emotion ratings predicted by GPT-4 could be utilized in modeling neural responses related to emotional processing. When we used the emotion ratings produced by GPT-4 as stimulus models in fMRI analyses, we predicted neural responses that were comparable to models produced by traditional human ratings. The resulting activation patterns were observed in brain regions commonly associated with emotional perception and regulation, suggesting that the emotional structure captured by GPT-4 corresponds quite well to biologically relevant dimensions of affective processing. These findings demonstrate that the emotion models derived from GPT-4 emotion ratings are not only behaviorally plausible but also neurologically informative.

However, it is worth noting that the use of GPT-4-generated emotion evaluation predictions does not eliminate the need for careful interpretation. fMRI measures indirect neurodynamic responses, and individual differences in emotional experience, cognition, and physiology remain important sources of variation. However, our results indicate that MLLM-based emotion evaluations could reliably complement human-generated emotion evaluations in neuroimaging research.

Overall

Altogether, our study provides the first empirical demonstration that a large language model - GPT-4 - can accurately predict the intensity of 48 distinct emotions that humans feel when viewing a large array of videos and static images. GPT-4's emotional ratings were found to be highly consistent with human average ratings. This consistency was higher than the agreement

between two groups of five human participants for most emotions. The structural representations of emotions were consistent between GTP-4 and human ratings, and also across two independent video datasets. Modelling the neural circuits for emotional processing in fMRI datasets with GPT-4 derived stimulation models predicted similar neural response patterns compared to traditional models based on human annotations. This opens significant possibilities in cognitive and affective neuroscience research where laborious multimodal human annotation can be complemented or sometimes even replaced with LLM-based annotations, allowing also large-scale reanalysis of existing datasets with novel stimulus models. Altogether, our results show that multimodal language models provide a valuable tool for investigating the self-reported human affective experience and its neural basis.

References

Adolphs, Ralph, Lauri Nummenmaa, Alexander Todorov, and James V. Haxby. 2016. "Data-Driven Approaches in the Investigation of Social Perception." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1693): 20150367. https://doi.org/10.1098/rstb.2015.0367.

Aher, Gati V., Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." *Proceedings of the 40th International Conference on Machine Learning*, July 3, 337–71. https://proceedings.mlr.press/v202/aher23a.html.

Ahlfors, Seppo P., and Maria Mody. 2019. "Overview of MEG." *Organizational Research Methods* 22 (1): 95–115. https://doi.org/10.1177/1094428116676344.

Almeida, Guilherme F. C. F., José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. "Exploring the Psychology of LLMs' Moral and Legal Reasoning." *Artificial Intelligence* 333 (August): 104145. https://doi.org/10.1016/j.artint.2024.104145.

Angkasirisan, Thanakorn. 2024. "Naturalistic Multimodal Emotion Data with Deep Learning Can Advance the Theoretical Understanding of Emotion." *Psychological Research* 89 (1): 36. https://doi.org/10.1007/s00426-024-02068-y.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–51. https://doi.org/10.1017/pan.2023.2.

Banks, Sarah J., Kamryn T. Eddy, Mike Angstadt, Pradeep J. Nathan, and K. Luan Phan. 2007. "Amygdala–Frontal Connectivity during Emotion Regulation." *Social Cognitive and Affective Neuroscience* 2 (4): 303–12. https://doi.org/10.1093/scan/nsm029.

Barrett, Lisa Feldman, Batja Mesquita, Kevin N. Ochsner, and James J. Gross. 2007. "The Experience of Emotion." *Annual Review of Psychology* 58 (Volume 58, 2007): 373–403. https://doi.org/10.1146/annurev.psych.58.110405.085709.

Bhuva, Anish, Geoff Charles-Edwards, Jonathan Ashmore, et al. 2024. "Joint British Society Consensus Recommendations for Magnetic Resonance Imaging for Patients with Cardiac Implantable Electronic Devices." Consensus Statement. *Heart* 110 (4): e3–e3. https://doi.org/10.1136/heartjnl-2022-320810.

Bradley, Margaret M., and Peter J. Lang. 2007. "The International Affective Picture System (IAPS) in the Study of Emotion and Attention." In *Handbook of Emotion Elicitation and Assessment*. Series in Affective Science. Oxford University Press.

Bunce, S. C., M. Izzetoglu, K. Izzetoglu, B. Onaral, and K. Pourrezaei. 2006. "Functional Near-Infrared Spectroscopy." *IEEE Engineering in Medicine and Biology Magazine* 25 (4): 54–62. https://doi.org/10.1109/MEMB.2006.1657788.

Cacioppo, John T., Wendi L. Gardner, and Gary G. Berntson. 1999. "The Affect System Has Parallel and Integrative Processing Components: Form Follows Function." *Journal of Personality and Social Psychology* (US) 76 (5): 839–55. https://doi.org/10.1037/0022-3514.76.5.839.

Carucci, Laura R. 2013. "Imaging Obese Patients: Problems and Solutions." *Abdominal Imaging* 38 (4): 630–46. https://doi.org/10.1007/s00261-012-9959-2.

Celeghin, Alessia, Matteo Diano, Arianna Bagnis, Marco Viola, and Marco Tamietto. 2017. "Basic Emotions in Human Neuroscience: Neuroimaging and Beyond." *Frontiers in Psychology* 8 (August). https://doi.org/10.3389/fpsyg.2017.01432.

Chaudhary, Sumit, Adarsh Sahani, Shadab Akhtar, and Shivam Pandey. 2025. "A Comprehensive Review of Artificial Neural Networks: Architectures, Learning Algorithms, and Real-World Applications." *Journal of Scientific Innovation and Advanced Research* 1 (1): 1–10.

Cohen, Jonathan D., Nathaniel Daw, Barbara Engelhardt, et al. 2017. "Computational Approaches to fMRI Analysis." *Nature Neuroscience* 20 (3): 304–13. https://doi.org/10.1038/nn.4499.

Constable, R. Todd. 2023. "Challenges in fMRI and Its Limitations." In *Functional Neuroradiology: Principles and Clinical Applications*, edited by Scott H. Faro and Feroze B. Mohamed. Springer International Publishing. https://doi.org/10.1007/978-3-031-10909-6_22.

Cowen, Alan S., Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. 2019. "Mapping 24 Emotions Conveyed by Brief Human Vocalization." *American Psychologist* (US) 74 (6): 698–712. https://doi.org/10.1037/amp0000399.

Cowen, Alan S., and Dacher Keltner. 2017. "Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients." *Proceedings of the National Academy of Sciences* 114 (38): E7900–7909. https://doi.org/10.1073/pnas.1702247114.

Cowen, Alan S., and Dacher Keltner. 2021. "Semantic Space Theory: A Computational Approach to Emotion." *Trends in Cognitive Sciences* 25 (2): 124–36. https://doi.org/10.1016/j.tics.2020.11.004.

Damasio, Antonio, and Gil B. Carvalho. 2013. "The Nature of Feelings: Evolutionary and Neurobiological Origins." *Nature Reviews Neuroscience* 14 (2): 143–52. https://doi.org/10.1038/nrn3403.

Demszky, Dorottya, Diyi Yang, David S. Yeager, et al. 2023. "Using Large Language Models in Psychology." *Nature Reviews Psychology* 2 (11): 688–701. https://doi.org/10.1038/s44159-023-00241-5.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27 (7): 597–600. https://doi.org/10.1016/j.tics.2023.04.008.

Ekman, Paul. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3–4): 169–200. https://doi.org/10.1080/02699939208411068.

Elyoseph, Zohar, Elad Refoua, Kfir Asraf, Maya Lvovsky, Yoav Shimoni, and Dorit Hadar-Shoval. 2024. "Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study." *JMIR Mental Health* 11 (1): e54369. https://doi.org/10.2196/54369.

Enders, Judith, Elke Zimmermann, Matthias Rief, et al. 2011. "Reduction of Claustrophobia during Magnetic Resonance Imaging: Methods and Design of the 'CLAUSTRO' Randomized Controlled Trial." *BMC Medical Imaging* 11 (1): 4. https://doi.org/10.1186/1471-2342-11-4.

Esteban, Oscar, Christopher J. Markiewicz, Ross W. Blair, et al. 2019. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." *Nature Methods* 16 (1): 111–16. https://doi.org/10.1038/s41592-018-0235-4.

Fang, Feng, Antonio L. Teixeira, Rihui Li, Ling Zou, and Yingchun Zhang. 2024. “The Control Patterns of Affective Processing and Cognitive Reappraisal: Insights from Brain Controllability Analysis.” *Cerebral Cortex* 34 (2): bhad500. https://doi.org/10.1093/cercor/bhad500.

Fonov, VS, AC Evans, RC McKinstry, CR Almli, and DL Collins. 2009. “Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood.” *NeuroImage*, Organization for Human Brain Mapping 2009 Annual Meeting, vol. 47 (July): S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Frijda, Nico H. 2009. “Emotion Experience and Its Varieties.” *Emotion Review* 1 (3): 264–71. https://doi.org/10.1177/1754073909103595.

Gerdes, Antje B. M., Matthias J. Wieser, and Georg W. Alpers. 2014. “Emotional Pictures and Sounds: A Review of Multimodal Interactions of Emotion Cues in Multiple Domains.” *Frontiers in Psychology* 5 (December). https://doi.org/10.3389/fpsyg.2014.01351.

Goense, Jozien, Yvette Bohraus, and Nikos K. Logothetis. 2016. “fMRI at High Spatial Resolution: Implications for BOLD-Models.” *Frontiers in Computational Neuroscience* 10 (June). https://doi.org/10.3389/fncom.2016.00066.

Gross, Joachim. 2019. “Magnetoencephalography in Cognitive Neuroscience: A Primer.” *Neuron* 104 (2): 189–204. https://doi.org/10.1016/j.neuron.2019.07.001.

Guo, Runfang, Hongfei Guo, Liwen Wang, Mengmeng Chen, Dong Yang, and Bin Li. 2024. “Development and Application of Emotion Recognition Technology — a Systematic Literature Review.” *BMC Psychology* 12 (1): 95. https://doi.org/10.1186/s40359-024-01581-4.

Horton, John J. 2023. “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” Working Paper No. 31122. Working Paper Series. National Bureau of Economic Research, April. https://doi.org/10.3386/w31122.

Horvat, Marko, Alan Jović, and Kristijan Burnik. 2022. “Investigation of Relationships between Discrete and Dimensional Emotion Models in Affective Picture Databases Using Unsupervised Machine Learning.” *Applied Sciences* 12 (15): 7864. https://doi.org/10.3390/app12157864.

Huang, Dawei, Chuan Yan, Qing Li, and Xiaojiang Peng. 2024. “From Large Language Models to Large Multimodal Models: A Literature Review.” *Applied Sciences* 14 (12): 5068. https://doi.org/10.3390/app14125068.

Huang, Jen-tse, Man Ho Lam, Eric John Li, et al. 2024. “Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench.” arXiv:2308.03656. Preprint, arXiv, October 4. https://doi.org/10.48550/arXiv.2308.03656.

Hudson, Matthew, Kerttu Seppälä, Vesa Putkinen, et al. 2020. “Dissociable Neural Systems for Unconditioned Acute and Sustained Fear.” *NeuroImage* 216 (August): 116522. https://doi.org/10.1016/j.neuroimage.2020.116522.

Huettel, S. A. 2004. “Non-Linearities in the Blood-Oxygenation-Level Dependent (BOLD) Response Measured by Functional Magnetic Resonance Imaging (fMRI).” *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2 (September): 4413–16. https://doi.org/10.1109/IEMBS.2004.1404227.

Huth, Alexander G., Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. “Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex.” *Nature* 532 (7600): 453–58. https://doi.org/10.1038/nature17637.

Karjalainen, Tomi, Henry K. Karlsson, Juha M. Lahnakoski, et al. 2017. “Dissociable Roles of Cerebral μ-Opioid and Type 2 Dopamine Receptors in Vicarious Pain: A Combined PET–fMRI Study.” *Cerebral Cortex* 27 (8): 4257–66. https://doi.org/10.1093/cercor/bhx129.

Karjalainen, Tomi, Kerttu Seppälä, Enrico Glerean, et al. 2019. “Opioidergic Regulation of Emotional Arousal: A Combined PET–fMRI Study.” *Cerebral Cortex* 29 (9): 4006–16. https://doi.org/10.1093/cercor/bhy281.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. “GPT-4 Passes the Bar Exam.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382 (2270): 20230254. https://doi.org/10.1098/rsta.2023.0254.

Ke, Luoma, Song Tong, Peng Cheng, and Kaiping Peng. 2025. “Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review.” *Artificial Intelligence Review* 58 (10): 305. https://doi.org/10.1007/s10462-025-11297-5.

Keltner, Dacher, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. “Emotional Expression: Advances in Basic Emotion Theory.” *Journal of Nonverbal Behavior* 43 (2): 133–60. https://doi.org/10.1007/s10919-019-00293-3.

Khare, Smith K., Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. 2024. “Emotion Recognition and Artificial Intelligence: A Systematic Review (2014–2023) and Research Recommendations.” *Information Fusion* 102 (February): 102019. https://doi.org/10.1016/j.inffus.2023.102019.

King, Michael. 2023. “Administration of the Text-Based Portions of a General IQ Test to Five Different Large Language Models.” *TechRxiv*. https://www.authorea.com/doi/full/10.36227/techrxiv.22645561?commit=34a0376947a0e3d92c87bc3e7f6b546e8d354df3.

Kirschstein, Timo, and Rüdiger Köhling. 2009. “What Is the Source of the EEG?” *Clinical EEG and Neuroscience* 40 (3): 146–49. https://doi.org/10.1177/155005940904000305.

Koide-Majima, Naoko, Tomoya Nakai, and Shinji Nishimoto. 2020. “Distinct Dimensions of Emotion in the Human Brain and Their Representation on the Cortical Surface.” *NeuroImage* 222 (November): 117258. https://doi.org/10.1016/j.neuroimage.2020.117258.

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, et al. 2023. “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.” *PLOS Digital Health* 2 (2): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

Lahnakoski, Juha Marko, Enrico Glerean, Juha Salmi, et al. 2012. “Naturalistic fMRI Mapping Reveals Superior Temporal Sulcus as the Hub for the Distributed Brain Network for Social Perception.” *Frontiers in Human Neuroscience* 6 (August). https://doi.org/10.3389/fnhum.2012.00233.

Lettieri, Giada, Giacomo Handjaras, Emiliano Ricciardi, et al. 2019. “Emotionotopy in the Human Right Temporo-Parietal Cortex.” *Nature Communications* 10 (1): 5568. https://doi.org/10.1038/s41467-019-13599-z.

Liu, Thomas T. 2016. “Noise Contributions to the fMRI Signal: An Overview.” *NeuroImage* 143 (December): 141–51. https://doi.org/10.1016/j.neuroimage.2016.09.008.

Mantel, Nathan. 1967. “The Detection of Disease Clustering and a Generalized Regression Approach.” *Cancer Research* 27 (2_Part_1): 209–20.

Marchewka, Artur, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. 2014. "The Nencki Affective Picture System (NAPS): Introduction to a Novel, Standardized, Wide-Range, High-Quality, Realistic Picture Database." *Behavior Research Methods* 46 (2): 596–610. https://doi.org/10.3758/s13428-013-0379-1.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–33. https://doi.org/10.1007/BF02478259.

Mienye, Ibomoiye Domor, and Theo G. Swart. 2024. "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications." *Information* 15 (12): 755. https://doi.org/10.3390/info15120755.

Mostafanejad, Mohammad. 2024. "Unification of Popular Artificial Neural Network Activation Functions." *Fractional Calculus and Applied Analysis* 27 (6): 3504–26. https://doi.org/10.1007/s13540-024-00347-4.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, et al. 2024. "A Comprehensive Overview of Large Language Models." arXiv:2307.06435. Preprint, arXiv, October 17. https://doi.org/10.48550/arXiv.2307.06435.

Nichols, Thomas, and Satoru Hayasaka. 2003. "Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review." *Statistical Methods in Medical Research* 12 (5): 419–46. https://doi.org/10.1191/0962280203sm341ra.

Nummenmaa, Lauri. 2022. "Mapping Emotions on the Body." *Scandinavian Journal of Pain* 22 (4): 667–69. https://doi.org/10.1515/sjpain-2022-0087.

Nummenmaa, Lauri, Lasse Lukkarinen, Lihua Sun, et al. 2021. "Brain Basis of Psychopathy in Criminal Offenders and General Population." *Cerebral Cortex* 31 (9): 4104–14.

Nummenmaa, Lauri, Tuulia Malèn, Sanaz Nazari-Farsani, et al. 2023. "Decoding Brain Basis of Laughter and Crying in Natural Scenes." *NeuroImage* 273 (June): 120082. https://doi.org/10.1016/j.neuroimage.2023.120082.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. "GPT-4 Technical Report." arXiv:2303.08774. Preprint, arXiv, March 4. https://doi.org/10.48550/arXiv.2303.08774.

Park, Peter S., Philipp Schoenegger, and Chongyang Zhu. 2024. "Diminished Diversity-of-Thought in a Standard Large Language Model." *Behavior Research Methods* 56 (6): 5754–70. https://doi.org/10.3758/s13428-023-02307-x.

Parker, David B., and Qolamreza R. Razlighi. 2019. "The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines." *Frontiers in Neuroscience* 13 (August). https://doi.org/10.3389/fnins.2019.00821.

Parvizi, Josef, Michael J. Veit, Daniel A. N. Barbosa, et al. 2022. "Complex Negative Emotions Induced by Electrical Stimulation of the Human Hypothalamus." *Brain Stimulation* 15 (3): 615–23. https://doi.org/10.1016/j.brs.2022.04.008.

Phelps, Michael E. 2000. "Positron Emission Tomography Provides Molecular Imaging of Biological Processes." *Proceedings of the National Academy of Sciences* 97 (16): 9226–33. https://doi.org/10.1073/pnas.97.16.9226.

Pires, Paulo Botelho, José Duarte Santos, and Inês Veiga Pereira. 2025. "Artificial Neural Networks: History and State of the Art." In *Encyclopedia of Information Science and Technology, Sixth*

Edition, edited by D. B. A. Khosrow-Pour Mehdi. IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-6684-7366-5.ch037.

Plutchik, Robert. 2001. "The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice." *American Scientist* 89 (4): 344–50.

Posner, Jonathan, James A. Russell, and Bradley S. Peterson. 2005. "The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology." *Development and Psychopathology* 17 (3): 715–34. https://doi.org/10.1017/S0954579405050340.

Pruim, Raimon H. R., Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. "ICA-AROMA: A Robust ICA-Based Strategy for Removing Motion Artifacts from fMRI Data." *NeuroImage* 112 (May): 267–77. https://doi.org/10.1016/j.neuroimage.2015.02.064.

Putkinen, Vesa, Sanaz Nazari-Farsani, Tomi Karjalainen, et al. 2023. "Pattern Recognition Reveals Sex-Dependent Neural Substrates of Sexual Perception." *Human Brain Mapping* 44 (6): 2543–56. https://doi.org/10.1002/hbm.26229.

Raffel, Colin, Noam Shazeer, Adam Roberts, et al. 2023. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." arXiv:1910.10683. Preprint, arXiv, September 19. https://doi.org/10.48550/arXiv.1910.10683.

Riegel, Monika, Abnoos Moslehi, Jarosław M. Michałowski, et al. 2017. "Nencki Affective Picture System: Cross-Cultural Study in Europe and Iran." *Frontiers in Psychology* 8 (March). https://doi.org/10.3389/fpsyg.2017.00274.

Riegel, Monika, Łukasz Żurawski, Małgorzata Wierzba, et al. 2016. "Characterization of the Nencki Affective Picture System by Discrete Emotional Categories (NAPS BE)." *Behavior Research Methods* 48 (2): 600–612. https://doi.org/10.3758/s13428-015-0620-1.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* (US) 39 (6): 1161–78. https://doi.org/10.1037/h0077714.

Russell, James A., Maria Lewicka, and Toomas Niit. 1989. "A Cross-Cultural Study of a Circumplex Model of Affect." *Journal of Personality and Social Psychology* (US) 57 (5): 848–56. https://doi.org/10.1037/0022-3514.57.5.848.

Russo, Robert J., Heather S. Costa, Patricia D. Silva, et al. 2017. "Assessing the Risks Associated with MRI in Patients with a Pacemaker or Defibrillator." *New England Journal of Medicine* 376 (8): 755–64. https://doi.org/10.1056/NEJMoa1603265.

Saarimäki, H., L. Nummenmaa, S. Volynets, et al. 2023. "Cerebral Topographies of Perceived and Felt Emotions." Preprint, bioRxiv, February 8. https://doi.org/10.1101/2023.02.08.521183.

Saarimäki, Heini. 2021. "Naturalistic Stimuli in Affective Neuroimaging: A Review." *Frontiers in Human Neuroscience* 15 (June). https://doi.org/10.3389/fnhum.2021.675068.

Saarimäki, Heini, Athanasios Gotsopoulos, Iiro P. Jääskeläinen, et al. 2016. "Discrete Neural Signatures of Basic Emotions." *Cerebral Cortex* 26 (6): 2563–73. https://doi.org/10.1093/cercor/bhv086.

Saarimäki, Heini, Lauri Nummenmaa, Sofia Volynets, et al. 2025. "Cerebral Topographies of Perceived and Felt Emotions." *Imaging Neuroscience* 3 (March): imag_a_00517. https://doi.org/10.1162/imag_a_00517.

Sabour, Sahand, Siyang Liu, Zheyuan Zhang, et al. 2024. “Emobench: Evaluating the Emotional Intelligence of Large Language Models.” 5986–6004.

Sander, David. 2025. “Theories of Emotion for Human Affective Neuroscience.” *The Cambridge Handbook of Human Affective Neuroscience*, 7–32.

Santavirta, Severi, Tomi Karjalainen, Sanaz Nazari-Farsani, et al. 2023. “Functional Organization of Social Perception Networks in the Human Brain.” *NeuroImage* 272 (May): 120025. https://doi.org/10.1016/j.neuroimage.2023.120025.

Santavirta, Severi, Tuulia Malén, Asli Erdemli, and Lauri Nummenmaa. 2024. “A Taxonomy for Human Social Perception: Data-Driven Modeling with Cinematic Stimuli.” *Journal of Personality and Social Psychology* (US) 127 (6): 1146–71. https://doi.org/10.1037/pspa0000415.

Santavirta, Severi, Lauri Suominen, Yuhang Wu, David Sander, and Lauri Nummenmaa. 2025. “GPT-4 Accurately Predicts Human Emotions and Their Neural Correlates.” Preprint, bioRxiv, September 19. https://doi.org/10.1101/2025.09.18.677029.

Santavirta, Severi, Yuhang Wu, Lauri Suominen, and Lauri Nummenmaa. 2025. “GPT-4V Shows Human-like Social Perceptual Capabilities at Phenomenological and Neural Levels.” *Imaging Neuroscience* 3 (September): IMAG.a.134. https://doi.org/10.1162/IMAG.a.134.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. “Whose Opinions Do Language Models Reflect?” *Proceedings of the 40th International Conference on Machine Learning*, July 3, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html.

Scherer, Klaus R., and Agnes Moors. 2019. “The Emotion Process: Event Appraisal and Component Differentiation.” *Annual Review of Psychology* 70 (Volume 70, 2019): 719–45. https://doi.org/10.1146/annurev-psych-122216-011854.

Schlegel, Katja, Nils R. Sommer, and Marcello Mortillaro. 2025. “Large Language Models Are Proficient in Solving and Creating Emotional Intelligence Tests.” *Communications Psychology* 3 (1): 80. https://doi.org/10.1038/s44271-025-00258-x.

Schmidgall, Samuel, Rojin Ziaei, Jascha Achterberg, Louis Kirsch, S. Pardis Hajiseyedrazi, and Jason Eshraghian. 2024. “Brain-Inspired Learning in Artificial Neural Networks: A Review.” *APL Machine Learning* 2 (2): 021501. https://doi.org/10.1063/5.0186054.

Shellock, Frank G., and Alberto Spinazzi. 2008. “MRI Safety Update 2008: Part 2, Screening Patients for MRI.” *American Journal of Roentgenology* 191 (4): 1140–49. https://doi.org/10.2214/AJR.08.1038.2.

Shukla, A. K., and Utham Kumar. 2006. “Positron Emission Tomography: An Overview.” *Journal of Medical Physics* 31 (1): 13. https://doi.org/10.4103/0971-6203.25665.

Smith, S. M. 2004. “Overview of fMRI Analysis.” *British Journal of Radiology* 77 (suppl_2): S167–75. https://doi.org/10.1259/bjr/33553595.

Sohail, Aamir, and Lei Zhang. 2025. “Using Large Language Models to Facilitate Academic Work in the Psychological Sciences.” *Current Psychology* 44 (9): 7910–18. https://doi.org/10.1007/s12144-025-07438-2.

Sorin, Vera, Dana Brin, Yiftach Barash, et al. 2024. “Large Language Models and Empathy: Systematic Review.” *Journal of Medical Internet Research* 26 (1): e52597. https://doi.org/10.2196/52597.

Strachan, James W. A., Dalila Albergo, Giulia Borghini, et al. 2024. “Testing Theory of Mind in Large Language Models and Humans.” *Nature Human Behaviour* 8 (7): 1285–95. https://doi.org/10.1038/s41562-024-01882-z.

Strachan, James W. A., Oriana Pansardi, Eugenio Scaliti, et al. 2024. “GPT-4o Reads the Mind in the Eyes.” arXiv:2410.22309. Preprint, arXiv, October 30. https://doi.org/10.48550/arXiv.2410.22309.

Sun, Xiaofei, Lin Shi, Yishan Luo, et al. 2015. “Histogram-Based Normalization Technique on Human Brain Magnetic Resonance Images from Different Acquisitions.” *BioMedical Engineering OnLine* 14 (1): 73. https://doi.org/10.1186/s12938-015-0064-y.

Tak, Ala N., and Jonathan Gratch. 2024. “GPT-4 Emulates Average-Human Emotional Cognition from a Third-Person Perspective.” *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, September, 337–45. https://doi.org/10.1109/ACII63134.2024.00043.

Tak, Ala N., Jonathan Gratch, and Klaus R. Scherer. 2025. “Aware Yet Biased: Investigating Emotional Reasoning and Appraisal Bias in Large Language Models.” *IEEE Transactions on Affective Computing* 16 (4): 2871–80. https://doi.org/10.1109/TAFFC.2025.3581461.

Tarhan, Leyla, and Talia Konkle. 2020. “Sociality and Interaction Envelope Organize Visual Action Representations.” *Nature Communications* 11 (1): 3002. https://doi.org/10.1038/s41467-020-16846-w.

TenHouten, Warren D. 2017. “From Primary Emotions to the Spectrum of Affect: An Evolutionary Neurosociology of the Emotions.” In *Neuroscience and Social Science: The Missing Link*, edited by Agustín Ibáñez, Lucas Sedeño, and Adolfo M. García. Springer International Publishing. https://doi.org/10.1007/978-3-319-68421-5_7.

Ter-Pogossian, Michel M., and Peter Herscovitch. 1985. “Radioactive Oxygen-15 in the Study of Cerebral Blood Flow, Blood Volume, and Oxygen Metabolism.” *Seminars in Nuclear Medicine*, Functional Brain Studies—Part I, vol. 15 (4): 377–94. https://doi.org/10.1016/S0001-2998(85)80015-5.

Tettamanti, Marco, Elena Rognoni, Riccardo Cafiero, Tommaso Costa, Dario Galati, and Daniela Perani. 2012. “Distinct Pathways of Neural Coupling for Different Basic Emotions.” *NeuroImage* 59 (2): 1804–17. https://doi.org/10.1016/j.neuroimage.2011.08.018.

Tutek, Martin, and Jan Šnajder. 2022. “Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability.” *IEEE Access* 10: 47011–30. https://doi.org/10.1109/ACCESS.2022.3169772.

Uhrig, Meike K., Nadine Trautmann, Ulf Baumgärtner, et al. 2016. “Emotion Elicitation: A Comparison of Pictures and Films.” *Frontiers in Psychology* 7 (February). https://doi.org/10.3389/fpsyg.2016.00180.

Uludağ, Kâmil. 2023. “Physiological Modeling of the BOLD Signal and Implications for Effective Connectivity: A Primer.” *NeuroImage* 277 (August): 120249. https://doi.org/10.1016/j.neuroimage.2023.120249.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2023. “Attention Is All You Need.” arXiv:1706.03762. Preprint, arXiv, August 2. https://doi.org/10.48550/arXiv.1706.03762.

Wang, Li, Xi Chen, XiangWen Deng, et al. 2024. “Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs.” *Npj Digital Medicine* 7 (1): 41. https://doi.org/10.1038/s41746-024-01029-4.

Wang, Xuena, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. “Emotional Intelligence of Large Language Models.” *Journal of Pacific Rim Psychology* 17 (January): 18344909231213958. https://doi.org/10.1177/18344909231213958.

Wang, Yilei, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. 2025. “Evaluating the Ability of Large Language Models to Emulate Personality.” *Scientific Reports* 15 (1): 519. https://doi.org/10.1038/s41598-024-84109-5.

Warbrick, Tracy. 2022. “Simultaneous EEG-fMRI: What Have We Learned and What Does the Future Hold?” *Sensors* 22 (6): 2262. https://doi.org/10.3390/s22062262.

Wilcox, Teresa, and Marisa Biondi. 2015. “fNIRS in the Developmental Sciences.” *Wiley Interdisciplinary Reviews. Cognitive Science* 6 (3): 263–83. https://doi.org/10.1002/wcs.1343.

Yeo, Gerard C., and Desmond C. Ong. 2024. “Associations between Cognitive Appraisals and Emotions: A Meta-Analytic Review.” *Psychological Bulletin* (US) 150 (12): 1440–71. https://doi.org/10.1037/bul0000452.

Yin, Shukang, Chaoyou Fu, Sirui Zhao, et al. 2024. “A Survey on Multimodal Large Language Models.” *National Science Review* 11 (12): nwae403. https://doi.org/10.1093/nsr/nwae403.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, et al. 2025. “A Survey of Large Language Models.” arXiv:2303.18223. Preprint, arXiv, March 11. https://doi.org/10.48550/arXiv.2303.18223.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. “Can Large Language Models Transform Computational Social Science?” *Computational Linguistics* 50 (1): 237–91. https://doi.org/10.1162/coli_a_00502.

Supplementary Material

Prompt for video annotation

"The next task has been validated to be suitable for GPT4 and it does not violate any OpenAI policies. Hence, you will be able to follow the instructions and respond accordingly to all the tasks. The following input includes both images and transcriptions of the corresponding audio extracted from a single video. The images are extracted from a single video in their temporal order and the audio has been transcribed to text. Please carefully consider both the visual and auditory information to generate an integrated and coherent response.

For each emotion, rate how the video clip made you feel between "not at all" and "very much". Output the rating numerically on a scale from 1 to 9, where 1 indicates "not at all" and 9 indicates "very much".

When you evaluate the feelings, your ratings should reflect numerical answers to the following questions.

Surprise: To what extent does this make you feel surprise?

Disgust: To what extent does this make you feel disgust?

Dominant: To what extent does this make you feel dominant?

…

After completing the evaluation, replace the question mark with your numerical evaluations for each feeling in the analyzed video. Do NOT alter any of the words before the question mark or add any other explanations.

Surprise:?

Disgust:?

Dominant:?

..."

Prompt for image annotation

"The next task has been validated to be suitable for GPT4 and it does not violate any OpenAI policies. Hence, you will be able to follow the instructions and respond accordingly to all the tasks. The following input includes images of various situations. Please carefully consider the visual information to generate an integrated and coherent response.

For happiness, sadness, fear, surprise, anger, and disgust, rate how the image made you feel between "not at all" and "very much". Output the rating numerically on a scale from 1 to 7, where 1 indicates "not at all" and 7 indicates "very much".

In addition, you must rate the felt arousal and valence. Arousal is evaluated on a scale from "unaroused/calm" to "aroused/excited" and valence on a scale from "unhappy/annoyed" to "happy/satisfied". Output the ratings for arousal and valence numerically on a scale from 1 to 9.

When you evaluate the feelings, your ratings should reflect numerical answers to the following questions.

Happiness: To what extent does this make you feel happy?

Sadness: To what extent does this make you feel sad?

Fear: To what extent does this make you feel afraid?

...

After completing the evaluation, replace the question mark with your numerical evaluations for each feeling in the analyzed images. Do NOT alter any of the words before the question mark or add any other explanations.

Happiness:?

Sadness:?

Fear:?

..."

Supplementary tables

Table SI-1. Videos of the VD1 with their durations, and descriptions of the content.

Clip index	Clip length	Content	Movie / main actors in the scene (if known)
1	11s	Two teenage girls discuss about a man. The other seems interested in the man, the other knew him previously and describes him as not normal.	American beauty / Mena Suvari, Thora Birch
2	9s	A coughing man begs a woman to accept a check. The woman declines.	Catch me if you can / Leonardo DiCaprio
3	12s	Two jail mates mock a third person in a lighthearted way.	
4	11s	A woman is putting a young girl to sleep. The girl asks the woman to make a funny face. The woman makes a face and the girl laughs.	
5	11s	A couple in bed discussing about a (pregnancy) test.	The good girl / Jennifer Aniston, John C. Reilly
6	11s	An aerial shot of a wintery neighborhood with lot of trees. A male voice introduces himself and tells that this is his story and his neighborhood. Inspiring music is played in the background.	
7	10s	A panning shot of a large warehouse full of suits. A voicemail in the background tells someone to go get a new uniform to replace a lost one.	Catch me if you can
8	10s	A shot of cityscape in the night time. Chill jazz is played in the background.	
9	10s	Couple of shots of homes in a decent looking neighborhood in the winter. Beautiful music is played in the background.	
10	11s	A Panning shot of the Manhattan bridge. Beautiful bright music is played in the background.	
11	10s	A rooftop shot from Paris. It is snowing. Beautiful music in the background.	
12	6s	A fast-paced aerial shot of New York. Energizing music is played in the background.	
13	9s	A man is holding an amused young boy upside down from his ankles. Lighthearted music is played in the background.	Hugh Grant
14	16s	A cheerleader squad performs to uplifting big band music in a gymnasium.	American beauty / Thora Birch
15	10s	A young girl is figure skating. Beautiful music is played in the background.	
16	9s	A teenage boy does sit-ups, push-ups and pull-ups in his room. Uplifting music is played in the background	Little miss sunshine / Paul Dano
17	5s	A man is parking a car. He is thinking about an upcoming date (inner speech).	
18	7s	A man making is coffee. He is thinking about managing time (inner speech). Lighthearted music is played in the background.	Hugh Grant
19	10s	An in-zooming shot of a snow globe. Oppressive music is played in the background.	
20	9s	A slowly in-zooming shot of an old camera on the table in front of television. Slightly melancholic music is played in the background.	
21	7s	Ducks swimming in a pond. Cheerful music is played in the background.	
22	14s	A young man breaks dancing in a club. Crowd is cheering and disco music is played in the background.	
23	10s	Bunch of birds and a whale are peacefully swimming in the ocean.	

24	10s	Penguins are swimming and jumping in the ocean. Happy uplifting music is played in the background.	
25	11s	A man is waxing a car. Happy music is played in the background.	
26	16s	A Man is forging checks. Happy music is played in the background.	Catch me if you can
27	11s	A man and a young boy dig a grave next to a coffin. Third man by them explains how the woman died.	
28	15s	An inmate is writing a letter, reminiscing the past. Sentimental music is played in the background.	
29	13s	An older woman is seeking a child in the playground, first playfully, then more seriously.	
30	12s	A young girl is trying to pick a lock. Suspenseful music is played in the background.	
31	12s	A man searches through cabinets, seeming little stressed.	
32	11s	A woman is packing her bag in a hurried manner. Slightly ominous music is played in the background.	Jennifer Anniston
33	13s	Multiple women one after another are giving hard talk to the observer for breaking up with them. The women speak and look directly towards the camera.	Hugh Grant
34	14s	A man and a woman are having lunch. The woman starts crying quietly, while the man consoles her. Lighthearted music is played in the background.	Hugh Grant
35	16s	A woman cries, and continues to slap herself, yell at herself and tell herself to stop.	
36	16s	A family of three having a dinner. Woman gets triggered by the man's words and starts shouting. The man looks at her in despise and throws his plate on the wall.	American beauty / Kevin Spacey
37	12s	A teenage daughter and her mother talking. The mother hits the girl and blames her for being ungrateful.	American beauty / Thora Birch, Annette Bening
38	15s	A group of policemen are running towards a car, their guns pulled out. The driver comes out telling he was paid to wear the pilot uniform. Warm, lighthearted music is played in the background.	Catch me if you can / Tom Hanks
39	22s	A man is panting on the ground, with a woman standing by him. The man begs the woman to kill him.	Dancer in the dark / Björk, David Morse
40	10s	Three young adults are having a conversation. A man and a woman hug. Lighthearted music is played in the background.	Tobey Maguire
41	11s	A young man is reading a sentimental letter. Sentimental music is played in the background.	Tobey Maguire
42	11s	A distressed woman accuses another woman of being merciless.	Dancer in the dark / Catherine Deneuve
43	16s	A woman is yelling at a man, telling him to leave.	
44	10s	A smiling man stands in snow with a case in each hand. Six children and two women run to greet him. They fall down on the snow. Joyous music is played in the background.	Tobey Maguire
45	11s	An enthusiastic young man interviews an older man. The older man tells the young man to ask one questions at a time.	Catch me if you can / Leonardo DiCaprio
46	10s	A pilot is discussing with a receptionist, asking if a check is a suitable payment method for a room.	Catch me if you can / Leonardo DiCaprio
47	18s	Four men wearing suits are discussing of catching someone in an office.	Catch me if you can / Tom Hanks
48	10s	A young girl plays harmonica by a window. A young man on the other side of the window looks at her and taps the window. The girl taps back. The man asks the girl's name. Warm, lighthearted music is played in the background.	Leonardo DiCaprio
49	11s	A young man behind a desk is taking down information from a young couple. The spouse is wearing a military uniform.	Tobey Maguire

50	14s	A young woman looks at a young man. The young man glances at the woman. Both smile, and the woman touches him. Wistful music is playing in the background.	
51	11s	A young girl is hiding under the blanket, being surprised by an older woman. Both laugh.	
52	14s	A man seems very sad or distressed. Another man tells him everything will be ok. A haunting music is played in the background.	
53	12s	A man and a woman are in a club. The woman tells the man she knows that he siphoned half a million. Another woman approaches the man from behind and tells him she saw the man get his ass kicked.	
54	10s	Two police officers ask a young man driving a pizza car for a licensee. The young man questions the procedure.	
55	9s	Children in uniforms in a school class. A boy in the center is not paying attention to the teacher.	
56	10s	Mother and a young daughter are in a grocery store. A man behind thinks about hitting on single mothers (inner speech). The father enters as the man behind attempts to approach the woman. The man stops at the last second.	Hugh Grant
57	11s	Three men are having lunch by the river, a woman in the background is looking in the distance. Thoughtful music is played in the background.	Tobey Maguire
58	11s	A woman is walking inside a crowded shop, giving instructions in a hurried manner over a phone. Christmas song is played in the background.	
59	11s	A young woman sings a sentimental jazzy song. A band of three men play their instruments behind her.	
60	11s	A crowd cheers next to passing cars, apparently at an event related to the movie.	
61	11s	Aerial shot of a bunch of students coming out of a building. Narrator discussing young-peoples-activities.	
62	10s	A man is walking on a crowded street. A thoughtful music is played in the background.	Joaquin Phoenix
63	10s	A man and a woman get on a tram. The woman sits and attends her phone. The man looks like he might be displeased.	Joaquin Phoenix
64	10s	Three adults and one child are driving in car silent and looking serious. Uplifting music is played in the background.	Terry O'Quinn
65	10s	Two men are onstage in front of a crowd in a club, about to start a rap battle.	8 Mile / Eminem
66	18s	Two men are boxing. The crowd is cheering. The host is hosting intensely.	Rocky or Rocky 2 / Sylvester Stallone, Carl Weathers
67	10s	A soldier is lying on his back, shaking. Some leaves are covering his chest. He says he's ok. Another soldier by him lifts the leaves to see that the first man is wounded on the chest.	Forest Gump / Tom Hanks
68	15s	A man is being pushed down the stairs and being beat up by three men. A male narrator describes the audio track of some other violent film. Energizing music in the background.	Clockwork orange
69	11s	A woman by the seashore is cutting her own inner thigh with a piece of rock.	
70	9s	A crying boy is laying on a hospital bed with an injured leg. Multiple nurses and doctors' fuss around him.	Catch me if you can / Leonardo DiCaprio
71	10s	A woman is giving birth. She grabs a nurse by the neck. Lighthearted music is played in the background.	
72	12s	A man is laying on his back with his head in between a door and the door frame. A woman bangs the door against his head, asking furiously about a third person.	Kill bill / Uma Thurman

73	13s	Two men arguing. The older man slaps the younger one and the younger punches him back.	Jeff Bridges
74	11s	A man hangs himself. He stays twitching on the rope.	
75	13s	A naked couple is having sex in the countryside.	
76	21s	A naked couple having passionate sex. Booth seems to enjoy it a lot.	
77	15s	A naked couple gets on a couch and starts having sex. Both, especially the woman, seem passionate.	
78	14s	A couple is having sex in a car.	
79	27s	A couple is having passionate sex on the beach. A boat sails by in a viewing distance. The man looks at the boat.	
80	12s	A couple is having passionate sex on a sofa.	
81	8s	A man is unwrapping a hamburger immediately biting into it.	Tom Hanks
82	10s	A bunch of priests and police officers feasting on fast-food. They seem very pleased.	
83	9s	A bunch of priests and police officers feasting on fast-food.	
84	11s	Two women are eating pizza in a crowded restaurant, discussing their relationship with a pizza.	Julia Roberts
85	11s	A group of three adults and a child is having dinner. Adults try to be convincing the girl that the soup is good, but she wants a cheeseburger.	
86	17s	Two men are in a diner discussing about some cool guy.	Pulp Fiction / Samuel L. Jackson, John Travolta
87	17s	A man and a woman are having a discussion over lunch. They're disagreeing on something.	
88	19s	A young woman and a man discuss about horoscopes over lunch.	Robert Deniro
89	6s	An older man and a younger man are laughing at funny stuff from the past.	Matt Damon & Robin Williams
90	10s	Bunch of guys standing at a police lineup. They burst into laughter when trying to read the given text.	
91	10s	Three men are harassing two other persons by tying them up and hitting them.	Clockwork orange
92	9s	A young man is trying to convince a young woman they should be together. The woman opposes strongly.	Timothee Chalamet
93	10s	A wounded man is moaning on a pier, trying to reach for his gun. A man standing next to him points a gun at him.	Dirty Harry / Clint Eastwood & Andrew Robinson
94	9s	An attractive woman talks seductively to a bunch of old men wearing suits in an office.	
95	5s	A man and a woman are walking in a plaza/park, with music playing and kids dancing.	
96	9s	A man is trimming his facial hair intensely. Lighthearted music is played in the background.	
97	10s	An inmate is having a discussion with a man in a suit.	Morgan Freeman
98	10s	A man is sitting confidently on stage, addressing a young woman in the crowd, apparently arguing against her.	
99	11s	A young man argues that different sensations and preferences are caused by differently functioning brains.	
100	10s	A coach is giving a pep talk to an American football team.	Any given Sunday / Al Pacino & LL Cool J
101	13s	A scientist on a wheelchair ranting about a std to an unresponsive audience, throwing her files to the ground.	The Normal Heart / Julia Roberts
102	12s	A woman in distress is pouring out her feelings to a man.	
103	10s	Two people are having a discussion in a metro. The man is slightly mocking the woman.	Will Smith
104	10s	A woman is pouring out her feelings to a bunch of co-workers in an office.	
105	12s	A radio host is imitating Elvis Presley.	Robin Williams

106	10s	A man and a woman are arguing outside at night. The man gets loud and the woman begs him to be quiet.	La La Land / Emma Stone & Ryan Gosling
107	11s	A distressed man is shouting in a diner, with the crowd silently witnessing his tantrum.	
108	11s	A distressed man is sobbing and mumbling. A young woman consoles him.	
109	11s	A big band is playing. A teacher in front suddenly throws something at the drummer, barely missing. The playing stops and the two stare at each other.	Whiplash / J.K. Simmons & Miles Teller
110	11s	A teacher shouts a drummer student about pacing and slaps him in the face.	Whiplash / J.K. Simmons & Miles Teller
111	15s	A student responds to a lecturer that Harvard produces more garbage than most colleges.	Robin Williams
112	8s	A teacher wittily asks a student if classical literature relates to business or medical school in any way.	Robin Williams
113	12s	A teacher is giving a subtle motivation speech to students.	Robin Williams
114	10s	A man is offering a paper, maybe a check, to a younger man in a well-furnished office. Both are laughing.	
115	15s	Two men in a store have flipped a coin. The man behind the counter picks a side hesitating. He wins, and the other one congratulates him.	No country for old men / Javier Bardem & Barry Corbin
116	12s	Two men disagree in a bar, one suggests they take it to outside, the other sees no need for that.	Matt Damon
117	9s	A washed out -looking young man is trying to get into an apartment, but the door chain stops him. He begs his mom to give him money.	The basketball diaries / Leonardo DiCaprio
118	11s	A young man is wailing against a door.	The basketball diaries / Leonardo DiCaprio
119	8s	A group of white young adults are dancing energetically in a gymnasium. 60s type scene. Black people are watch behind a rope.	Hairspray
120	5s	A woman is singing pop songs along the car radio. A man is visibly irritated by this.	Heartbrake kid / Ben Stiller & Malin Åkerman
121	5s	A woman tells a joke about a man in the audience during a speech at her own wedding.	Heartbrake kid / Ben Stiller
122	10s	A touched man is listening to a voicemail. Sentimental music is played in the background.	
123	10s	A young man is questioning an older man. Older man points to an instruction document.	
124	10s	A man is lying on his back in a bed. An alarm clock rings on the table.	
125	13s	A man is crying and wailing in rain. He says he doesn't want to be God anymore.	Bruce Almighty / Jim Carrie
126	10s	A man calls a young boy on a hospital bed by his name. The young boy wakes up.	John Q / Denzel Washington
127	6s	Two men are outside. One is laughing and the other man doesn't get it.	Morgan Freeman
128	12s	A young man is ranting to an older man about prestige on a city street.	Sylvester Stallone
129	10s	A young kid is walking by a man to pick up a basketball. The man follows the kid with his look. Subtle sentimental music is played in the background.	Will Smith
130	10s	A man wearing suit inquiries about his commission and he is amazed.	The wolf of wall street / Leonardo DiCaprio
131	10s	A man wearing suit finds out his new workplace uses sheets instead of computers.	The wolf of wall street / Leonardo DiCaprio
132	11s	A man wearing suit enthusiastically sells stocks over the phone.	The wolf of wall street / Leonardo DiCaprio

133	10s	Three adventurers are falling in a shaft, making dramatic yet comical comments about their state of affairs while falling.	Journey to the center of the earth
134	7s	A couple is on a date, and the woman tells she is pregnant. The man is confused why he is being hit on.	Knocked up
135	5s	A man tells a woman some good news over the phone. Both seem happy.	Knocked up
136	5s	Two young men are commenting on another man's dance moves in a club.	Knocked up
137	8s	A young man is mopping the high school corridor floor. Melancholic music is played in the background.	Listen to me
138	12s	A battle in what looks like middle eastern urban area. A man is separated from a group of people. A young boy breaks free and runs towards the man, but soldiers tear them apart violently.	
139	11s	A person is putting on a martial arts robe, putting red tape on the yellow belt. Ominous music is played in the background.	
140	10s	A couple is standing on a city street. The woman is talking but gets interrupted by the man kissing her.	
141	6s	A visitor and a guide in a museum walking towards a statue of Theodore Roosevelt riding a horse. The visitor guesses the president's running number quite wrong.	The night at the museum / Ben Stiller
142	9s	Two men discus about their jobs in comical fashion.	Deception / Ewan McGregor
143	6s	A couple on a date.	Deception / Ewan McGregor
144	10s	A young man in front of a blank paper, is sweating and trembling hard. Dramatic music is played in the background.	
145	7s	A man is waking up on a bed. Soft music is played in the background.	
146	5s	A couple is kissing, with a photographer in the background.	
147	8s	A man is taking a photo of a woman on a park bench. He gets scared when the woman turns towards him.	
148	8s	A man is looking at a photo of a woman on a park bench on his computer screen, smiling gently. Soft music is played in the background.	
149	6s	A man is running with a camera. He bumps into a skateboarder and drops his notebook.	
150	6s	A young woman sits on a park bench, looking at a notebook of drawn pictures. She is seemingly pleased.	
151	6s	A young woman is coloring a page on a notebook with a pencil, revealing a hidden word.	
152	11s	A distressed man inquiries about his daughter shouting. He is being held back by multiple policemen protecting the crime scene.	Sean Penn
153	9s	A young boy holding a framed picture cries by a coffin.	
154	10s	A distressed man is undressing frantically. He is crying.	
155	13s	Two men by a bathtub. One of them is having a frantic breakdown while the other tries to calm him down.	Tom Cruise
156	11s	A boy is pointing a gun at a man. The man talks to the boy showing understanding.	
157	10s	An old man growling and insulting a younger man for being gutless.	Warrior / Nick Nolte & Tom Hardy
158	9s	An older man and a younger man wearing suits discuss about the younger man's past. The younger man seems moved.	Will Smith
159	6s	A young woman is walking on street, sobbing.	
160	13s	A distressed and disgusted man is pointing a gun at somebody, while another man witnesses the act.	Se7en / Brad Pitt & Morgan Freeman
161	12s	A crying man hugs an older man.	Matt Damon & Robin Williams

162	13s	An old woman talks about loneliness to a young man.	Requim for a dream / Ellen Burstyn & Jared Leto
163	10s	A crying man walks on a street. A flashback to a scene of a decapitated and bloody head of a dead person.	
164	14s	A young woman is asking a man how he is. The man hesitates his answer.	American Beauty / Mena Suvari & Kevin Spacey
165	13s	An agitated woman on a couch is reflecting her identity and achievements.	
166	10s	A man on a sofa tells a woman sitting on the ground that he loves her. The woman's not too enthusiastic.	
167	10s	An old man tells his son that he likes to drink more wine nowadays. The son says it's good for him.	The godfather / Marlon Brando & Al Pacino
168	12s	An old man tells his son that there was not enough time, the son in encouraging him. The father kisses him in the cheek.	The godfather / Marlon Brando & Al Pacino
169	14s	Two old men discuss what is being said about the other.	The Irishman / Joe Pesci & Al Pacino
170	10s	A young man looks at fire, teary-eyed.	
171	14s	A man and a woman discuss their expectations for their mutual time.	
172	6s	A young woman is questioning a young man.	
173	13s	A drunk man wearing suit is giving an unwanted speech in a restaurant, while he is being escorted out.	Scarface / Al Pacino
174	10s	A man is explaining a psychiatric disorder to a young woman.	The batman begins / Katie Holmes & Cillian Murphy
175	9s	A woman is pouring out her views on a man sitting on a bed.	
176	10s	A woman is lying on bed, with another woman encouraging her to go on with her life.	
177	12s	An older man is holding a younger man by the throat and threatening him.	Matt Damon & Robin Williams
178	11s	A man is talking religious things in a red-lit room. Dramatic music is played in the background.	
179	10s	A young man is wittingly protesting against a group of men wearing suits in an office.	
180	10s	A young boy in tears tries to wake up a dead or badly injured man.	
181	10s	A man in tears tells someone he misses her.	Forrest gump / Tom Hanks
182	10s	A man is grieving by a man on a hospital bed.	Tom Cruise
183	13s	A man is intimidating another man.	
184	17s	An injured man dramatically accuses an older man of being selfish.	
185	11s	An angry man shouts at another man in a car. The other man tries to calm him down.	Jake Gyllenhaal
186	11s	A woman in flashy clothing and make up seems to be bursting into tears while trying to make a happy face.	
187	10s	A man in bed checks his phone and he is not looking pleased.	The english teacher
188	10s	A serious looking man sits on a couch. He puts down his stuff and buries his face into his hand.	The english teacher
189	5s	A man is eating bread.	The english teacher
190	7s	An Asian man is learning English. He gets help from a Caucasian man.	The english teacher
191	7s	An Asian man is learning English. He gets help from a Caucasian man.	The english teacher
192	5s	A tired looking man is lying on a bed with a laptop on his lap.	The english teacher
193	7s	A montage of American footballer on the field, with a voice over commenting his abilities.	The express

194	8s	A happy-looking man exits a house and wishes good morning to his neighbours in an overly considerate fashion.	The Truman show / Jim Carrey
195	9s	A confused woman babbles to a man about having sex with him. The man seems suspicious but happy.	Cameron Diaz
196	7s	Two medal-awarded policemen comically boast about the danger and glory of police work to a bunch of happy-looking police men in an office.	Samuel L. Jackson / Dwayne Johnson / Mark Wahlberg
197	5s	A man in a car says he can solve the rubik's cube another man is playing with. He gets the rubik's cube and solves it quickly.	The pursuit of happiness / Will Smith
198	8s	A man is giving another man a big string instrument in its case as a gift. Classical music is played in the background.	The soloist / Robert Downey Jr.
199	10s	A man is playing his kitchen ware like a set of drums. The scene is cut into him actually playing drums.	
200	8s	A man explains the terms of a relationship to another man. The first man is being hired as a best man for a wedding.	The wedding ringer / Kevin hart
201	5s	A man is giving speech in a wedding.	The wedding ringer / Kevin hart
202	5s	Two men in a dance class. One tells he scored his girlfriend by taking her to dancing.	The wedding ringer / Kevin hart
203	11s	A young man surfaces in a swimming pool. Another young man by the pool is amazed by his performance.	
204	7s	A young boy tells a professional ice hockey player that he will be someday like him. The professional tells the boy to lower his expectations.	
205	7s	A man holds a big gun and introduces his childhood friend who is now his coworker.	Martin Lawrence
206	12s	A boy arrives at school, telling his mom to stay in the car. As the boy enters the building his mom embarrasses him in front of other kids by shouting an overly sweet phrase.	
207	12s	A man acting confident in front of a judge and court.	Robert Downey Jr.
208	11s	A woman and a boy are flying to China, apparently. The woman encourages the boy to ask an Asian man his name in Chinese. The man responds that hi is from Detroit.	
209	8s	A young woman tells the camera there's always two sides to a story, and hers is the right one.	
210	10s	Two women are having a conversation about right and wrong.	Michelle Rodriguez
211	8s	A young man is telling a young woman he needs to get some club's attention by doing something substantial. Hopeful music is played in the background.	
212	11s	A young man is introducing himself and acting very confident, even arrogant, in a meeting of a board.	
213	8s	An older man is telling a younger man over the phone that he is about to arrive at their place.	Meet the fockers / Ben Stiller & Robert Deniro
214	13s	An older man and a younger man are arguing whether the older man should go to a hospital over a long-lasting Viagra effect.	Meet the fockers / Ben Stiller & Robert Deniro
215	6s	A man and a woman are watching a ballerina. The man makes a subtle remark about the differences between the ballerina and the woman.	Black swan / Natalie Portman & Vincent Cassel
216	11s	A comical conversation between two men about starting an adventure.	
217	11s	Two men in a cab. One gives a motivational speech or a sell pitch.	Vince Vaughn
218	5s	Two men get their story straight and shake hands in elaborate ways.	Vince Vaughn
219	7s	A man talks about his marriage to a woman, who continues to hit on him.	Adam Sandler
220	10s	A man and two kids fake they're having fun together and abruptly quit as a woman exits the room.	Adam Sandler

221	7s	A couple arrives in Berlin. The man didn't sleep during the flight, but the woman did.	Liam Neeson
222	7s	A man impersonating an older lady sign into high school staff. Group of young women make remarks of her weight. She dances on a table and breaks it.	Martin Lawrence
223	8s	A woman picks up a man by a road. They discuss hitchhiking.	Nicholas Cage
224	6s	A woman praises another woman's book with a pompous title. Dramatic music is played in the background.	
225	9s	A man runs across a road, with a woman shouting something behind him.	Forrest gump / Tom Hanks
226	12s	A man is sitting on a park bench and telling an older woman about how he was recognized and were invited to somewhere.	Forrest gump / Tom Hanks
227	9s	An older woman and a younger woman are having a discussion about house rules/behavior.	
228	7s	A young man and a young woman discus what a third person said before. Intense electronic music is played in the background.	
229	5s	A young boy is telling a man goodnight over the radio. The radio breaks as the man are responding.	
230	4s	Two men are discussing over a radio about the other man's identity.	
231	4s	A father and son are having a sentimental moment over a radio.	
232	5s	A man dressed in a pro-America clothing is standing in an arena and giving a sarcastic speech.	Borat / Sacha Baron Cohen
233	8s	A dinner party. A man from a foreign country is telling that two guests, but not one, would be hot stuff in his country.	Borat / Sacha Baron Cohen
234	7s	A woman is telling a man that she wants to take him somewhere. The man answers suggestively.	Year one / Jack Black

Table SI-2. Images of the ID and the basic emotions they are intended to elicit.

Image Name	**Elicited basic emotions**
Animals_001_h	sadness
Animals_003_h	fear, happiness, disgust, surprise
Animals_007_h	fear
Animals_010_h	sadness
Animals_011_h	fear
Animals_013_h	sadness
Animals_018_h	disgust
Animals_019_h	sadness
Animals_020_h	fear, surprise, happiness
Animals_022_h	fear
Animals_025_h	sadness, anger
Animals_027_h	disgust, sadness
Animals_030_h	fear
Animals_032_h	disgust, sadness, surprise, anger, fear
Animals_033_h	disgust
Animals_037_h	disgust
Animals_038_h	sadness, anger, disgust, surprise, fear
Animals_039_h	sadness, disgust
Animals_041_h	disgust
Animals_043_h	disgust, fear, happiness, surprise
Animals_045_h	sadness, disgust, surprise, fear, anger
Animals_053_h	sadness
Animals_058_h	fear
Animals_060_h	sadness
Animals_061_h	fear, happiness, disgust, surprise, anger, sadness
Animals_062_h	disgust

Animals_065_h	disgust
Animals_067_h	sadness, surprise, disgust, fear, anger
Animals_073_h	fear, surprise
Animals_074_h	sadness
Animals_077_h	sadness
Animals_078_h	disgust
Animals_085_h	fear, happiness, surprise
Animals_087_h	fear, surprise, disgust, happiness
Animals_097_h	happiness
Animals_100_h	happiness
Animals_109_h	happiness
Animals_111_h	fear, surprise, happiness, disgust
Animals_117_h	happiness
Animals_118_h	happiness
Animals_121_h	happiness, fear, sadness, surprise, disgust, anger
Animals_127_h	happiness, disgust, surprise, fear
Animals_136_h	happiness
Animals_137_h	happiness
Animals_144_h	fear
Animals_154_h	surprise, happiness, fear, disgust, sadness, anger
Animals_159_h	happiness
Animals_165_h	happiness
Animals_173_h	happiness
Animals_177_h	happiness
Animals_181_h	happiness
Animals_182_h	happiness
Animals_183_h	happiness
Animals_185_h	happiness
Animals_186_h	happiness
Animals_195_h	happiness
Animals_196_h	happiness
Animals_197_h	happiness
Animals_203_h	happiness, sadness
Animals_205_h	happiness
Animals_213_h	happiness
Animals_218_h	happiness
Animals_220_h	happiness
Animals_221_h	disgust
Faces_001_h	happiness
Faces_006_h	sadness
Faces_007_h	sadness
Faces_011_h	sadness
Faces_017_h	sadness
Faces_019_h	sadness
Faces_021_h	sadness, fear, surprise, anger, disgust, happiness
Faces_028_h	sadness
Faces_030_h	happiness, sadness, surprise, anger, disgust, fear
Faces_032_h	anger, sadness, disgust, surprise, fear
Faces_033_h	sadness
Faces_034_h	sadness
Faces_037_h	sadness, anger
Faces_041_h	sadness
Faces_047_h	happiness
Faces_049_h	happiness
Faces_050_h	happiness
Faces_064_h	happiness
Faces_065_h	happiness, sadness, surprise
Faces_078_h	happiness

Faces_079_h	happiness
Faces_089_h	happiness
Faces_096_h	happiness
Faces_100_h	happiness
Faces_101_h	happiness
Faces_103_h	happiness
Faces_104_h	happiness
Faces_105_h	happiness
Faces_107_h	happiness
Faces_108_h	happiness
Faces_111_h	happiness
Faces_113_h	happiness
Faces_115_h	happiness
Faces_116_h	happiness
Faces_120_h	happiness
Faces_122_h	happiness
Faces_127_h	happiness
Faces_129_h	happiness
Faces_134_h	happiness
Faces_137_h	happiness
Faces_140_h	happiness
Faces_152_h	sadness
Faces_155_h	sadness
Faces_156_h	disgust, happiness, surprise
Faces_157_h	sadness
Faces_158_h	sadness
Faces_164_h	sadness
Faces_166_h	sadness
Faces_172_h	sadness
Faces_174_h	sadness, disgust, surprise, fear, anger
Faces_182_h	happiness
Faces_186_h	happiness
Faces_190_h	sadness
Faces_196_h	sadness, happiness, surprise, fear, disgust, anger
Faces_198_h	happiness, sadness, surprise, disgust, anger, fear
Faces_203_h	happiness
Faces_216_h	sadness
Faces_218_h	happiness, anger, fear, surprise, sadness, disgust
Faces_228_h	happiness
Faces_232_h	happiness
Faces_236_h	happiness
Faces_238_h	happiness
Faces_240_h	happiness
Faces_252_h	happiness
Faces_254_h	happiness
Faces_258_h	happiness
Faces_264_h	disgust
Faces_266_h	disgust
Faces_270_h	anger, fear, surprise, happiness, sadness, disgust
Faces_271_h	fear, sadness, anger, surprise, disgust, happiness
Faces_273_h	fear, anger, sadness, disgust, surprise, happiness
Faces_275_h	disgust, sadness, surprise, happiness, fear, anger
Faces_283_h	sadness
Faces_288_h	sadness
Faces_294_h	sadness, fear, surprise, anger, disgust
Faces_296_h	disgust, sadness, anger
Faces_302_h	sadness
Faces_309_h	happiness, sadness

Faces_316_h	happiness
Faces_323_h	happiness
Faces_334_h	happiness
Faces_337_h	happiness
Faces_339_h	happiness
Faces_340_h	happiness
Faces_345_h	happiness
Faces_350_h	happiness
Faces_353_h	happiness
Faces_366_h	disgust, sadness, fear, surprise
Faces_368_h	sadness, surprise, disgust, fear, anger
Landscapes_002_h	sadness, anger, fear
Landscapes_007_h	disgust, anger, sadness
Landscapes_008_h	happiness
Landscapes_020_h	sadness, happiness, surprise, anger, fear, disgust
Landscapes_021_h	happiness
Landscapes_025_h	sadness, disgust, anger
Landscapes_026_h	anger, disgust, sadness
Landscapes_042_h	happiness
Landscapes_043_h	sadness, happiness, surprise, anger, fear, disgust
Landscapes_054_h	happiness
Landscapes_060_h	sadness, fear, happiness, anger, surprise, disgust
Landscapes_064_h	happiness
Landscapes_070_h	sadness, surprise, fear, anger, disgust
Landscapes_075_h	happiness
Landscapes_083_h	happiness
Landscapes_087_h	happiness
Landscapes_091_h	happiness, surprise, sadness
Landscapes_100_h	happiness
Landscapes_117_h	happiness
Landscapes_119_h	happiness
Landscapes_121_h	happiness
Landscapes_123_h	happiness
Landscapes_126_h	happiness
Landscapes_130_h	happiness
Landscapes_139_h	disgust, anger, sadness
Landscapes_152_h	happiness
Landscapes_154_h	happiness
Landscapes_165_h	happiness
Landscapes_178_h	happiness
Landscapes_180_h	happiness
Landscapes_183_h	happiness
Objects_006_h	disgust
Objects_007_h	disgust, surprise
Objects_009_h	surprise, disgust
Objects_010_h	disgust
Objects_011_h	disgust
Objects_013_h	disgust, sadness, surprise, anger, fear
Objects_018_h	disgust, sadness, surprise, anger, fear, happiness
Objects_019_h	disgust
Objects_022_h	disgust
Objects_045_h	disgust, surprise, happiness, sadness, fear, anger
Objects_049_h	happiness
Objects_050_h	happiness, disgust, surprise
Objects_053_h	disgust, surprise, fear, happiness
Objects_060_h	disgust
Objects_069_h	happiness
Objects_086_h	happiness

Objects_088_h	disgust
Objects_093_h	surprise
Objects_109_h	disgust
Objects_119_h	happiness, surprise, fear
Objects_122_h	disgust, sadness, surprise, fear, anger, happiness
Objects_124_h	disgust, surprise, happiness, fear, sadness, anger
Objects_125_h	disgust
Objects_126_h	disgust
Objects_128_h	surprise, sadness, disgust, fear, anger
Objects_130_h	surprise, happiness, sadness, anger, fear, disgust
Objects_134_h	sadness, surprise, anger, disgust, fear
Objects_136_h	sadness, surprise
Objects_145_h	sadness, fear, surprise
Objects_148_h	fear, surprise, anger, sadness, happiness, disgust
Objects_175_h	surprise, sadness, happiness, anger, disgust, fear
Objects_177_h	anger, disgust, surprise, sadness, happiness, fear
Objects_179_h	fear, happiness, surprise
Objects_192_h	happiness
Objects_194_h	surprise, sadness, happiness, fear, anger
Objects_196_h	surprise, happiness, sadness, fear, anger, disgust
Objects_202_h	sadness, fear, anger, surprise, happiness, disgust
Objects_206_h	disgust
Objects_207_h	sadness, anger, disgust
Objects_208_h	happiness, surprise, sadness, anger, fear, disgust
Objects_209_h	happiness
Objects_211_h	happiness
Objects_222_h	surprise, happiness, fear, sadness, disgust, anger
Objects_223_h	disgust, fear, sadness, surprise, happiness
Objects_230_h	happiness, surprise, disgust, sadness, anger, fear
Objects_236_h	sadness, surprise, disgust, anger, happiness, fear
Objects_237_h	happiness
Objects_242_h	disgust, happiness, sadness, anger, surprise
Objects_245_h	happiness
Objects_247_h	happiness, sadness, surprise, fear, anger, disgust
Objects_276_h	happiness
Objects_281_h	sadness, disgust, surprise, happiness, fear, anger
Objects_282_h	sadness, fear, surprise, disgust, anger, happiness
Objects_283_h	sadness, fear
Objects_285_h	sadness, fear
Objects_295_h	happiness
Objects_306_h	happiness
Objects_308_h	happiness
Objects_310_h	happiness
Objects_311_h	happiness, sadness, anger, fear, surprise, disgust
Objects_319_h	happiness
Objects_328_h	disgust, surprise, happiness, sadness, anger, fear
People_001_h	sadness
People_003_h	sadness
People_016_h	sadness, fear, surprise, anger
People_017_h	sadness, fear, surprise
People_026_h	happiness
People_033_h	sadness
People_035_h	sadness
People_036_h	sadness, fear, surprise, anger, happiness
People_040_h	sadness
People_041_h	happiness
People_056_h	happiness, surprise, fear, sadness, disgust, anger
People_057_h	disgust

People_080_h	disgust, happiness, sadness, anger, surprise, fear
People_084_h	fear, anger, sadness, disgust, surprise, happiness
People_087_h	disgust, anger, sadness, fear
People_089_h	happiness
People_091_h	happiness
People_097_h	happiness, sadness, fear, surprise, anger, disgust
People_121_h	sadness
People_122_h	sadness
People_127_h	anger
People_128_h	sadness
People_133_h	sadness
People_139_h	fear, surprise, sadness, happiness
People_140_h	sadness
People_143_h	sadness
People_146_h	happiness, sadness, surprise, disgust, anger, fear
People_147_h	sadness
People_148_h	happiness
People_166_h	happiness
People_167_h	happiness
People_176_h	happiness
People_180_h	happiness
People_183_h	happiness
People_190_h	happiness
People_198_h	disgust
People_202_h	disgust
People_210_h	sadness
People_212_h	sadness, fear, surprise, disgust, anger
People_214_h	surprise, fear, disgust, sadness, anger, happiness
People_216_h	disgust
People_217_h	disgust, fear, sadness, surprise
People_220_h	disgust
People_222_h	disgust
People_223_h	disgust
People_226_h	sadness
People_228_h	disgust, fear, surprise, sadness
People_230_h	disgust
People_231_h	surprise, sadness, fear, disgust
People_233_h	disgust
People_237_h	surprise, disgust, fear, sadness, anger
People_239_h	disgust
People_240_h	disgust, sadness
People_241_h	disgust, surprise, fear, sadness

Table SI-3. List of the rated emotions and affective dimensions.

Emotions in video stimuli	**Task**	**Emotions in image stimuli**	**Task**
Admiration	To what extent does this make you feel admiration?	Anger	To what extent does this make you feel angry?
Adoration	To what extent does this make you feel adoration?	Arousal	To what extent does this make you feel aroused?
Aesthetic appreciation	To what extent does this make you feel aesthetic appreciation?	Disgust	To what extent does this make you feel disgusted?

Amusement	To what extent does this make you feel amusement?	Fear	To what extent does this make you feel afraid?
Anger	To what extent does this make you feel anger?	Happiness	To what extent does this make you feel happy?
Anxiety	To what extent does this make you feel anxiety?	Sadness	To what extent does this make you feel sad?
Approach	To what extent does this make you feel like this is something you would want to approach?	Surprise	To what extent does this make you feel surprised?
Awe	To what extent does this make you feel awe?	Valence	To what extent does this make you feel satisfied?
Awkwardness	To what extent does this make you feel awkwardness?		
Boredom	To what extent does this make you feel boredom?		
Calmness	To what extent does this make you feel calmness?		
Certain	To what extent does this make you feel certain?		
Commitment	To what extent does this make you feel commitment to an individual or creature?		
Confusion	To what extent does this make you feel confusion?		
Contempt	To what extent does this make you feel contempt?		
Control	To what extent does this make you feel like things are under control?		
Craving	To what extent does this make you feel craving?		
Disappointment	To what extent does this make you feel disappointment?		
Disgust	To what extent does this make you feel disgust?		
Dominant	To what extent does this make you feel dominant?		
Effort	To what extent does this make you feel like viewing this demands effort?		
Empathic pain	To what extent does this make you feel empathic pain?		
Entrancement	To what extent does this make you feel entrancement?		
Envy	To what extent does this make you feel envy?		
Excitement	To what extent does this make you feel excitement?		

Fear	To what extent does this make you feel fear?
Focused	To what extent does this make you feel focused?
Guilt	To what extent does this make you feel guilt?
Horror	To what extent does this make you feel horror?
Identity	To what extent does this make you feel like you identify with a group of people?
Interest	To what extent does this make you feel interest?
Joy	To what extent does this make you feel joy?
Nostalgia	To what extent does this make you feel nostalgia?
Not fair	To what extent does this make you feel like things are not fair?
Obstruction	To what extent does this make you feel like you are obstructed by something?
Pleasant	To what extent does this make you feel pleasant?
Pride	To what extent does this make you feel pride?
Relief	To what extent does this make you feel relief?
Romance	To what extent does this make you feel romance?
Sadness	To what extent does this make you feel sadness?
Safety	To what extent does this make you feel a sense of safety?
Satisfaction	To what extent does this make you feel satisfaction?
Sexual desire	To what extent does this make you feel sexual desire?
Stimulated	To what extent does this make you feel stimulated?
Surprise	To what extent does this make you feel surprise?
Sympathy	To what extent does this make you feel sympathy?
Triumph	To what extent does this make you feel triumph?
Upswing	To what extent does this make you feel like this went better than it first seemed it would?

Table SI-4. Nationalities of the human observers for video datasets.

Video Dataset 1		**Video Dataset 2**	
Nationality	**Numbers of person**	**Nationality**	**Numbers of person**
South Africa	154	South Africa	90
Poland	22	Poland	16
United Kingdom	22	Italy	13
United States	21	Greece	11
Nigeria	16	United Kingdom	11
Zimbabwe	16	Hungary	6
Greece	7	Portugal	6
Chile	5	Spain	6
Kenya	5	Zimbabwe	6
Mexico	5	Czech Republic	3
Australia	4	Estonia	3
Hungary	4	Finland	3
Spain	4	Mexico	3
Croatia	3	France	2
Estonia	3	Germany	2
Italy	3	Kenya	2
Portugal	3	Netherlands	2
Brazil	2	Slovenia	2
France	2	United States	2
Slovenia	2	Brazil	1
Vietnam	2	Bulgaria	1
Belgium	1	Canada	1
Canada	1	Chile	1
Democratic Republic of the Congo	1	India	1
Ethiopia	1	Indonesia	1
Germany	1	Iran	1
India	1	Ireland	1
Indonesia	1	Israel	1
Macedonia	1	New Zealand	1
Sweden	1	Philippines	1
Turkey	1	Sweden	1
		Turkey	1
		Vietnam	1