Affective computing has changed: the foundation model disruption

Björn Schuller, Adria Mallol-Ragolta, Alejandro Peña Almansa, Iosif Tsangko, Mostafa M. Amin, Anastasia Semertzidou, Lukas Christ, Shahin Amiriparian

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Adria Mallol-Ragolta, Alejandro Peña Almansa, Iosif Tsangko, Mostafa M. Amin, Anastasia Semertzidou, Lukas Christ, and Shahin Amiriparian. 2026. "Affective computing has changed: the foundation model disruption." *npj Artificial Intelligence* 2 (1): 16. https://doi.org/10.1038/s44387-025-00061-3.

npj | artificial intelligence

Article

https://doi.org/10.1038/s44387-025-00061-3

Affective computing has changed: the foundation model disruption

Björn Schuller[1,2,3,4,5] ✉, Adria Mallol-Ragolta[1,3], Alejandro Peña Almansa[5,6], Iosif Tsangko[1,3,5], Mostafa M. Amin[1,5,7], Anastasia Semertzidou[1], Lukas Christ[1] & Shahin Amiriparian[1]

The dawn of Foundation Models has on the one hand revolutionised a wide range of research problems, and, on the other hand, democratised the access and use of AI-based tools by the general public. We even observe an incursion of these models into disciplines related to human psychology, such as the Affective Computing domain, suggesting their affective, emerging capabilities. In this work, we aim to raise awareness of the power of Foundation Models in the field of Affective Computing by synthetically generating and analysing multimodal affective data, focusing on vision, linguistics, and speech (acoustics). We also raise awareness of evaluation problems related to the use of Foundation Models in this research area.

"*The world of Affective Computing has changed. I see it in the vision modality. I read it in the linguistic modality. I hear it in the speech modality. Much that once was is outdated...*" This quote, which might sound slightly familiar to the J. R. R. Tolkien's readers, aims to literary exemplify how the disruption of Foundation Models (FM) might be impacting the Affective Computing research as we knew it. Before centring the discussion on this topic, we summarise where we come from.

Early works on affect recognition relied on conventional Machine Learning (ML) pipelines, in which expert-crafted features capturing emotional content were first extracted from the raw data, and then processed utilising traditional methods; e.g., Support Vector Machines (SVM). The success of Deep Learning (DL) in the early 2010's shifted the paradigm towards using representational learning via Deep Neural Networks (DNN). Consequently, feature engineering took a backseat when data-driven approaches unleashed their potential, leading to the first disruption: the learning of representations. The efforts were ultimately shifted from experts crafting representations to experts choosing model architectures to learn these representations. The data-driven disruption also contributed to improving emotional data synthesis[1–6]. A second, albeit less noted and exploited potential disruption came with the possibility of (neural) architecture search by reinforcement learning[7] or more efficient approaches[8]. This meant that in principle, once having affective (labelled) data, the representation could be learnt to then analyse affective data as well as the best architecture to do so.

Yet, a critical issue was the acquisition of such reliable, annotated data. The subjectivity of measuring inner emotion through self-assessment shifted the focus to perceived emotion. However, 'measuring' outer perceived emotion usually requires several labellers to reduce uncertainty, hence coming at high effort and cost. Moreover, the lack of spontaneous data sources—e.g., due to privacy restrictions—favoured the use of acted/elicited, non-spontaneous data samples[9,10]. Whilst non- or less-spontaneous data eased the problem of data availability, it came with the drawback that the analysis of real-world emotion struggles with subtlety not met in training. The acquisition of data from Internet sources (e.g., social media, films) allowed the collection of large "in-the-wild" databases[11,12], whose annotations were obtained through semi-automatic methods, crowdsourcing, or based on the criteria of experts in affect.

A third disruption is taking place nowadays in the community, as—based on current developments—even specialised annotated affective data for training the models may no longer be needed, since affective computing abilities start to emerge in (general large data) pre-trained FMs. Nevertheless, curating high-quality sets of annotated data to some extent remains crucial to assess the performance of the models. New FMs[13–16] have demonstrated surprising capabilities using prompt-based instructions, to the point that they can generate realistic data samples or perform zero-shot classification. The extent of the affective capabilities of these models, and the potential they open up, is still uncertain. Herein, we aim to shed some light on this topic, and explore how the emergence of FMs is influencing the Affective Computing community.

Results

We first summarise the datasets and prompt templates, models, and metrics used to probe FM emergence, then report per-modality findings. Two of the main characteristics of the FMs are that i) they are trained on a broad range

[1]CHI – Chair of Health Informatics, TUM University Hospital, Munich, Germany. [2]MDSI – Munich Data Science Institute, Munich, Germany. [3]MCML – Munich Center for Machine Learning, Munich, Germany. [4]GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK. [5]EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany. [6]School of Engineering, Universidad Autonoma de Madrid, Madrid, Spain. [7]AI R&D Team, SyncPilot GmbH, Munich, Germany. ✉e-mail: schuller@ieee.org

Table 1 | Attributes defined in the input prompts to synthesise emotional facial images with Stable Diffusion XL[16]

Attribute	Values
Prompt template	Face image of a <*age*> <*sex*> with <*skin*> skin, with a <*emotion*> face, in a <*style*> style, realistic eyes, white background, ultra quality, frontal picture, looking at camera
Negative prompt	Disfigured, unrealistic eyes, blurry, b&w, <*style*>
Emotion	*Neutral, fear and terror, anger and rage, happiness and joy, sadness and grief, disgust and loathing, surprise and amazement*
Age	*Young, middle-aged, old*
Sex	*Man, woman*
Skin tone	*White, brown, black*
Style	*Photorealistic, cartoon and painting, anime, 3D Pixar animation*

The prompt template considers three different sources of variation: the emotion, the style, and the demographic group. The latter is determined by three different demographic attributes: age, biological sex, and skin tone. We have also utilised a negative prompt, which includes all the styles that are not desired in the current synthesis.

Table 2 | Summary of the face images generated with the Stable Diffusion XL model[16]

Emotion	Photorealistic	Cartoon	Anime	3D
Neutral	233	132	131	143
Fear	185	55	94	136
Anger	183	165	119	119
Happiness	223	202	118	142
Sadness	179	156	137	99
Disgust	173	90	56	40
Surprise	184	147	120	139
Σ	1360	947	775	818

The prompts for the generation are detailed in Table 1.

of data, so that the resulting models can be utilised in a wide range of problems, and ii) they exploit large amounts of learning parameters. Given sufficient learning material, from a certain number of such parameters hence well trained, knowledge 'emerges' in the FMs, and they achieve competitive performances in tasks they have not been specifically trained for. This can, however, be difficult to predict[17]. We aim to investigate the 'emergent' affective capabilities of FMs. Focusing on the vision (cf. Section 2.1), linguistics (cf. Section 2.2), and speech (acoustics) (cf. Section 2.3) modalities, we assess the capabilities of current FMs to i) generate synthetic affective samples, from which we infer the conveyed emotions with pre-trained emotion recognition classifiers (note that in principle, this can lead to a 'closed loop', as we cannot be sure whether the data used in the pre-trained emotion classifiers has not also been used in the training of the FMs, but generally, it is unlikely, as high-quality affective data are rarely freely available on the Internet due to their privacy restrictions), and ii) analyse well-established datasets in the field in a zero-shot manner. To favour the comparability among the different modalities investigated, we focus on the 'Big Six' Ekman emotions[18] (i.e., fear, anger, happiness, sadness, disgust, and surprise), in addition to the neutral state.

The research gap this work aims to fill is the assessment of the affective capabilities of current FMs, including a first impression of to which extent these represent a breakthrough in the field of Affective Computing. Since affect plays a pivotal role in human-related interactions, it is of paramount importance for AI systems—and for any system interacting with humans—to be able to understand and properly respond to the emotional state of the user. Understanding how FMs react to human emotions and how well they conform to or deviate from emotion theory is a key step towards the integration of this technology at the core of AI-powered agents. The scope of our analysis is constrained by the fundamental nature of our emotion model, which is based on Ekman's theory of discrete emotions. However, as we have discussed, this approach allows us to unify the evaluation of the three modalities addressed here. Before exploring complex models—such as the characterisation of emotions in the dimensional (i.e., valence and arousal) space—, it is important to start the analysis of the capabilities of FMs with a model that, despite limited, is extremely useful in a wide range of tasks and applications. Moreover, the use of discrete emotions could help us gain more insights into the model performances and associated weaknesses or biases that may have emerged from their training.

The vision modality has changed

We proceed to discuss the *generation* capabilities of current vision-based FMs. In the visual domain, data synthesis started to obtain pseudorealistic results in the last decade thanks to the Generative Adversarial Network-(GAN) based models[1,6]. Nowadays, a boost in the quality of the synthesised images has been achieved via text prompt inputs-based models, due to i) the CLIP model[19] and ii) the Diffuser model[20]. The former was presented as a model to predict how well a given caption describes an image. The latter learns to reconstruct images by removing an added Gaussian noise through a Markov Chain. During inference, the model generates new images from Gaussian noise, being more efficient than other generative architectures. Models such as Stable Diffusion (SD)[21] or DALL-E 2[22] integrate the CLIP and the Diffuser models to efficiently synthesise images with high semantic control.

We have leveraged one of the latest versions of SD—i.e., Stable Diffusion XL (SDXL)[16]—to synthetically generate a face emotion dataset. This dataset is generated utilising predefined prompts with three sources of variation: i) the emotion, ii) the style (photorealistic, cartoon-painting, anime, and 3D), and iii) the demographic group. The template, along with the values explored for each attribute, are detailed in Table 1. Table 2 presents a summary of the gathered dataset. Although our emotion model is based on the 'Big Six' Ekman emotions[18] plus the neutral state, we employed these basic emotions together with a higher intensity variation—as defined in Plutchik's model[23]—to further emphasise the desired affective states. The generation process spans 18 different demographic groups; determined by age, biological sex, and skin tone. Visual examples for each emotion and style are provided in Fig. 1 for the demographic group <*young*, *woman*, *white*>. For the case of the photorealistic style, we played with the background to generate some samples in realistic scenarios (e.g., outdoors, office, park). The generation process involved two experts, the SDXL base model and the SDXL refiner, with a total of 40 steps (80% in the base model, 20% in the refiner) and a guidance value of 7.5. Together with the desired prompt, we utilised a negative prompt to highlight what we did not want to see in the output. Once the images were generated, one annotator curated the data according to four principles: the presence of dissatisfactions or artifacts, nudity, prompt compliance and plausibility of the style.

We now aim to automatically verify the affective quality of the generated facial images with Face Emotion Recognition (FER) models, which we train employing the manually annotated subset of the AffectNet dataset[11]. The images belonging to this dataset are annotated in terms of eleven emotions. Nevertheless, we select the images corresponding to the

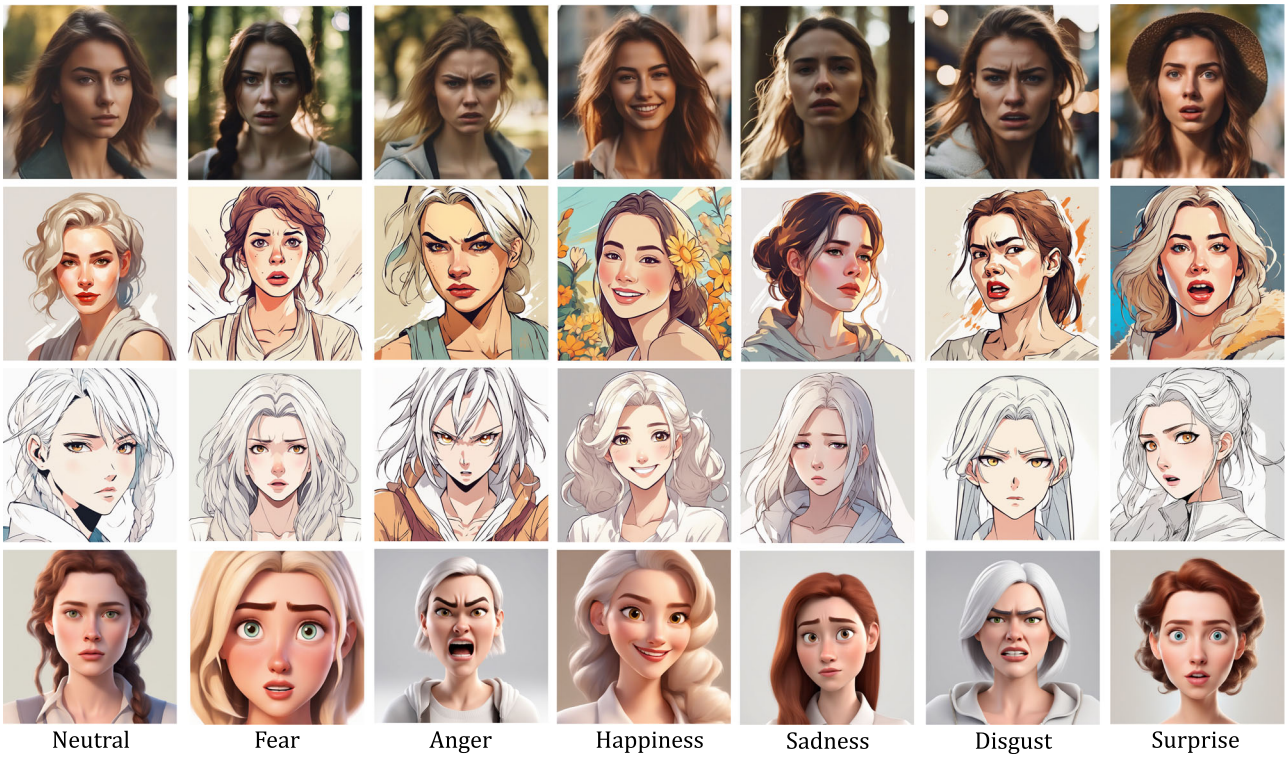


Fig. 1 | Synthetic facial images of a white-skin, young woman conveying the 'Big Six' Ekman emotions[18], in addition to the neutral state. All the images were generated with Stable Diffusion XL[16], conditioned on four different styles, namely photorealistic (first row), cartoon-painting (second row), anime (third row), and 3D (fourth row).

Table 3 | Summary of the face images selected from AffectNet[11] in the training and the validation partitions

Emotions	Training	Validation
Neutral	74,873	500
Fear	6378	500
Anger	24,881	500
Happiness	134,411	500
Sadness	25,458	500
Disgust	3803	500
Surprise	14,090	500
Σ	283,894	3500

Table 4 | Performance summary of the trained Support Vector Classifier-based models for Face Emotion Recognition on the validation partition of the considered subset of AffectNet

Model				ACC (%)
	Kernel	*Weighted Samples*	*Optimal C*	
SVC	Linear	✗	10	27.3
	Linear	✓	10^2	28.7
	Rbf	✗	10^2	28.5
	Rbf	✓	10^2	29.7
ViT – FER				43.9
Chance level				14.3

We also include the performance of a state-of-the-art Vision Transformer for Facial Expression Recognition (ViT–FER). We select the accuracy (ACC) as the metric to assess the model performances.

'Big Six' Ekman emotions in addition to the neutral class. We process the selected images with OpenFace[24] to extract features related to a subset of the Action Units (AU) defined in the Facial Action Coding System (FACS)[25]. Specifically, OpenFace extracts 35 features per facial image, indicating the presence (0 or 1) and the intensity (in a scale from 0 – not present – to 5 – present with maximum intensity) of a subset of the AUs. We discard the images that OpenFace fails to process; for instance, due to the absence of a face in the image. Table 3 summarises the resulting data.

We start our preliminary investigation by training FER models with Support Vector Classifiers (SVC), as these are considered a standard machine learning technique with excellent results in a wide range of problems. We compare their performance when utilising a linear and a Radial Basis Function (RBF) kernel. To overcome the imbalanced training samples (cf. Table 3), we weigh the training data, so that the samples corresponding to the least represented classes have more importance when training the models. We fine-tune our models optimising the regularisation parameter $C \in [10^{-2}, 10^{-1}, 1, 10, 10^2]$. The performance of the optimal models on the validation partition is depicted in Table 4.

We also contrast the performance of the SVC-based FER models with a state-of-the-art Vision Transformer[26] for Facial Expression Recognition (ViT–FER), trained on the Facial Expression Recognition 2013 (FER-2013) dataset[27]. We use the pre-trained model off-the-shelf—without fine-tuning—and evaluate its performance on the validation partition of our subset of AffectNet. The obtained results are reported in Table 4.

The best performance on the validation partition (cf. Table 4) is obtained with the ViT–FER model. We use this model to assess the affective quality of the generated facial images. The results obtained exemplify the breakthrough of working with end-to-end approaches, operating on the raw images instead of on the features extracted from them. Nevertheless, it is worth mentioning that AffectNet was collected in the wild, which may complicate the estimation of the AUs from the facial images and, in turn, worsen the performance of the SVC-based models.

Table 5 summarises the results obtained when analysing the generated facial images with the four different styles utilising the ViT–FER pre-trained

model. Figure 2 presents the confusion matrices computed, comparing the model inferences and the ground truth. Across styles, the worst results are those of the photorealistic style, as denoted by both accuracy—Weighted Average Recall (WAR)—and Unweighted Average Recall (UAR). UAR is the sum of recall per class divided by the number of classes—this reflects imbalances and is a standard measure in the field. In contrast, the best results are obtained with the 3D style. Interestingly, in all the 4 styles both the neutral and the happiness emotions consistently obtain the best results. These classes are traditionally over-represented in face datasets (e.g., see the training set distribution of AffectNet in Table 3), probably due to the nature of the data sources[28]. Thus, it is expected for the generative models to be biased towards these emotions, making their recognition easier. As can be observed from Fig. 2, the other emotions obtain low recognition rates, not even reaching a 50% rate in most of the cases. An interesting, recurrent mistake across all 4 styles is the models tendency towards predicting the neutral class. This result aligns with the bias associated with the imbalanced emotional datasets. The limited amount of samples from some concrete emotions available for training the generative models makes the generation of samples representative of these classes harder. By "harder" we do not only refer to the correct generation according to the prompted emotion, but also to convincingly convey the intensity of the desired emotion.

We proceed to discuss the *analysis* capabilities of current vision-based FMs. To evaluate the affective analytical capabilities of FMs in the image domain, we explore their performance in a zero-shot emotion recognition task under different configurations. Our experiments are conducted on the considered validation set of AffectNet[11], as it is balanced, in-the-wild, and manually annotated. We compare three different approaches relying on model-prompting. In the first two, we provide the presence of a subset of the AUs and their intensities in a textual format as input to a FM. The third approach consists in directly feeding the images within the prompt of a FM. The first two approaches can be addressed with a Language Foundation Model (LFM), for which we select the LLaMA2 7B model. We utilise a Multimodal Foundation Model (MFM) for the third scenario; specifically, the LLaVA1.5 7B model[29,30]. The selected models have the same number of parameters.

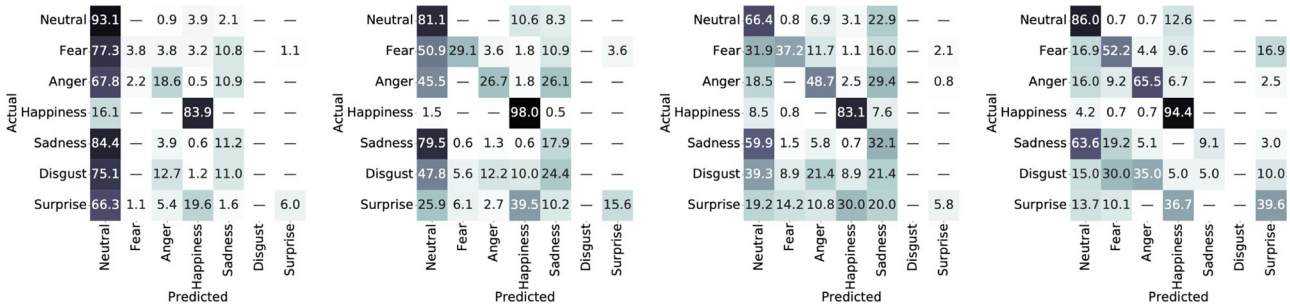
In Table 7, we present the results of the aforementioned scenarios. The prompts utilised are detailed in Table 6. We also include in Table 7 the results of the ViT–FER model for comparability purposes. We obtain the best results when feeding the prompts with the raw images. Both AU-based prompt approaches exhibit accuracy results close to chance. The LLaVA model achieves an accuracy only 4 points below the ViT–FER model, which was explicitly trained on the FER-2013 dataset to recognise emotions. This is an interesting result, which suggests that the LLaVA1.5 model presents emergent affective capabilities, despite not being specifically trained on affective computing tasks.

Table 5 | Accuracy (ACC) and Unweighted Average Recall (UAR) scores obtained when analysing the facial images generated in the four different styles with the ViT–FER pre-trained model

	Generated styles			
	Photorealistic	**Cartoon-painting**	**Anime**	**3D**
ACC (%)	35.0	43.9	42.5	57.5
UAR (%)	30.9	38.3	39.1	49.5

The linguistic modality has changed

We proceed to discuss the *generation* capabilities of current language-based FMs. The introduction of the Transformer model[31] has revolutionised the Natural Language Processing (NLP) field and marked the



(a) Photorealistic (b) Cartoon-painting (c) Anime (d) 3D

Fig. 2 | Confusion matrices obtained by analysing the facial images generated according to the four different styles with the ViT–FER pretrained model. a–d Correspond to the photorealistic, cartoon-painting, anime, and 3D styles, respectively.

Table 6 | Prompts employed to perform zero-shot emotion recognition with Foundation Models in different scenarios

Approach	Prompt template
AU presence	<s> [INST] <<SYS>> You are a highly skilled Affective Computing system with an expertise in accurately predicting emotion classes from Action Units. I will provide you a list of Action Units present in a face. Your task is to answer the most likely emotion class, without any further explanation. Please, provide only one of the following classes as answer: Neutral, Fear, Anger, Happiness, Sadness, Disgust, Surprise. The question is, which is the most likely emotion if the following Action Units are present <</SYS>> "{AU}".[/INST] ### Response:
AU intensity	<s> [INST] <<SYS>> You are a highly skilled Affective Computing system with an expertise in accurately predicting emotion classes from Action Units. I will provide you a JSON object with the intensities of the Action Units present in a face. Your task is to answer the most likely emotion class, without any further explanation. Please, provide only one of the following classes as answer: Neutral, Fear, Anger, Happiness, Sadness, Disgust, Surprise. The question is, which is the most likely emotion if the following Action Units are present <</SYS>> "{AU}".[/INST] ###Response:
Image	<image> USER: You are provided with a face image of a person. Classify the most likely emotional state depicted into one of the classes between brackets [Neutral, Fear, Anger, Happiness, Sadness, Disgust, Surprise] ASSISTANT:

The prompts including (i.e., first two rows) AU information are injected in LLaMA2[39], while the prompt including the raw image is injected to a LLaVA1.5 model[30].

Table 7 | Accuracy scores obtained with the LLaMA2[39] and LLaVA1.5[30] Foundation Models on the validation set of AffecNet[11]

Model	Input	ACC (%)
LLaMA2 7B	Prompt with information of AU presence	18.7
LLaMA2 7B	Prompt with description of AU presence	13.4
LLaMA2 7B	Prompt with information of AU intensity	17.9
LLaMA2 7B	Prompt with description of AU intensity	16.8
LLaVA1.5 7B	Prompt with image	39.3
LLaVA1.5 7B	Prompt with image and AU presence	20.5
ViT – FER	Image	43.9
Chance level		14.3

We have included as well the performance obtained with the ViT–FER model[26] trained on FER-2013[27] for comparison purposes.

Table 8 | Affective style transfer example towards the emotion 'surprise' with the neutral phrase: "*The weather is clear and sunny.*"

Model	Affective phrase
Mixtral	*Wow! What a surprise! The sky is astonishingly bright and clear today!*
Mistral	*The sudden emergence of unobstructed sunlight has taken me by complete astonishment!*
LLaMA	*It comes as quite a shock to discover that the sky has transformed itself into such crystal clarity!*

Table 9 | Statistics of the considered subset of the GoEmotions dataset

Emotion	Training	Validation	Test
Neutral	12,823	1592	1606
Fear	515	66	77
Anger	3878	485	520
Happiness	12,920	1668	1603
Sadness	2121	241	259
Disgust	498	61	76
Surprise	3553	435	449
Σ	36,308	4548	4590

Table 10 | Performance scores of the implemented models when inferring the emotions (Ekman's 'Big Six' in addition to the neutral state) conveyed by the sentences belonging to the test set of the GoEmotions dataset

Model	ACC (%)	UAR(%)
BiLSTM	53.53	51.44
RoBERTa	**69.22**	**62.82**
Chance level	14.29	14.29

The bold values indicate the highest performance (best score) for each metric.

beginning of the Large Language Models (LLM) era. OpenAI's work with the Generative Pre-trained Transformer (GPT) models[13,32]—culminating with the recent GPT-4[33]—is the greatest exponent of this revolution, advancing text generation. From an Affective Computing perspective, LLMs present a novel approach to inject and transfer emotional content in linguistic data, with recent works demonstrating their intrinsic emotional capabilities in a variety of domains[34–38].

We investigate and leverage the affective style transfer capabilities of cutting-edge open-source LLMs. We select LLaMA2[39] and Mistral[40]—comprising 7 billion parameters each—, given their proven high performance in both language understanding and text generation tasks. Additionally, we include the Mixtral LLM[41], which utilises the Sparse Mixture of Experts (SMoE) method[42]. We first compile a corpus comprised of 122 human-curated neutral phrases. The dataset encompasses various topics, from mundane personal activities to formal professional interactions. Utilising the gathered corpus, we task the aforementioned LLMs (i.e., LLaMA2, Mistral, and Mixtral) to generate six emotional phrases from each original neutral phrase, conveying one of the 'Big Six' Ekman emotions. We generate three synthetic sentences for each combination of text and emotion, yielding a corpus of 2194 emotional phrases (the 2 missing sentences were discarded during preprocessing due to a corrupted generation). The generation parameters used a temperature of 0.9, top-p of 0.6, and a repetition penalty of 1.2. An example of the generated sentences is shown in Table 8.

To investigate the quality of the generated sentences, we implement two baseline models trained on the GoEmotions dataset[43], a well-renowned corpus in the field and commonly utilised for benchmarking purposes due to its comprehensive labelling and categorisation of the emotions. The GoEmotions dataset consists of English Reddit comments annotated according to 27 distinct emotions, plus the neutral state, by 3 or 5 labellers each. Due to the nature of the annotations in the GoEmotions dataset, we begin by tailoring the data to meet the specific requirements of our experiments. First, we select instances from the dataset annotated with a single emotion, to tackle the task as a single-label classification problem, instead of a multi-label classification one. As previously, we adopt the 'Big Six' Ekman emotions[18], in addition to a seventh neutral state. This restructuring of the GoEmotions taxonomy to the Ekman taxonomy is achieved by aggregating the original labels into the targeted, broader categories[43]—i.e., emotions like annoyance and irritation, originally distinct, were grouped under 'anger' to fit the Ekman model. Table 9 summarises the statistics of the considered subset of the GoEmotions dataset.

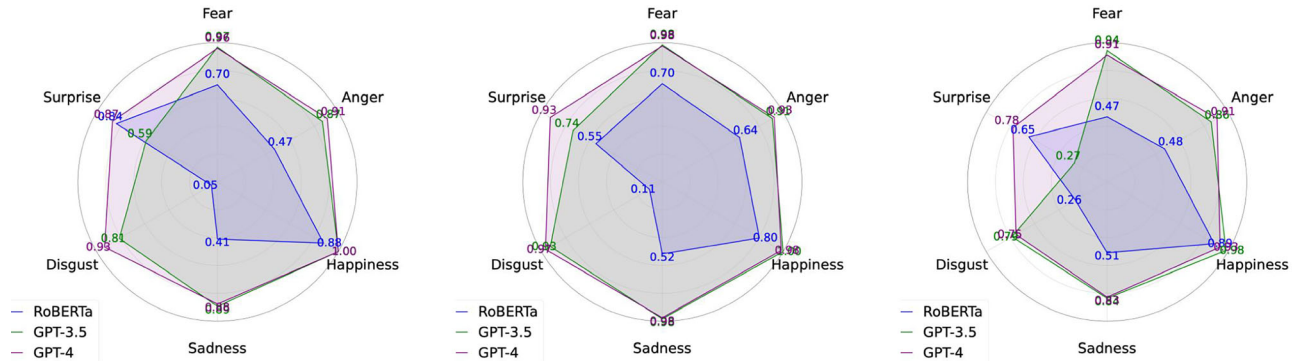
We employ two different architectures as baselines: a Bi-directional Long Short-Term Memory (BiLSTM) network and a fine-tuned version of the RoBERTa-base model[44]. Both are trained on the selected subset of the GoEmotions dataset. The BiLSTM consists of two bidirectional Long Short-Term Memory (LSTM) layers with 128 units each. It is trained with a learning rate of 5×10^{-3} and a batch size of 96 for 40 epochs, while RoBERTa-base was fine-tuned at a conservative learning rate of 5×10^{-5} and a smaller batch size of 12. The models are trained on the training partition of the dataset, and the weights yielding the highest validation UAR are selected for each model. The test scores for both baseline models (BiLSTM and RoBERTa) on the test partition of the GoEmotions dataset are shown in Table 10. Given the superior performance of RoBERTa, we utilise it for analysing the synthetically generated emotional sentences.

To evaluate the performance of the various LLMs on the emotion injection task, we test the generated sentences with the RoBERTa baseline model. We also explore GPT-4 as an approximation for human evaluation, and its weaker variant GPT-3.5. Table 11 demonstrates the prompt templates used for the GPT models, inspired by previous works[37,45]. The versions of GPT variants used are 'gpt-3.5-turbo-0125' for GPT-3.5 and 'gpt-4-turbo-2024-04-09' for GPT-4. The results of this evaluation are depicted in Fig. 3. These results reflect a better agreement between models than with the ground truth labels, which are not human-annotated. However, the results of GPT-4 should be the closest to the human evaluations[37,46].

The GPT-4 model achieves high UAR scores on all six emotions. Its inferior model GPT-3.5 achieves slightly worse results in most cases, but it experiences a performance drop in the recognition of surprise. In comparison, RoBERTa has a different behaviour. It is generally much worse than GPT-4 and GPT-3.5 in most of the cases, but it obtains a higher score than GPT-3.5 for the surprise emotion with the LLaMA2- and the Mixtral-

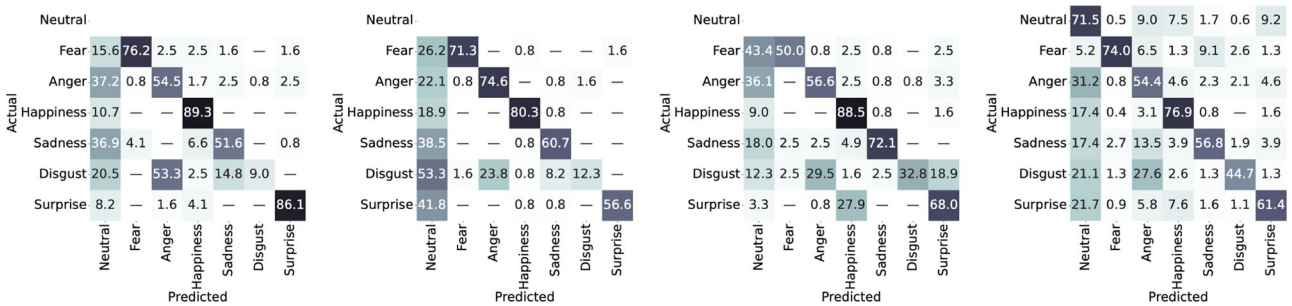
Table 11 | Prompts to use LLMs for zero-shot emotion recognition, following a similar pattern as in refs. 37 and 45

Prompt template
You are an expert at affective computing. Given a text by the user, analyze which emotion is most dominant in the given text. Only classify one of the seven Ekman emotions, namely: 'neutral', 'fear', 'anger', 'happiness', 'sadness', 'disgust', 'surprise'. You are only allowed to answer with EXACTLY ONE word corresponding to the aforementioned seven emotions.
In case of multiple emotions, use ONLY the ONE emotion you are most confident about.
Use the following format:
* You are only allowed to answer with one of the following seven words:
"neutral", "fear", "anger", "happiness", "sadness", "disgust", "surprise".
* Don't write an explanation of the answer.
* Don't write things like "My guess is...", or "I think ...".
Just write the emotion, and nothing else.



(a) LLaMA2-based synthesis. (b) Mistral-based synthesis. (c) Mixtral-based synthesis.

Fig. 3 | UAR scores obtained with the RoBERTa, GPT-3.5, and GPT-4 models when recognising the emotions conveyed by the synthetic sentences generated by LLaMA2 (left), Mistral (centre), and Mixtral (right). a–c Correspond to LLaMA2-based, Mistral-based, and Mixtral-based synthesis, respectively.



(a) LLaMA2 (b) Mistral (c) Mixtral (d) GoEmotions

Fig. 4 | Confusion matrices showing the performance (in %) of the fine-tuned RoBERTa baseline on the synthesised benchmarks, generated by LLaMA2, Mistral, and Mixtral, respectively, in addition to the GoEmotions test benchmark. a–d Correspond to the LLaMA2-based, Mistral-based, Mixtral-based synthetic benchmarks and the GoEmotions test benchmark, respectively.

generated sentences. Additionally, RoBERTa is showing very low performance for disgust.

Figure 4 depicts the confusion matrices obtained with the RoBERTa baseline model on the LLaMA2-, the Mistral-, and the Mixtral-generated sets. We also include its performance on the test set of GoEmotions as a reference. A common issue with the RoBERTa model is the confusion between anger and disgust. Analysing the confusion matrices, we also observe an interesting effect: most of the model's mispredictions are assigned to the neutral class.

Further analysing the results presented in Fig. 4, we observe that three systematic error patterns emerge. First, a noticeable "neutral-fallback" effect: although the synthetic test sets do not contain neutral ground-truth sentences, RoBERTa assigns the neutral label to a sizeable share of inputs with negative valence—e.g., with LLaMA2, Mistral and Mixtral a 15.6%, 26.2%, and 43.4% of the fear instances are misclassified into the neutral class, respectively; a 36.9%, 38.5%, and 18.0% of the sadness instances are misclassified into the neutral class, respectively; and a 20.5%, 53.3%, and 12.3% of the disgust instances are misclassified into the neutral class, respectively. Second, a one-way anger-disgust drift: more than half of the disgust instances generated by LLaMA2 are misclassified into the anger class (53.3%), with consistent though smaller shifts for Mistral (23.8%) and Mixtral (29.5%). The reverse confusion remains below 3% in all the three scenarios. Third, a robust performance of positive emotions: happiness is recognised with ≥80% recall across all synthetic sets, and the recognition of

Table 12 | Performance scores of the different LLMs tested on a zero-shot fashion for recognising the corresponding emotion on the sentences belonging to the test partition of the GoEmotions dataset

Model	Recall (%)							UAR	ACC
	Neutral	Fear	Anger	Happiness	Sadness	Disgust	Surprise	(%)	(%)
LLaMA2-7B	4.51	33.77	42.31	66.60	39.00	40.79	58.04	40.72	38.78
LLaMA2-13B	11.89	40.26	**67.31**	69.73	37.45	40.79	28.12	42.22	42.39
Mistral	57.95	57.14	47.31	22.08	**58.30**	40.79	9.60	41.88	39.20
Mixtral	53.50	66.23	48.65	52.85	50.19	30.26	14.96	45.24	48.59
GPT-3.5	23.84	**75.32**	56.35	65.54	40.54	53.95	18.08	47.66	43.85
GPT-4	38.42	68.83	42.50	50.28	36.68	**65.79**	26.56	47.01	42.74
RoBERTa	**71.53**	74.03	54.42	**76.86**	56.76	44.74	**61.38**	**62.82**	**69.22**

The bold values indicate the highest performance (best score) for each metric.

surprise leads to limited confusion—achieving recall rates of 86.1%, 56.6%, and 68.0%, respectively. Overall, the confusion matrices indicate that, while prompt-based style transfer still under-specifies nuanced negative emotions, it embeds strong, distinctive cues for high-arousal emotions.

We proceed to discuss the *analysis* capabilities of current language-based FMs. We analyse the zero-shot sentiment analysis capabilities of the following LLMs: Mistral, Mixtral, and two versions of LLaMA2 (7 billion and 13 billion parameters). We assess their zero-shot capabilities on the test partition of the GoEmotions dataset. We design a prompt that requires the selected LLMs to predict the corresponding emotion, including the neutral state (cf. Table 11). To minimise randomness and increase confidence in the predictions, we reduce the temperature setting to 0.1. This lower temperature sharpens the probability distribution, ensuring that the predicted classes reflect those that the LLMs are predicting with the highest probabilities[45]. However, the LLM outputs sometimes include irrelevant or multiple emotions. To address this without intrusively altering the model's outputs, we select the first listed emotion as the most reliable prediction.

Table 12 summarises the comparative performance of the tested LLMs. We include the performance of the RoBERTa baseline model trained on the GoEmotions dataset for benchmarking purposes. The first observation is that the UAR scores obtained by all the investigated LLMs surpass the chance level (14.3%), underscoring the emergent affective capabilities of the LLMs tested in a zero-shot manner. Nevertheless, none of the LLMs outperforms the RoBERTa baseline model, fine-tuned on the GoEmotions dataset, which indicates that model-specific tuning is advantageous. Nevertheless, it is worth highlighting that the difference in the UAR scores obtained by the best-performing GPT-3.5 and GPT-4 models in comparison to the RoBERTa baseline model is around 15%. This is an interesting result, as these models have not been trained on the GoEmotions dataset, but still obtained a competitive performance on the emotion recognition task, reinforcing, one more time, the emergent affective capabilities of the models.

The speech modality has (not yet) changed

We proceed to discuss the *generation* capabilities of current speech-based FMs. Research on generating affective speech has been conducted for more than three decades[47]. The first approaches were rule-based, while contemporary methods typically rely on DL. All of these methods are explicitly engineered to produce emotional speech. Thus, this line of research can be dubbed as a subfield of 'traditional' Affective Computing, while technically belonging to the Text-To-Speech (TTS) field. In recent years, similar to the developments in NLP and Computer Vision (CV), research on TTS has heavily been influenced by the success of the FM paradigm[48,49].

UniAudio[48] is a Transformer-based general-purpose audio synthesis system, pretrained on seven generative audio tasks including TTS. In its pretrained state, however, it does not support affective speech synthesis. The authors demonstrate that their pretrained model serves as a basis for adaptation to different downstream tasks which opens up the possibility to equip the model with affective speech synthesis capabilities by mere fine-tuning. Another recent generative audio FM, PromptTTS2[49], synthesises speech based on text prompts that include descriptions of the voice to be generated. The controllable attributes of the synthesised speech must be defined during the training time. The authors of ref. 49 do not investigate emotion as one such attribute, though their proposed framework would permit this. To the best of our knowledge, no evidence for affective speech synthesis as an emergent capability of FMs has been published so far. At the moment, although the demos for GPT-4o claim affective speech synthesis capabilities, neither a technical report on GPT-4o, nor a systematic evaluation of affective speech produced by GPT-4o is available.

Considering the development towards releasing large pretrained models in the NLP and CV areas, we assume that in the near future powerful speech synthesis models will also be made publicly available and, consequently, investigated more thoroughly, allowing us to carry out according experiments to the above for vision and linguistics. In addition, we expect a continuing trend toward truly multimodal FMs that may not just take in but also produce natural speech with controllable prosodic properties. Several multimodal models that also produce audio output data have been proposed[50–52]. To the best of our knowledge, none of them exhibit emotional speech synthesis capabilities.

We proceed to discuss the *analysis* capabilities of current speech-based FMs. A system capable of analysing arbitrary affective properties of speech data without any tuning must ingest both audio and text inputs. Several FM approaches fulfilling this requirement have been proposed. However, the vast majority of them are not pretrained on speech data at all.

Examples include AnyMAL[53], X-InstructBLIP[54], and ModaVerse[55]. In only a few models, speech is part of the pretraining data. QWEN-Audio's training data comprises several labelled speech datasets, including emotionality already. Hence, QWEN-Audio[56] in the proposed form is not a candidate for exploring 'emergent' affective recognition capabilities. X-LLM[57] processes video, text, and audio inputs and is explicitly designed to process speech. The authors, however, do not report any experiments related to predicting affect in speech. As of now, the pretrained X-LLM model is not publicly available, hence, unfortunately again, not allowing us to carry out experiments analogous to the vision and the linguistic ones.

Similar to the affective speech synthesis problem, near-future multimodal FMs can be expected to be capable of analysing affective speech in a zero-shot manner, even if not explicitly pretrained for this task. As of now, however, we are not aware of any such model.

The evaluation is changing

One of the reasons to understand the impressive performance of currently available FMs is that they use massive amounts of data from "the Internet" for training. Nevertheless, the indiscriminate use of data poses the following challenge to the scientific community: can we guarantee that the data we feed to these models for their evaluation has not been used for training? In case of a negative answer, how fair and representative of the model capabilities can standard evaluation metrics be? Although we do not have a concrete answer yet, we hope these challenges engage the research community into looking for methods and metrics that allow a proper scientific evaluation of these

emerging FMs in the field of Affective Computing. As is, the current state of such models in Affective Computing may resemble a shell game: many different tools and approaches are shuffled and mixed until some partially less, partially more convincing performances are obtained. Especially because it is the popular 'Big Six' Ekman emotions we considered herein, chances are high that the models only reverberate with what they have already seen. Testing on more subtle models such as the dimensional approach or less considered affective states is therefore urgently needed.

Discussion

We analysed the affective capabilities of currently available FMs exploiting the vision, the linguistics, and the speech (acoustic) modalities. While the affective generation and analysis capabilities of the vision- and the linguistic-based FMs are plausible, the affective generation and analysis of speech-based FMs is not yet mature enough. Nonetheless, it is reasonable to imagine a not-too-distant future where this technology achieves similar results as with the other two modalities. Despite not being currently available, we also envision physiological-based FMs to be developed and explored in the near future.

The obtained results support the utilisation of FMs for not only creating, but also analysing affective data. The former could simplify the collection of affective data, reducing time and costs. Nevertheless, the utilisation of such synthetic datasets might not always be suitable, as—for example—the emotional variability in the generated data might be limited. Intuitively, it seems reasonable to think that the generated data convey a sort of prototypical, canonical representation of the prompted emotion; hence, reducing the richness and losing the nuances in the way how humans express and show emotions. Affective models trained exclusively with synthetic data might underperform in real-world, unconstrained scenarios because of this lack of diversity in the training data. The analysis capabilities of current FMs could simplify the deployment of affective models, providing off-the-shelf solutions. Despite the advantages of such one-size-fits-all solutions, an open question—among others—is whether such models would be able to capture the cultural emotional nuances in conveying emotions, which might limit their deployment and usage.

One of the main outcomes of this work is the collection of two synthetic affective corpora generated with FMs—one containing facial images and the other textual sentences, which will be publicly available. The models training and the analyses reported herein were performed assuming that the synthetically generated instances conveyed the prompted emotions. We acknowledge this could not always be the case. To overcome this limitation, we plan to run a data collection with human annotators to annotate the generated samples, assessing the affective capabilities of the selected FMs from a human perspective. The human annotations will also allow conducting further experiments to assess biases in the generated images. These could be centred on analysing the emotions conveyed by the synthetic images for each one of the 18 demographic groups included in the generated dataset, considering age, sex, and skin tone. Concerning the analysis capabilities of current FMs, future works could verify the generalisability of our findings exploiting other already-existing affective datasets.

Data availability

For further information and access to the dataset, the interested readers are kindly asked to contact the authors.

Received: 29 October 2024; Accepted: 28 November 2025;
Published online: 31 January 2026

References

1. Ding, H., Sricharan, K. & Chellappa, R. ExprGAN: facial expression editing with controllable expression intensity. In *Proceedings of the 32nd Conference on Artificial Intelligence*, 6781–6788 (AAAI, 2018).
2. Karras, T. et al. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 8110–8119 (CVF, 2020).
3. Balakrishnan, G., Xiong, Y., Xia, W. & Perona, P. Towards causal benchmarking of bias in face analysis algorithms. In *Deep Learning-Based Face Analytics. Advances in Computer Vision and Pattern Recognition* 327–359 (ECCV, 2021).
4. Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P. & Scherer, S. Affect-LM: a neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 634–643 (ACL, 2017).
5. Zhou, K., Sisman, B., Rana, R., Schuller, B. W. & Li, H. Speech synthesis with mixed emotions. *IEEE Trans. Affect. Comput.* **14**, 3120–3134 (2023).
6. Wang, Y. et al. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, 5180–5189 (PMLR, 2018).
7. Zhang, Z., Han, J., Qian, K. & Schuller, B. Evolving learning for analysing mood-related infant vocalisation. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, 142–146 (ISCA, 2018).
8. Rajapakshe, T. et al. emoDARTS: joint optimisation of CNN & sequential neural network architectures for superior speech emotion recognition. *IEEE Access* **12**, 110492–110503 (2024).
9. Lucey, P. et al. The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. In *Workshop Proceedings of the Conference on Computer Vision and Pattern Recognition*, 94–101 (IEEE, 2010).
10. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B. A database of german emotional speech. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association*, 1517–1520 (ISCA, 2005).
11. Mollahosseini, A., Hasani, B. & Mahoor, M. H. AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**, 18–31 (2019).
12. Kossaifi, J. et al. SEWA DB: a rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1022–1040 (2021).
13. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, 27730–27744 (NIPS, 2022).
14. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
15. Wang, C. et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE/ACM Trans. Audio Speech Lang. Process*. https://doi.org/10.1109/TASLPRO.2025.3530270 (2025).
16. Podell, D. et al. SDXL: improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, (ICLR, 2024).
17. Schaeffer, R., Miranda, B. & Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage? In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* (NIPS, 2023).
18. Ekman, P. & Friesen, W. V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **17**, 124–129 (1971).
19. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763 (PMLR, 2021).
20. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems* (NIPS, 2020).
21. Rombach, R. et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 10684–10695 (CVF, 2022).

22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. Preprint at https://arxiv.org/abs/2204.06125 (2022).
23. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution* (American Psychological Association, 2003).
24. Baltrušaitis, T., Robinson, P. & Morency, L. OpenFace: an open source facial behavior analysis toolkit. In *Proceedings of the Winter Conference on Applications of Computer Vision* (IEEE, 2016).
25. Ekman, P. & Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement* (Consulting Psychologist Press, 1978).
26. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations* (ICLR,Virtual Conference, 2021).
27. Goodfellow, I. J. et al. Challenges in representation learning: a report on three machine learning contests. In *Proceedings of the International Conference on Neural Information Processing*, 117–124 (Springer, 2013).
28. Peña, A., Morales, A., Serna, I., Fierrez, J. & Lapedriza, A. Facial expressions as a vulnerability in face recognition. In *Proceedings of the International Conference on Image Processing*, 2988–2992 (IEEE, 2021).
29. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* (NIPS, 2023).
30. Liu, H., Li, C., Li, Y. & Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 26296–26306 (CVF, 2024).
31. Vaswani, A. et al. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems* (NIPS, 2017).
32. Radford, A. et al. Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised (2018).
33. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
34. Li, C. et al. Large language models understand and can be enhanced by emotional stimuli. Preprint at https://doi.org/10.48550/arXiv.2307.11760 (2023).
35. Broekens, J. et al. Fine-grained affective processing capabilities emerging from large language models. In *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction* (IEEE, 2023).
36. Amin, M., Cambria, E. & Schuller, B. W. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intell. Syst.* **38**, 15–23 (2023).
37. Amin, M. M., Mao, R., Cambria, E. & Schuller, B. W. A wide evaluation of ChatGPT on affective computing tasks. In *IEEE Transactions on Affective Computing* (IEEE, 2024).
38. Wang, X., Li, X., Yin, Z., Wu, Y. & Liu, J. Emotional intelligence of large language models. *J. Pac. Rim Psychol*. **17**, 18344909231213958 (2023).
39. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
40. Jiang, A. Q. et al. Mistral 7B. Preprint at https://arxiv.org/abs/2310.06825 (2023).
41. Jiang, A. Q. et al. Mixtral of experts. Preprint at https://arxiv.org/abs/2401.04088 (2024).
42. Fedus, W., Zoph, B. & Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23**, 1–39 (2022).
43. Demszky, D. et al. GoEmotions: a dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054 (ACL, 2022).
44. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
45. Amin, M. M. & Schuller, B. W. On prompt sensitivity of ChatGPT in affective computing. In *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction* (IEEE, 2024).
46. Zheng, L. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* (NIPS, 2023).
47. Triantafyllopoulos, A. et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proc. IEEE* **111**, 1355–1381 (2023).
48. Yang, D. et al. UniAudio: an audio foundation model toward universal audio generation. Preprint at https://arxiv.org/abs/2310.00704 (2023).
49. Leng, Y. et al. PromptTTS 2: describing and generating voices with text prompt. In *Proceedings of the 12th International Conferenceon Learning Representations (ICLR)*, (ICLR, 2024).
50. Wu, J., Gan, W., Chen, Z., Wan, S. & Philip, S. Y. Multimodal large language models: a survey. In *Proceedings of the International Conference on Big Data*, 2247–2256 (IEEE, 2023).
51. Zhang, D. et al. MM-LLMs: recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 12401–12430 (ACL, 2024).
52. Triantafyllopoulos, A. et al. Computer audition: from task-specific machine learning to foundation models. In *Proceedings of the IEEE*, vol. 113, 317–343 (2025).
53. Moon, S. et al. AnyMAL: an efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conferenceon Empirical Methods in Natural Language Processing: Industry Track*, 1314–1332 (ACL, 2024).
54. Panagopoulou, A. et al. X-InstructBLIP: a framework for aligning image, 3D, audio, video to LLMs and its emergent cross-modal reasoning. In *Computer Vision – ECCV 2024*. *Lecture Notes in Computer Science*, vol. 15103, 177–197 (Springer, Cham, 2025).
55. Wang, X., Zhuang, B. & Wu, Q. ModaVerse: efficiently transforming modalities with LLMs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 26606–26616 (CVF, 2024).
56. Chu, Y. et al. Qwen-audio: advancing universal audio understanding via unified large-scale audio-language models. Preprint at https://arxiv.org/abs/2311.07919 (2023).
57. Chen, F. et al. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. Preprint at https://arxiv.org/abs/2305.04160 (2023).

Acknowledgements

This project has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), and from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101060660 (SHIFT).

Author contributions

B.S. conceptualised the study. A.M.-R. and A.P.A. focused on the vision modality, summarising the literature, creating the synthetic facial images, and running the experimental analyses; I.T., M.A., and A.S. contributed to the linguistic modality, summarising the literature, gathering and curating the neutral, baseline sentences, synthesising the affective phrases, and running the dedicated experiments; L.C. addressed the speech modality, investigating the state-of-the-art methods emphasising on their current possibilities and limitations. B.S., A.M.-R., and A.P.A. edited the manuscript for coherence and consistency. B.S. and S.A. coordinated and supervised the study. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Björn Schuller.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.