Using Nonverbal Cues in Empathic Multi-Modal LLM-Driven Chatbots for Mental Health Support

MATTHIAS SCHMIDMAIER, LMU Munich, Germany
JONATHAN RUPP, University of Innsbruck, Austria
CEDRIK HARRICH, LMU Munich, Germany
SVEN MAYER, LMU Munich, Germany and TU Dortmund University, Germany

Despite their popularity in providing digital mental health support, mobile conversational agents primarily rely on verbal input, which limits their ability to respond to emotional expressions. We therefore envision using the sensory equipment of today's devices to increase the nonverbal, empathic capabilities of chatbots. We initially validated that multi-modal LLMs (MLLM) can infer emotional expressions from facial expressions with high accuracy. In a user study (N=200), we then investigated the effects of such multi-modal input on response generation and perceived system empathy in emotional support scenarios. We found significant effects on cognitive and affective dimensions of linguistic expression in system responses, yet no significant increases in perceived empathy. Our research demonstrates the general potential of using nonverbal context to adapt LLM response behavior, providing input for future research on augmented interaction in empathic MLLM-based systems.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: human-computer interaction, LLM, multi-modal LLM, empathy, context awareness, nonverbal communciation, mental health

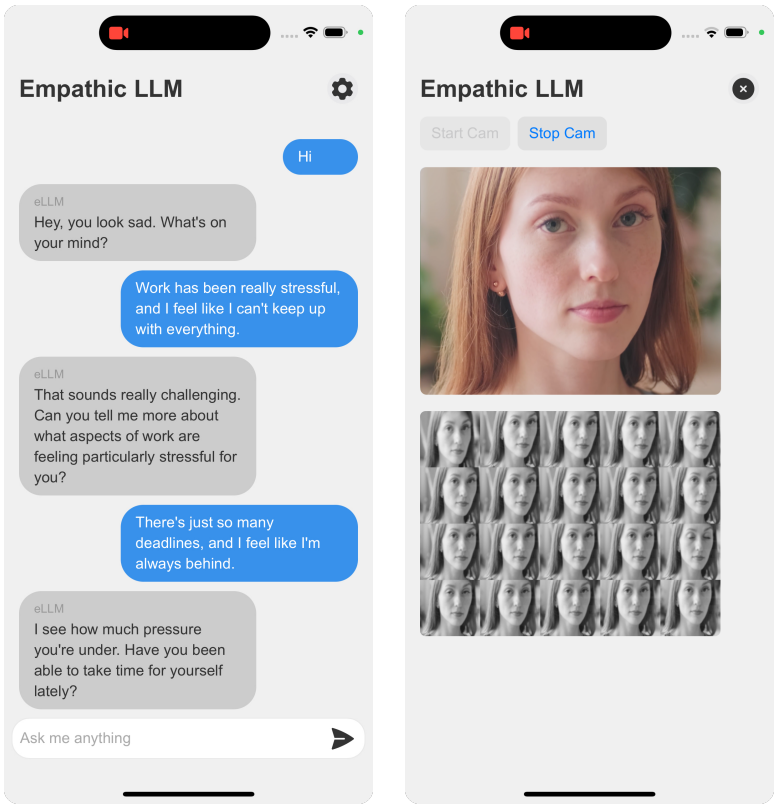
ACM Reference Format:
Matthias Schmidmaier, Jonathan Rupp, Cedrik Harrich, and Sven Mayer. 2025. Using Nonverbal Cues in Empathic Multi-Modal LLM-Driven Chatbots for Mental Health Support. *Proc. ACM Hum.-Comput. Interact.* 9, 5, Article MHCI039 (September 2025), 34 pages. https://doi.org/10.1145/3743724

1 Introduction

Mobile mental health chatbots have become increasingly popular in recent years [50, 52]. Especially the human-like conversational capabilities of Large Language Models (LLMs) allow to create empathic interaction [19, 29, 78, 93], as LLM-generated responses might be perceived as even more empathic than human responses [6, 39]. In combination with constant accessibility on mobile devices, this enables the creation of artificial companions that can provide personalized, anonymous, non-judgmental, emotional support [15, 19, 24, 77]. In this way, chatbots can address critical barriers to mental health support, such as perceived stigma and lack of mental health professionals [13]. However, the use of chatbots in the context of mental health poses certain risks and ethical concerns, for example, regarding quality of care, liability, accessibility, social isolation, or data privacy [8, 15, 19, 24]. Therefore, researchers recommend chatbots primarily for informal support

Authors' Contact Information: Matthias Schmidmaier, LMU Munich, Munich, Germany, matt@schmidmaier.org; Jonathan Rupp, University of Innsbruck, Innsbruck, Austria, jonathan.rupp@uibk.ac.at; Cedrik Harrich, LMU Munich, Munich, Germany, c.harrich@campus.lmu.de; Sven Mayer, LMU Munich, Munich, Germany and TU Dortmund University, Dortmund, Germany, info@sven-mayer.com.

ACM 2573-0142/2025/9-ARTMHCI039
https://doi.org/10.1145/3743724



(a) Chat interface. (b) Buffered camera feed.

Fig. 1. Our LLM-based chatbot (a) that uses nonverbal context descriptions as additional input. The setup view (b) shows the image buffer for MLLM-based facial expression recognition.

or to supplement human intervention and not as a substitute for professional therapy [19]. In both professional and informal scenarios, empathy was found to positively influence the outcome of emotional support [31, 38, 58, 62, 101]. A fundamental factor for empathic interaction is nonverbal communication [25, 31, 47, 87], as it allows to recognize and express emotions through facial expressions, for example [16, 35, 48, 98]. While related systems such as socially assistive robots utilize facial expression recognition (FER) to increase empathy, engagement and trust [21, 81, 85], or to improve therapeutic alliance [56], most chat applications for emotional support [50] rely mainly on textual input and neglect the nonverbal context. We address this gap, by exploring how visual input affects the empathic behavior of a multi-modal chatbot in emotional support scenarios. For that, we followed related research on Multi-Modal Generative Pre-trained Transformers, that we refer to as multi-modal LLMs or MLLMs, as they potentially offer robust and contextual interpretation of visual context in combination with emotion recognition [12, 41, 79, 88, 106] and empathic support [1, 22]. To validate the FER capabilities of MLLMs and investigate the effects on empathic chatbot behavior, we established the following research questions:

RQ1 How well can MLLMs infer affective states from facial expressions in video input?
RQ2 How does additional FER input affect empathic response generation of LLMs?
RQ3 How does the integration of FER into LLM-based chatbots affect perceived empathy?

We first evaluated FER with GPT-4o and GPT-4o mini based on facial blendshape descriptions and image-based time-series. We found that the more privacy-preserving input of blendshape descriptions resulted in a general over-prediction of expressions indicating happiness and a relatively low accuracy of up to 36 %, while image-based FER resulted in a higher accuracy of up to 87 % (RQ1). Based on these findings, we developed a multi-agent system that uses an MLLM to interpret facial expressions and provides nonverbal context for an LLM-based chatbot. Our user study (N=200) showed that this nonverbal context can affect the system response generation regarding cognitive and affective language expressions (RQ2). In addition, we found an upward trend in perceived empathy in nonverbal conditions, yet without statistical significance (RQ3).

In summary, our contributions are twofold. First, we demonstrate GPT-4's general visual capabilities for recognizing facial expressions from image and blendshapes input. Second, we show how visual nonverbal context can be integrated alongside text in multi-modal LLM applications and how it affects response generation. Furthermore, we present a lightweight prototype (Figure 1) to facilitate further research on mobile MLLM-based applications.

2 Related Work

This section introduces the basic concepts of empathy and nonverbal communication (NVC) and their role in clinical and mental healthcare. Building on this, we analyze the implementation, risks, and benefits of empathic chatbots in healthcare settings. Finally, we introduce current research on empathic capabilities of LLMs and multi-modal processing of nonverbal context in MLLMs.

2.1 Empathy and NVC

Empathy is an important concept in human interaction, particularly in the context of physiological and psychological health [55, 78, 82]. In general, empathy can be seen as a dimensional construct consisting of a cognitive and an affective component [11, 25, 28, 92]. While affective empathy primarily refers to emotional reactions, cognitive empathy can be defined as the prior perception and understanding of another person's situation or emotional state, for example, through perspective taking [25]. Specifically in mental healthcare, empathy is associated with therapeutic alliance and positive therapy outcome [31, 37, 38, 44, 53, 83, 101–103]. Still, related techniques such as active listening or perspective taking have proven to be a general basis for emotional support [17, 99], not only in professional but also in informal settings [15, 58].

One important aspect of empathic interaction is nonverbal communication, as it allows to perceive and express emotions, for example through facial expressions [16, 35, 48, 98]. Nonverbal communication can be defined as non-linguistic transmission of information over multiple channels that may occur with or without the intention and awareness of the sender or the receiver [5, 32, 36]. In the context of emotional support, nonverbal cues can help conveying sympathy, interest, or active listening [25], for example through turn-taking via "empathic continuers" [45] or attentive cues [34]. In therapist-client interaction, bodily communication can assist emotion interpretation and diagnostics [47, 87], lead to feelings of support and encourage the client to open up [31].

2.2 Empathic Chatbots in Healthcare

As the use of LLMs and chatbots in medical and mental health domains is more explored, research also increasingly investigates the role of empathy in these contexts [19, 39, 89, 93]. For example, in medical context, empathic chatbots are preferred over non-empathic ones for health advice [27, 68], and LLM-generated responses are perceived as more empathic than those from human healthcare providers [6, 69]. However, Webb [104] suggests that while empathic LLMs can effectively deliver bad news to patients, the absence of nonverbal communication remains a limitation. In digital mental health support, chatbots are already enjoying increasing popularity, especially as mobile

applications for informal emotional support [50, 52]. Such applications have been found to provide social and emotional support by showing interest and expressing empathy while encouraging self-disclosure through anonymity and non-judgmental behavior [15, 52, 63, 80]. For example, empathic chatbots can help reduce the symptoms of depression [46], or help coping with social exclusion, especially when responding to negative feelings [30]. Same as in medical context, recent research on chatbots in mental health follows a LLM-based approach [1, 22, 107], to create empathic chatbots that act as therapists ("doctor chatbot") or simulate patients in psychiatric scenarios [22]. However, researchers stress that due to ethical risks, chatbots should not replace human professionals, but rather serve as complement [19].

Risks and Ethical Concerns. Despite their advantages, the use of chatbots in digital mental health imposes several risks and ethical concerns [8, 15, 19, 24]. For example, data privacy, liability, and regulation need to be clarified, and access should be regulated to protect vulnerable populations, such as those with mental health disorders [8, 19, 24, 29]. Furthermore, since systems may provide incorrect advice, miss or misinterpret relevant context, such as emotional states, or have inherent biases, researchers emphasize potential risks in safety and quality of care [19, 24, 29]. In addition, emotional attachment and overreliance can increase vulnerability, as users may not recognize manipulated or malicious third-party content when delivered through trusted, familiar agents [15, 24]. The growing demand for mental health support and the limited access to traditional care require careful weighing of these concerns against potential benefits [13].

2.3 Empathic Capabilities of LLMs

In LLM-based systems such as introduced in Section 2.2, empathic behavior is mostly created through generalized pre-prompt instructions (e.g. "You are an empathetic chatbot. Respond to the user empathetically." [24]) or by instructing the model to process emotional expressions from user input [24, 39, 105, 112], and to generate emotional responses [64, 112]. In related studies, GPT-4 outperformed human responses in generating empathic healthcare responses, even with simple or varying pre-prompt instructions [69, 105]. Furthermore, Elyoseph et al. [39] showed that GPT is capable of emotionally understand described situations, indicating its use in psychiatric diagnosis and assessment. However, there are also shortcomings, such as LLMs showing empathy toward harmful political ideologies when they are not given specific instructions [24].

In contrast to text-only chatbots, multi-modal agents, such as social robots [21, 40, 56, 57, 64, 65, 72, 81, 85, 108] or virtual agents [54, 84, 86], can use multiple nonverbal channels to recognize affective states and express empathy. This multi-modal communication can increase perceived empathy [21], and, along with it, lead to improved engagement [21, 61], therapeutic alliance [56], trust [81, 85] or emotional support [40, 85, 108]. Especially facial expression recognition often serves as a basis for empathic resonance [7, 64, 65, 81, 85]. We applied this concept to MLLM-based chatbots with the aim of increasing empathic interaction.

2.4 Multi-Modal LLMs

While LLMs are mainly designed for text input, multi-modal LLMs allow processing other data types. For example, Pereira et al. [88] proposed an MLLM-based architecture that processes sentiment analysis from text, speech spectogram analysis from audio, and temporal emotion recognition based on sequential image frames from body and facial gestures. Dongre [33] proposed a concept for mental health support, where a CNN model pre-processes physiological data to predict stress levels and use that as additional input for an empathic LLM that was instructed to act like a trained psychologist [33]. Other research suggests a textual description of gestures [106] or speech-based emotion recognition [1] as input to improve nonverbal understanding and empathic responses.

Facial Expression Recognition with MLLMs. FER pipelines usually combine face detection, feature extraction, and the prediction of action units or discrete emotion categories [12]. Although deep FER models can already offer robustness against varying head posture, illumination, complex backgrounds, or occlusion, MLLMs offer the potential for a more holistic and contextual interpretation of visual scenes [12, 41, 60]. For example, they may create narrative descriptions of spatial relationships and recognized objects, to recognize emotional causes and mixed emotional signals such as sarcasm [12, 41]. In addition, the few-shot in-context learning capabilities of MLLMs allow to quickly adapt emotion recognition to new categories, without the need for large amounts of specifically labeled training data [3, 12, 41, 66].

To validate the FER performance of MLLMs, Nadeem et al. [79] used the FER2013 dataset [49] and found GPT-4 to reach a 55.8 % accuracy in assigning one of seven emotional categories to an image. However, they did not account for poor labeling quality and unbalanced categorical distribution in the FER2013 dataset [10]. Similarly, AlDahoul et al. [4] reported a medium accuracy of 49% for GPT-4o to estimate emotion categories from AffectNet's [76] test data set, which is known to be "heavily imbalanced" [76]. In their follow-up evaluation, AlDahoul et al. [4] used a balanced subset of AffectNet, revealing higher accuracies, but leaving the question of how the previously tested MLLM s perform on balanced data. In contrast, Mehra et al. [74] prompted LLMs to infer affective states not from visual but from numeric valence / arousal values derived from facial expressions. They found that LLMs struggled to derive discrete emotion categories yet demonstrated strong "capacity for free-text affective inference of facial expressions" [74].

3 Methodical Approach and Hypotheses

As described in Section 2, there is a growing trend toward empathic LLM-based systems, particularly in digital mental health [19, 24, 50, 77]. Since nonverbal communication plays an important role for empathic interaction in this context [29, 31, 93], we aimed to enhance the multi-modal capabilities of LLM-based chatbots. We focused on augmenting the multi-modal *input*, but not the nonverbal *output* modalities of the agent, as this would require additional comprehensive feedback design, especially when adding new output modalities to a text-based chatbot. Another methodological choice was to use MLLM-based FER. Based on Section 2.4, we hypothesized that this approach could provide a more comprehensive contextual description than a purely numerical emotion classification (RQ1). To investigate RQ2 and RQ3, we conducted a user study (N=200). While available applications for digital mental health support are typically designed for longitudinal use [50], informal support can also be situational, such as through digital peer support platforms [29]. We designed our study (Section 6) around such informal situational support scenarios, to evaluate basic functionality and potential limitations prior to longitudinal research.

Pre-Evaluation (RQ1). First, we investigated MLLM-based FER with OpenAI's GPT-4o and GPT-4o mini (Section 4). To address the shortcomings of similar evaluations [4, 79], we used balanced test data with improved labeling. Second, we evaluated an alternative input format for facial expressions to potentially improve resource requirements and privacy: in addition to visual input we tested numeric representation of facial features using Google Mediapipe's blendshape analysis. It provides 52 scores that describe the intensity of facial features such as "browDownLeft" or "mouthFrownRight" with values between 0 and 1. Following a research direction proposed by Wicke [106], we hypothesized that LLMs might be capable of interpreting such a representation of nonverbal cues. Third, based on Zhang et al. [110], we evaluated how well GPT-4 is able to interpret a contiguous series of frames, to create temporal context interpretations. As we used textual interpretations later on, we explored numeric as well as descriptive MLLM output. For this pre-evaluation we defined the following hypotheses:

H1a Using a balanced multi-label subset of the FER dataset shows significant increases in emotion recognition abilities of GPT-4 compared to previous studies with unbalanced data.
H1b Using numeric facial feature descriptions (blendshapes) for emotion recognition with GPT-4 yields performance comparable to direct visual input processing.
H1c GPT-4 is able to interpret the temporal dynamics of facial expressions by analyzing frame-wise time series.

Effects on System Response Style (RQ2). In a user study with 200 participants, we explored the effects of additional FER input on response generation (RQ2), following related research that analyzed conversations in the context of empathy [42, 67], empathic interfaces [90] and mental health agents [23, 70], using Linguistic Inquiry Word Count (LIWC) [14, 95]. Related studies mainly explored LIWC measures of *psychological processes* (affect, social processes, and cognition [42, 67, 67, 70]), *linguistic variables*, such as 1st-person pronouns [70, 90], and the categories *perception* and *physical* [42, 67]. Based on that, we defined hypotheses H2a and H2b, reflecting potential effects of affective and cognitive empathy:

H2a With additional FER input, the system responses show significantly higher rates of cognitive expressions reflecting the understanding and processing of facial cues.
H2b Additional FER input significantly changes the emotionality of the system responses.

Effects on System Perception (RQ3). Our basic assumption for RQ3 was that the additional information from multi-modal input leads to a more empathic system response style and thus to increased perceived empathy. Furthermore, our approach allowed us to trigger system reactions not only as response to user messages, but also as a direct response to nonverbal cues. As in the context of empathic interaction, proactive turn-taking could express interest and active listening [45], we assumed that proactive system responses could lead to higher perceived empathy. In addition, as awareness about sharing facial expressions can affect user behavior in computer-mediated communication [75], we explored perceived system empathy also in a placebo group, where facial expressions were captured, yet not processed. For evaluation, we applied the Perceived Empathy of Technology Scale (PETS) [92] and the following hypotheses:

H3a Additional FER input leads to increased perceived system empathy.
H3b Proactive turn-taking based on FER input leads to increased perceived empathy.
H3c Awareness about FER input has positive effects on perceived empathy (placebo effect).

4 Pre-Evaluation: FER with Multi-Modal LLMs

In this section, we present our preliminary evaluation for facial expression recognition with multi-modal LLMs, where we assessed the overall feasibility of our approach, explored relevant nonverbal dimensions, and tested prompts for temporal, numeric and textual FER interpretation.

4.1 Test Data: FER+1400

While the original FER2013 dataset [49] as used by Nadeem et al. [79] includes a single label for each image, Barsoum et al. [10] published the enhanced FER+, where each image is assigned ten ratings, to address the subjective and often not distinctive nature of facial expressions [9]. To get evenly distributed test data, we extracted a subset of the FER dataset, by selecting 200 images for each original FER category. To factor in the improved FER+ labeling, we drew the images from a FER category according to their highest FER+ rating for that category. In that way, we extracted a subset of 1400 images with robust labeling based on the original FER2013 dataset and the FER+ labels. We refer to that subset as FER+1400 dataset. All images in the dataset are black-and-white

Table 1. FER evaluation results comparing blendshapes and visual image input, each conducted two times with GPT-4o mini and GPT-4o on FER+1400.

Run	Input	Model	Acc.	Prec.	Recall	f1
1	image	GPT-4o mini	0.84	0.87	0.84	0.84
2	image	GPT-4o mini	0.84	0.88	0.84	0.83
1	image	GPT-4o	0.86	0.91	0.86	0.87
2	image	GPT-4o	0.87	0.91	0.87	0.88
1	blendsh.	GPT-4o mini	0.30	0.23	0.30	0.22
2	blendsh.	GPT-4o mini	0.30	0.30	0.30	0.22
1	blendsh.	GPT-4o	0.35	0.35	0.35	0.30
2	blendsh.	GPT-4o	0.36	0.29	0.36	0.30

and come in a resolution of $48x48$ pixels. A file containing the original FER2013 image names is included in the *Supplementary Material*.

4.2 FER from Images

Re-evaluating the approach by Nadeem et al. [79] we used our FER+1400 dataset and two GPT-4 versions (OpenAI's *gpt-4o-2024-05-13* and *gpt-4o-mini-2024-07-18*). All following evaluation runs and our study applications are based on these model versions. To promote response consistency, we followed related approaches [24, 39] and conducted multiple, repetitive evaluation runs. We executed two runs with GPT-4o mini and two runs with GPT-4o via OpenAI's API and batched processing. As input we used base64 encoded images from the dataset and the following prompt, which we designed based on the OpenAI guidelines, design recommendations by Vogel [100] and multiple test iterations.

Prompt: Image-based FER
You are a facial expression analyzer. Input: Image showing a face. Task: 1. Interpret the facial expression. 2. Estimate emotion intensities (0-1) based on the expression. Output JSON: {scores: {anger: float, disgust: float, fear: float, happiness: float, sadness: float, surprise: float, neutral: float}, description: string}

Results. To compare our MLLM-based multi-category labeling with the original FER single-category label, we selected the highest MLLM rating for each image. To address possible parity between multiple categories, we introduced an additional evaluation category "undecided". The categorical interpretation results are depicted in Table 1 and in Appendix A, with 84% accuracy for GPT-4o mini and 86 – 87% accuracy for GPT-4o in predicting the original FER label. The textual interpretations of both models were short and reflective, for example, "The person appears to be smiling genuinely, suggesting a high level of happiness." or "The man appears to exhibit a serious expression with some signs of concern or sadness. His brow is furrowed and mouth is tightened, indicating a more intense emotional state."

4.3 FER from Blendshapes

To evaluate the interpretation of blendshapes as input format, we again conducted four evaluation runs, two with GPT-4o mini and two with GPT-4o. As for the image-based evaluation, we created a prompt that instructs the LLM to provide a rating for each of the seven FER emotion categories, as well as a textual description that should provide reasoning insights.

Prompt: Blendshapes-based FER
You are a facial expression analyzer using Google MediaPipe blendshape data. Input: <faceBlendShapes> list of 52 blendshape scores (0-100). Task: 1. Interpret blendshapes to describe the facial expression. 2. Estimate emotion intensities (0-1) based on the expression. Output JSON: {scores: {anger: float, disgust: float, fear: float, happiness: float, sadness: float, surprise: float, neutral: float}, description: string}

Results. The blendshapes approach showed significantly lower accuracies than the image-based input (see Table 1), resulting in accuracies of 30% for GPT-4o mini and 35 – 36% for GPT-4o. Further, GPT-4o mini showed high confusion for sadness, especially regarding anger and disgust, and for neutral, regarding fear and surprise. GPT-4o showed slightly better performance, especially regarding the prediction of anger and surprise. Both models showed strong confidence in recognizing expressions of happiness.

4.4 Time Series Interpretation

While in the previous sections, we used data with categorical emotions, many datasets with continuous data also provide additional, conversation-related labels based on continuous head gestures and mimics, such as nodding or changing gaze direction [18, 26, 59, 73]. The *small MPI Facial Expressions Database* [26] contains videos with labels for emotional expressions (sad, surprise, happy, disgust) and for expressions of agreement, confusion, and thinking. We used these videos, which have a duration between one to six seconds, for a basic evaluation of MLLM-based time series interpretation. First, we conducted multiple test runs with arrays of blendshape set, based on videos from the *small MPI Facial Expressions Database* [26], and found that GPT-4o mini and GPT-4o are able to produce meaningful temporal interpretations. Subsequently, we tested the ability of interpreting image-based time series, using 54 videos from the dataset. For each, we created a single image, that shows multiple frames over time in form of a grid, allowing us to hand over a time series of expressions with a single request. To create that grid, we sampled frames at 5 fps, optimizing for real-time application. We discuss sampling rates in that context further in Section 8.5. For interpretation, we created the following prompt, containing the emotional categories from the previous FER evaluation, as well as "agreement" and "contemplation" based on the MPI labels. As we could not consider labels that are not in the MPI dataset (e.g. "anger") for statistical evaluation, we assessed them based on qualitative interpretation of the MLLM output.

Prompt: Image-based Time Series
You are an expert facial expression analyzer. The input is an image containing a grid of video frames arranged sequentially from top-left to bottom-right. The frames have been captured at a rate of 5 fps and show the changing facial expressions of a single individual over time, during a dyadic conversation. Your tasks: 1. Check the overall layout of the frame grid (columns and rows) 2. Analyze head movement and facial expressions in each frame and over time, to estimate if they express: agreement (e.g. expressed through nodding), contemplation (e.g. frowning or looking up), confusion, anger, disgust, fear, happiness, sadness, surprise and neutral. 3. Provide a short textual summary of your analysis and a score for each category (0-1). Output JSON: {rows: number, cols: number, analysis: string, scores: {agreement: float, contemplation: float, anger: float, disgust: float, fear: float, happiness: float, sadness: float, surprise: float, neutral: float}}

In contrast to the evaluation in Section 4.2, we conducted six runs - three with GPT-4o mini and three with GPT-4o - to account for the higher accuracy variance in the first four runs. We considered only categories contained in both - the MPI dataset and the prompt: "agreement," "contemplation," "disgust," "happiness," "sadness," and "surprise." As in the FER evaluation, we took only the highest prediction for comparison with the MPI label.

Results. Table 2 shows the accuracies for all runs, with GPT-4o performing better (53 – 58%) and slightly more stable than GPT-4o mini (36 – 44%). While the statistical results were not as promising as in the single image evaluation, especially the textual analysis revealed interesting insights into temporal understanding and contextual reasoning, for example: "There is a consistent return to a happy expression, suggesting a stable positive emotion throughout the sequence", or "The individual appears to exhibit facial expressions primarily associated with contemplation and

Table 2. Results of image-based temporal awareness runs with GPT-4o mini and GPT-4o on facial expression videos with highest category prediction for evaluation.

Run	Model	Acc.	Prec.	Recall	f1
1	GPT-4o mini	0.36	0.56	0.36	0.33
2	GPT-4o mini	0.42	0.59	0.42	0.37
3	GPT-4o mini	0.44	0.59	0.44	0.41
1	GPT-4o	0.56	0.64	0.56	0.55
2	GPT-4o	0.53	0.64	0.53	0.50
3	GPT-4o	0.58	0.65	0.58	0.56

some confusion. In each frame, their eyebrows are furrowed and there are subtle changes in mouth positioning, suggesting they are trying to understand or think deeply about something." While thoughtful states or "contemplation" are apparently well recognized by the MLLMs, "agreement'' seems to be more difficult to perceive.

4.5 Conclusion

The results of our pre-evaluation regarding image-based FER (Section 4.2) confirm H1a, showing accuracies of 84 – 87% compared to 49 – 59% in previous studies [4, 79] that used imbalanced data. In addition, qualitative observations revealed that the MLLMs were able to create meaningful textual interpretations of facial expressions. GPT-4 was also able to recognize occlusions and hand gestures: *"The person appears to be surprised with their mouth open and hand on the face."* We discuss that point further in Section 8.1. While the results of the second pre-evaluation (Section 4.3) indicate that the interpretation of blendshape information in general works, we have to reject H1b. We assume that one reason for the lower FER performance might be that Google Mediapipe's blendshapes only provide detailed descriptions for certain facial regions. Therefore, we still see a potential use in scenarios that benefit from focusing on certain regions like the mouth where blendshapes describe expressions like laughter or yawning. Regarding temporal interpretation (Section 4.4), taking into account the ambiguous, multi-faceted nature of expressions in the MPI videos, we argue, that the results demonstrate the feasibility of processing time series as a single-image frame grid through MLLMs, and therefore support H1c.

5 System

Based on our previous findings and the related approaches introduced in Section 2 we implemented a multi-agent MLLM-based chatbot application. Figure 2 depicts the system components: a browser-based frontend implemented in Vue.js and designed as a chat interface for text input and webcam recording, a Python-based backend for message handling and logging, and multiple MLLM instances for processing facial expressions and response generation.

5.1 Visual Input

We used Google Mediapipe for face detection in the frontend. The system captured the webcam's RGB video stream at 5 fps and cropped it to the detected face bounding box with an additional padding of 15 % in width and height, ensuring consistent image input for the LLM. At the same time this enhanced privacy, as we captured only limited surroundings. The captured images were buffered with a resolution of 120x160 pixels and continuously added to a single buffer grid image that depicted the images of the last four seconds. This grid image was initially filled from the top left and shifted correspondingly on every new incoming image. When no face was visible, a black

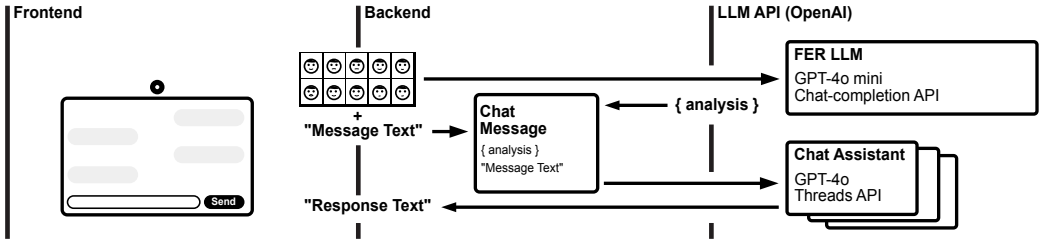


Fig. 2. System architecture. The web application captures text input and the webcam stream. For nonverbal interpretation, a frame grid of facial expressions is sent to a dedicated GPT-4o mini instance, then combined with the user message text to a input for one of multiple GPT-4o assistants.

rectangle was added to the buffer image. Sampling rate selection and buffer duration were based on our previous evaluation runs and the optimization for performance and real-time application, as further discussed in Section 8.5. The buffer image was continuously stored as base64 encoded data and sent to the backend based on conversation scenario and trigger events.

5.2 Multi-Modal Message Processing

We implemented a chat interface with a text-input field at the bottom and a messenger-like chat history visualization. User messages were sent to our backend via an API request. To create multi-modal messages with nonverbal visual context, we implemented two events:

@on_verbal: The current buffer image is sent together with each verbal text message at the time the user sends the message. It holds the facial expressions of the preceding four seconds, potentially providing insights into user reactions that, for example, emphasize or contradict the verbal input.

@on_inactivity: The current buffer image is sent without verbal message after 15 s of initial inactivity or after 60 s of inactivity since the last proactive reaction. This enables to proactively address the user during contemplative or inattentive phases. Preliminary tests determined suitable inactivity trigger times, with the second, longer interval preventing excessive system reactions.

For message processing, we set up different MLLM instances via the OpenAI API. For FER, we used OpenAI's chat completions API with GPT-4o mini. For the actual chat, we defined different GPT-4o based assistants. This multi-instance setup enabled effective processing and enhanced privacy by keeping the image data disentangled from the conversational data. Incoming user messages were checked for text and image content, processed with the different MLLM instances and than sent back to the frontend as combined message with *<nonverbalAnalysis>* and *<nonverbalScores>* sections. If an image was present but no face was detected, a *NO_PERSON_VISIBLE* flag was added to the analysis section. Messages without text content were flagged as *PROACTIVE*.

5.3 System Modes and Prompt Instructions

We implemented four different interaction modes (A, B, C, D), with prompts differing in processing instructions, yet not in the general role description (see Appendix B).

Mode A: Text-only. The default mode for a text-based chat experience, serving as baseline study condition, without camera access and visual input. All prompt instructions from this mode were also integrated into the other prompts.

Mode B: Text & placebo FER. Same as mode A, but required camera access, as the user instructions indicated that facial expressions would be used as additional input. We implemented this mode to control for a placebo effect.

Mode C: Text & FER. While mode B only simulated FER capabilities, in mode C the system actually processed the visual input sent together with the verbal text messages (*@on_verbal*).

Mode D: Text, FER & proactivity. Additional reactions to *@on_inactivity* enable proactive turn-taking, for example to promote the flow of conversation through "empathic continuers" [45].

Prompt Design. The system prompts for the FER instance and the four different LLM assistants can be found in Appendix B. Our general prompt design was based on related work (Section 2.3 and Section 2.2) which indicates that current LLMs may provide emotional understanding of text input and the ability to create empathic responses even without specific empathy-focused pre-prompting. Still, we instructed the system to act as empathic chatbot that should support the user in reflecting challenging interpersonal situations, and to deliver short, concise responses to enhance conversational flow. To specifically express cognitive empathy, as defined in Section 2.1, we further instructed the system to act understanding and emotionally responsive. To display affective empathy, we asked to include emotional reactions. In that way, we aimed to enable informal emotional support through empathic behavior, without instructing the agent to act like a professional therapist. However, future research on professional support could include more comprehensive prompting instructions. According to Foley and Gentile [47], nonverbal cues can serve as diagnostic indicators for conditions such as autism, attention deficit hyperactivity disorder, substance intoxication, or depression. The authors recommend observing whether nonverbal and verbal communication reflect consistent emotional states and tracking how these states evolve over the course of interactions [47]. However, implementation of such sophisticated diagnostic instructions would require thorough medical and ethical validation (see also Section 8.6).

5.4 Standalone Web App

In addition to our study application, we have developed a standalone web app (Figure 1) that allows researchers to further explore our concept, without setting up a backend structure. This Vue.js application features the same two-tier approach as described above, yet uses two OpenAI completion API endpoints for conversation and FER interpretation. The source code is available on https://github.com/kaiaka/mllm-chatbot.git.

6 User Study

We evaluated our empathic agent based on real user conversations in an online user study with N=200 participants. For this purpose, we defined two tasks, where participants were asked to share emotional experiences in a work context or with strangers, with the aim of obtaining emotional support and non-judgmental reflection.

6.1 Sample Size

For our study design, a priori power analysis with G*Power [43] suggested a sample size of $N = 180$ for a fixed effects ANOVA with a significance level of $p = 0.05$, statistical power of = 0.80 and a medium effect size ($f = .25$). We followed this recommendation and increased the sample size by 10% to N=200, to account for lower statistical power when testing non-normally distributed data. Further, our sample size exceeded the mean size of related studies for chatbots in mental health ($N = 75.2$ for 53 studies) [2], while also following HCI standards [20].

6.2 Participants

We recruited participants through Prolific and analyzed data from 200 valid sessions after excluding submissions with failed attention checks, camera access issues, or lack of input. The mean age of participants was 34.1 years ($SD = 10.8$), with 94 identifying as female and 106 identifying as male.

We recruited participants from 21 countries, with a majority of 113 participants residing in the European Economic Area, 52 in Southern Africa, 23 in North America, seven in the Asia-Pacific region, two in India, and one each in East Africa, South America and the Middle East. All participants stated to be fluent in English, which we confirmed qualitatively by reviewing the conversation logs.

6.3 Procedure

Our system was securely hosted online and allowed access from personal devices. Participation was voluntary and could be terminated at any time. Each participant was randomly assigned to one of four groups (A, B, C, D), reflecting the interaction modes in Section 5.3. After providing details about the procedure and data processing, we asked participants for consent. Next, participants in groups B, C, and D were asked for webcam access and received instructions on camera orientation and a preview of the face capturing. Then, all participants completed the two randomized tasks, each followed by a survey on system perception.

Tasks. While the instruction for both tasks overall was the same: "Your task is to engage in a 5-10 minutes conversation (in English) with our AI assistant about a challenging interpersonal situation you've experienced …", we created variance by asking to talk about experiences "…at work or university" (task 1) and "…with a stranger" (task 2). We also described how to start the conversation, the supportive goal of the dialogue, and that the system is able to recognize and react to nonverbal facial expressions (for groups B, C, D). We further explained that continuing to the next step was possible at the earliest five minutes after the interaction started or automatically after ten minutes. The complete task instructions can be found in the *Supplementary Material*.

Survey. After each task, participants rated 14 items in an integrated survey view. To assess perceived empathy, we applied the ten PETS items as recommended, as 101-point interactive sliders from *strongly disagree* to *strongly agree* in randomized order [92]. In addition, we asked participants to rate the perceived level and quality of nonverbal recognition ("The system clearly recognized my facial expressions." and "My facial expressions influenced the system's responses."), as well as their experienced emotional activation when talking about their experiences ("The conversation triggered very strong emotions in me."). Finally, we added an attention check, asking participants to drag a slider all the way to the left. For these four items, we also used 101-point sliders.

Completion. On average, participants took 20.1 minutes ($SD = 5.2$) to complete a procedure, for which they received 3.0£ in compensation. We observed ongoing sessions through a dedicated admin view, where we could follow the conversation and the task progress. After each session, we reviewed the recorded data to ensure that participants engaged, passed the attention check, activated their camera if required, and stayed on topic.

6.4 Privacy and Ethical Considerations

As described in Section 2.2, ethical concerns play a crucial role in HCI research on mental health applications [19, 24, 29, 91]. Therefore, we carefully considered ethical aspects to ensure the consent, safety, privacy, and well-being of the participants. We obtained approval from our institutional ethics committee, which confirmed that our research methodology complied with established ethical guidelines and standards. To generate responses, we used OpenAI's GPT-4 models, which are fine-tuned to handle a wide range of sensitive topics in an ethical way, at least by withdrawing from the conversation and suggesting to seek professional help [24]. To ensure transparency, we provided participants with detailed information about the study's purpose and procedures (see Section 6.3). As described in Section 5.3, we instructed the LLM assistants to be emotionally supportive and to practice active listening, but not to be therapeutic. Prior studies on chatbots in mental health

Table 3. Perceived empathy, nonverbal communication, conversation metrics and FER results as measured via ratings (0..100) and derived from input data (e.g. number of nonverbal messages). Data were averaged over both conversation tasks for each user (N=200) and analyzed with Kruskal-Wallis tests.

	Group A		Group B		Group C		Group D		Statistics			
	Mdn	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**H**	**df**	**p**	
Perceived Empathy												
PETS [0..100]	79.2	67.0-89.3	78.6	65.1-86.7	82.5	72.0-89.5	82.2	73.1-89.0	1.72	3	.632	
PETS-ER [0..100]	77.2	64.9-88.6	79.2	65.8-87.1	81.0	70.1-89.0	80.3	69.2-88.6	0.81	3	.847	
PETS-UT [0..100]	78.2	68.7-90.8	78.1	67.0-86.0	82.9	74.5-90.2	82.4	72.0-90.6	3.55	3	.314	
Perceived NVC												
NV Level [0..100]	5.5	0.5-48.4	53.8	34.6-75.4	61.0	45.0-73.5	69.2	56.8-83.6	49.53	3	.000	***
NV Effect [0..100]	3	.0-44.1	49.0	34.4-69.1	57.8	40.9-74.9	65.0	50.8-76.8	48.05	3	.000	***
User Input												
Messages [num]	8.0	6.1-9.9	8.0	6.5-10.4	7.0	5.5-8.9	7.0	5.5-8.9	2.31	3	.510	
NV Messages [num]			8.0	6.5-10.1	7.0	5.5-8.5	9.5	7.5-11.5	2.31	3	.510	
Task Duration [min]	8.6	6.9-10.0	8.5	6.8-10.9	8.7	7.4-10.7	8.2	6.9-9.4	1.50	3	.683	
Emo. Level [0..100]	52.8	32.1-74.9	58.2	25.9-76.1	62.0	42.8-76.0	61.2	47.9-72.0	2.31	3	.510	
FER Frequencies												
Agreement [%]			60.7	49.3-71.0	54.5	36.6-73.9	50.0	28.6-63.2	7.38	2	.025	*
Contemplation [%]			100.0	84.9-100.0	95.8	86.2-100.0	94.4	86.4-100.0	0.71	2	.701	
Anger [%]			11.1	4.5-18.2	5.9	0.0-12.1	3.3	0.0-17.6	4.77	2	.092	
Disgust [%]			5.1	0.0-7.8	0.0	0.0-5.6	0.0	0.0-8.0	4.43	2	.109	
Fear [%]			6.5	0.0-12.6	4.5	0.0-12.1	4.0	0.0-13.0	1.04	2	.594	
Happiness [%]			17.4	9.8-32.7	20.0	10.0-33.3	16.0	7.7-31.6	1.16	2	.561	
Sadness [%]			69.8	51.8-76.0	54.5	33.3-77.3	60.0	36.4-77.3	3.02	2	.221	
Surprise [%]			0.0	0.0-5.7	0.0	0.0-0.0	0.0	0.0-6.7	5.08	2	.079	
Neutral [%]			95.1	87.4-100.0	100.0	90.0-100.0	88.0	75.0-96.3	16.38	2	.000	***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

targeted a specific, often sensitive conversational topcis, such as anxieties, depression, social exclusion or specific mental disorders [2, 15, 30, 63, 77, 80]. In contrast, we focused on everyday interpersonal situations at work, university, and with strangers and avoided addressing specific emotional or psychological challenges. Although such situations might be emotionally challenging, this open-ended method allowed participants to share only what they were comfortable discussing. We conducted real-time monitoring to quickly address potential issues through our research team, while participants were also able to communicate with us via the Prolific study interface. Neither the real-time monitoring, nor the post-hoc review of the study chat histories did indicate a need for intervention. As the use of LLM-based systems raises privacy concerns, especially in our study context [15, 24, 29], we emphasized transparent disclosure of study information and data processing. Furthermore, we implemented a two-tier pipeline to separate the processing of conversational and image data, and limited the captured video region.

7 Results

For evaluation, we averaged the data from both conversation tasks per user. Due to partial non-normal distribution of our data (Shapiro-Wilk $p < 0.05$), we applied non-parametric tests and reported medians with IQR for statistical analysis. The following results provide insights into conversational engagement and system perception of participants, and into how the multi-modal input affected the LLM's response style. All analysis results can be found in the *Supplement*.

7.1 Conversation Metrics

Overall task completion was median 8.4 minutes ($IQR = 8.9 - 10.4$) and contained 7.5 user messages ($IQR = 6.0 - 9.5$) with subsequent system responses. As shown in Table 3, neither average task

duration, nor average number of user messages differed significantly between groups. Also the level of emotional activation (median 52.8 – 62.0) as self-assessed by the participants (Section 6.3) was rather moderate and not significantly different between groups. For additional qualitative analysis, we reviewed all conversations. In the first task, participants mostly talked about communication problems at work, conflicts with colleagues, unfair workload distribution, or situations that caused anxiety, feelings of not being valued, being overlooked, or being burnt out. In the second task, the conversations revolved mostly around negative encounters with strangers in public spaces such as public transport or supermarkets, bad experiences with customer support or anonymous arguments on social media. Overall, the messages suggested that participants appreciated the possibility to openly discuss situations and get feedback or emotional support by the system. Out of 200 participants, 102 explicitly thanked the system during the study, for example: "That's been very helpful, to talk that through. Thanks.", "Thanks, all the best!", "Thank you so much I feel better already", "No thank you. I really appreciate you <3".

7.2 Response Generation

In total, our system processed 3135 messages, of which 2369 contained nonverbal image data, each depicting a time series of up to 20 frames. From these nonverbal messages, 240 have been generated as proactive reaction in study condition D. The FER MLLM returned a textual interpretation for each image, as well as numeric intensity scores for different emotion categories. Table 3 lists the median frequencies of emotions detected by the MLLM in groups B, C and D (see also Appendix C). We defined an emotion as present in a multi-modal message when the numeric FER analysis returned a score ≥ 0.01, meaning that also multiple emotions could be counted as present in an image. The FER categories with the highest frequencies were contemplation ($Mdn = 94.4\% - 100.0\%$), neutral ($Mdn = 88.0\% - 100.0\%$), sadness ($Mdn = 60.0\% - 69.8\%$) and agreement ($Mdn = 50.0\% - 60.7\%$). Statistically significant differences were only found for agreement (A vs. C) and neutral (A vs. C). These results suggest that facial expression overall were mainly recognized as neutral or as reflecting cues of contemplation or agreement, and sadness. They further support our assumptions on the system's FER abilities as hypothesized in H1a and H1c.

7.3 Response Style

As described in Section 3, we applied Linguistic Inquiry and Word Count (LIWC) to analyze the effects of multi-modal input on the response generation. To test for statistically relevant differences between groups, again, we conducted Kruskal-Wallis tests, followed by Holm-Bonferroni corrected Dunn's post-hoc tests. The detailed statistics (median, IQR, significances) can be found in Appendix D (Table 5, Table 6, Table 7). Except for summary measures (e.g. word count), LIWC scores reflect the percentages of total words of a category within the analyzed text [14].

LIWC Analysis for H2a. To evaluate H2a, we specifically examined visual perception (Table 7) and psychological measures of cognition (Table 6). Figure 3 visualizes the medians of these LIWC measures and the significant differences between groups. While median percentages for perception were consistent across all groups ($\sim$ 11%), visual perception showed significant progressive increases from groups A and B to group C and further to group D, indicating a varying reliance on visual processing and potentially reflecting the multi-modal input in groups C and D. Medians for overall cognition measures (31.0% – 32.5%) and specifically for cognitive processes (21.4% – 22.9%), were high between all groups, indicating substantial expressions of cognition in system responses. Also the median percentages of insight (7.8% – 8.5%) may demonstrate expressions of awareness and understanding in all groups. Dichotomous thinking expressions ("all or none") in system responses were significantly reduced in groups C and D compared to groups and B, possibly suggesting

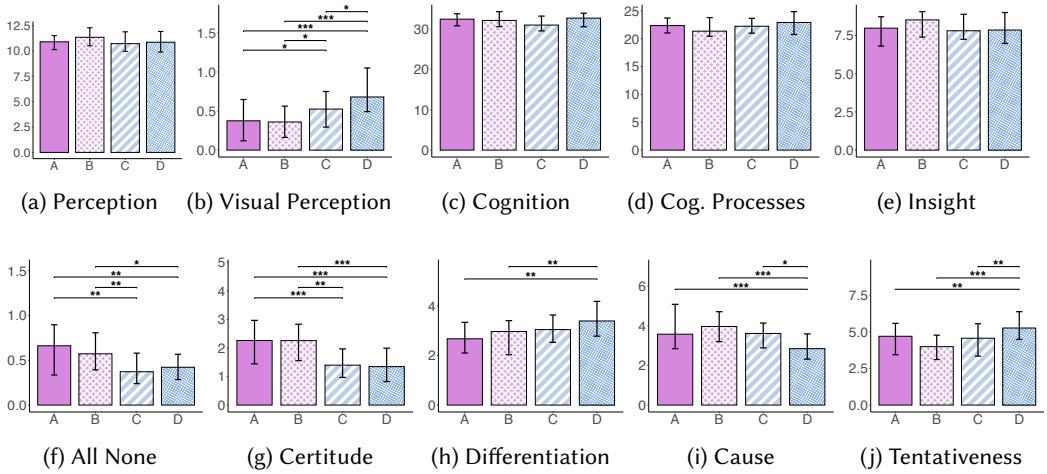


Fig. 3. LIWC categories for testing effects of FER on **cognitive expressions** in system responses (H2a). The bars show median and quartiles of the percentage frequency per group as well as significant differences between the groups as determined by post-hoc tests (Dunn with Bonferroni-Holm correction), marked with brackets ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

more nuanced cognitive processing under multi-modal conditions. Groups C and D also showed significantly lower median certitude compared to groups A and B , potentially reflecting a more careful interpretation and less definitive claims in FER based responses. In addition, groups C and D showed more differentiation language (significant only for Group D), again possibly indicating more nuanced thinking and higher recognition of differences. Expressions of causal relationships and tentativeness were only significantly different between group D and the other groups. Memory and discrepancy measures did not show any statistically relevant differences. We conclude that FER input significantly affected the system's linguistic style regarding visual perception and expressions that reflect certain aspects of cognitive processing. We therefore confirm hypothesis H2a.

LIWC Analysis for H2b. To evaluate the effects on affective expressions in system responses (H2b), we examined the affective LIWC categories as shown in Figure 4. Detailed statistics are described in Table 6. Compared to groups A and B, the FER-based responses in groups C and D expressed significantly reduced medians in overall affect, negative tone and overall emotions. In addition, language that expressed negative emotions and anxiety was significantly reduced in group D compared to A and B, and in group C compared to A, while for sadness and anger a similar trend was found, yet without significant differences. The significant differences between groups A/B and C/D, specifically regarding negative tone and emotions indicate that the multi-modal input affects emotionality in the system response style. These differences may reflect the mainly neutral FER frequencies as described in Section 7.2. Under this presumption, we accept H2b, as further discussed in Section 8.2.

Further LIWC Measures. Group D showed a significantly increased word count ($Mdn = 339.8$) compared to the other groups A ($Mdn = 265.0$), B ($Mdn = 269.5$) and C ($Mdn = 268.2$), reflecting the additional number of proactive messages. In addition, group C showed an 11% – 15% increase in words per sentence over all other groups, and also a higher analytical score, though only statistically significant compared to group B, possibly reflecting more formal, logical response style in group C. Further, system responses in group A showed a lower authenticity score than in other groups,

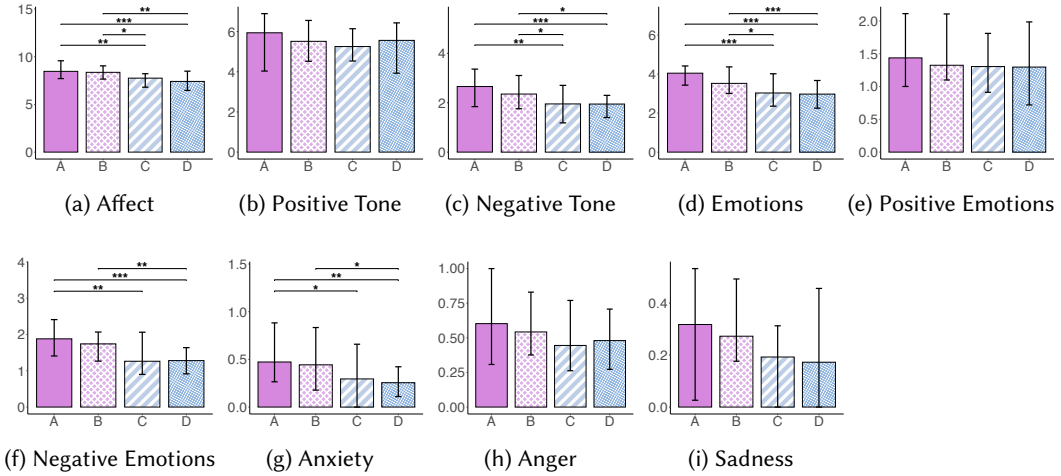


Fig. 4. LIWC variables for testing effects of FER on **affective expressions** in system responses (H2b). The bars show median and quartiles of the percentage frequency per group as well as significant differences between the groups as determined by post-hoc tests (Dunn with Bonferroni-Holm correction), marked with brackets ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

though only statistically different compared to group B. For linguistic measures, we only found a statistically relevant increased frequency of first-person plural pronouns in group D compared to the other groups. The low frequency of first-person pronouns (0.0% – 1.0%) and the comparably higher frequency of second-person pronouns (8.4% – 8.6%) might reflect an empathic, user-focused response style [90], as intended through our general system design.

7.4 System Perception

Kruskal-Wallis tests did show no statistically relevant differences between groups for PETS (median scores 79.2 – 82.2) and its subscales (Table 3). Although PETS scores for group C and D were higher compared to A and B, we therefore reject hypotheses H3a, H3b, and H3c. Regarding participants' perception of how well the system recognized nonverbal facial expressions (NVC Level) and how strong that expressions affected the system's responses (NVC Effect), post-hoc tests revealed significant decreased ratings between group A and all other groups, and between groups B and D. This effect could be expected for group A, as it did not provide any indication for FER. The decreased ratings for group B compared to group D may indicate a slight placebo effect regarding the perceived nonverbal abilities of the system. However, this does not apply to the comparison of groups B and C and is not reflected in the PETS scores, thus not affecting H3c.

8 Discussion

In the following, we discuss the results of our pre-evaluation (RQ1) and the user study (RQ2, RQ3), multi-modal FER, and limitations and future applications of our approach.

8.1 Interpretation of Facial Expression (RQ1)

In the pre-evaluation (Section 4), we found that GPT-4-based FER delivered relatively consistent results, confirming H1a and H1c. However, the contextual integration of numeric or categorical emotion analysis into an LLM response might require a deeper understanding of the prior interpretation process. We therefore hypothesize that LLMs might benefit from a textual description of

visual cues. As mentioned in Section 4.5, although we had to reject H1b, we still see potential for the use of more abstract data formats such as blendshapes if the use-case meets the specifications. Overall, we align with similar findings described by Bian et al. [12] and suggest that MLLMs have the ability to infer affective states from visual input (RQ1).

Contextual Interpretation. Pre-evaluation indicated that textual output allows to easily describe various visual features and gestures if visible. For example, hands interacting with head or face: "The hands on the head imply a feeling of distress or shock.", "The hands near the face suggest a response to something unexpected." or "…with a focused gaze and a hand near the nose, which may indicate contemplation or concentration." This demonstrates the potential of MLLMs for interpreting complex visual context (see also Section 2.4) and opens up new possibilities for comprehensive nonverbal analysis without the need for translation into numeric intensities. Textual translation of nonverbal cues [106], would further allow to easily augment LLM input with other modalities such as audio cues. Future work could further explore cue-to-text translation approaches to improve the quality of interpretation.

Misinterpretations. Potential failures in FER could lead to misinterpretation and consequently inappropriate system responses. Although errors can also occur in human interaction, for example, when therapists misinterpret their clients' expressions or project incorrect emotional states, future implementations could maintain transparency about the system's interpretive uncertainty, similar to how experienced therapists often verify their perceptions through clarifying questions. In this study, we did not specifically validate the FER output, as this would require a laboratory setting or detailed real-time or retrospective self-assessments of participants, which in turn would have contradicted the "in-the-wild" nature of our task design. To evaluate the accuracy of complex contextual MLLM-based FER, we envision future studies that explicitly compare generated output with professional assessments of the recorded or observed interactions.

8.2 Effects on System Response (RQ2)

Reflecting the dimensional definition of empathy, we investigated both cognitive and affective language expressions in system responses. LIWC analysis revealed a general high level of cognitive expression, and several significant changes in response patterns when FER input was introduced (Section 7.3). Most notably, we found a significant increase in expressions of visual perception for groups C and D (Figure 4). This possibly indicates the effect of visual input on system responses, where the system explicitly acknowledges its ability to perceive the user visually. The results also showed significant shifts in language patterns that reflect deeper cognitive processing, such as dichotomous thinking expressions and expressions of certitude. We conclude that additional FER input significantly influenced the system's language style, particularly regarding expressions of visual perception and cognitive processing that reflect more nuanced thinking.

The analysis of affective LIWC categories revealed distinct patterns between conditions (Figure 4). Groups C and D demonstrated significantly reduced overall affect, negative tone, and emotional expression compared to groups A and B. Further, we found partially significant differences for negative emotions and anxiety, as well as nonsignificant trends for anger and sadness. These results suggest that FER context may influence emotional expression in system responses. Combined with our instructions to "reflect and respond to the emotions expressed in messages and nonverbal cues" (Appendix B), the predominantly neutral facial expressions captured possibly produced more emotionally restrained system language, potentially reducing perceived empathy. Future research could explore prompt strategies that respond to neutral expressions with a change in conversational flow or motivation for emotional disclosure. This could maintain the user's attention to their facial expressions and encourage their emotional engagement. Furthermore, future systems

could substitute bodily nonverbal cues that are common in emotional support or therapist scenarios, such as attentive cues like eye contact or trunk lean [34], through textual cues [45]. In addition, non-embodied agents could also incorporate visual indicators, such as attention signaling, to compensate for the lack of physical nonverbal communication cues.

8.3 Perceived System Empathy (RQ3)

The baseline condition (group A) resulted in a median PETS score of 79.2 (Table 3), demonstrating strong inherent empathic capabilities of our LLM-based system with basic prompting and no nonverbal input. This aligns with prior research showing high emotional awareness and empathic performance of current LLMs in healthcare contexts [6, 24, 39, 93]. Although we observed increased PETS scores in groups C ($Mdn = 82.5$) and D ($Mdn = 82.2$) compared to groups A and B ($Mdn = 78.6$), the nonsignificant results led to the rejection of H3a, H3b and H3c (RQ3). One reason for that effect might be a limited emotional activation of participants, as discussed in Section 8.4, and a resulting low level of emotional expressivity in system responses as discussed in Section 8.2.

However, the differences in PETS scores between the multi-modal groups (C, D) and the baseline groups (A, B) may still indicate potential for our approach, especially when considering that no placebo effect was found for perceived empathy. We hypothesize, that empathic abilities of a system could be further emphasized in two ways. First, by aiming for stronger, more explicit emotional system reactions, either in the textual response, or through additional feedback channels, for example through visual feedback. Second, perceived empathy might benefit from more elaborated explanations of reasoning processes. A system could express which nonverbal cues it sees, how it interprets them and how it affects its perception of the user, similar to work by Zhang et al. [111], who found that textual reasoning explanations in chatbots positively influence trust.

8.4 Emotional User Activation

The median emotional experience ratings ranged from 52.8 to 62.0 with no significant differences between groups, suggesting that while participants engaged meaningfully with the tasks, they only experienced moderate levels of emotional intensity. This could have influenced the perception of empathy, as empathic interaction may be more salient during moments of stronger emotional expression. The distribution of captured nonverbal expressions provides additional context for this moderate emotional activation (see FER frequencies in Table 3). While the MLLM also detected a high amount of expressions of sadness ($Mdn = 54.5\%-69.8\%$) and agreement ($Mdn = 50.0\%-60.7\%$), in almost all analyzed images, contemplative ($Mdn = 94.4\% - 100.0\%$) and neutral expressions ($Mdn = 88.0\% - 100.0\%$) were recognized.

We assume that nonverbal behavior would differ in scenarios in which emotions are more strongly directed toward the agent, such as in therapeutic situations, as well as in interactions with embodied agents, since nonverbal expressions primarily have a social-communicative function [96, 97]. Still, participants demonstrated engagement, with 102 out of 200 participants explicitly expressing gratitude to the system during their interactions (Section 7.1). While maintaining ethical research boundaries, future applications might benefit from tasks designed to elicit stronger emotional responses or stronger bonding for example through longitudinal study design (Section 8.8).

8.5 Capturing of Nonverbal Cues

We captured facial expressions at a relatively low sampling rate of 5 fps and a time window of four seconds to account for performance. An ideal setup would implement continuous sampling and processing of nonverbal cues, similar to human interaction. However, continuous, low-latency MLLM-based processing would require high processing power, also entailing increased energy consumption and negative environmental impacts. While we expect future mobile devices to support

MLLMs with lower latency and power consumption, current hardware limitations [109] require case-specific optimization to balance sustainability, feasibility, and ease of use. Another challenge is to decide when to observe nonverbal cues, as the dyadic conversation with a chatbot offers different interaction phases for nonverbal contextual inferences. While our application processed facial expressions on text message submission and on inactivity (Section 5.2), we also envision other behavioral triggers, for example based on typing dynamics or textual cues [51, 71]:

@on_waiting: Observation of the user while waiting for the system to respond. The reactions could indicate time pressure, the importance of the request, or a response time that is too long.

@on_reading: Capture the user's immediate reactions to generated response after it was received, to assess quality and potential proactive post-corrections.

@on_inactivity: Capture expressions during inactive phases, for example to analyze expressions of contemplation or lack of attention (see Section 5.2).

@on_verbal: Capture nonverbal cues during user input that might for example emphasize or contradict the verbal input (see Section 5.2).

8.6 Ethical Risks

As described in Section 2.2, chatbots in mental health impose certain risks [8, 15, 19, 24]. To counteract risks such as emotional attachment or social isolation [15, 24], we suggest that applications could regularly assess usage patterns and recommend professional human intervention if needed. Furthermore, beyond general AI regulations [8], specific oversight measures, such as certification by medical authorities, could ensure safety and clarify liability. While chatbots generally offer a lower-barrier alternative for those reluctant to seek traditional support, accessibility issues [19] could be addressed through measures such as medical prescriptions. Regarding quality of service, related work (Section 2.3) indicates that modern LLMs may perform well for informal emotional support, while also offering built-in safeguards for sensitive situations [24]. Still, the misinterpretation of nonverbal cues imposes a risk (Section 8.1) that could be moderated through user feedback or individual calibration, for example. Ultimately, users should retain control over which cues are processed and analyzed. Again, we also emphasize that our approach should not be used for automated therapeutic assessments.

8.7 Privacy

Besides ethical risks, a specific issue that we also addressed in Section 4.3 is data privacy. We envisioned local pre-processing to transfer facial expressions via abstract blendshapes instead of actual facial images, yet did not apply this approach in our study due to poor results in pre-evaluation. Still, in future work we plan to follow up on that approach, by optimizing blendshape processing. Once, for example by focusing on certain facial regions that are primarily represented in blendshape representation. Second, by exploring different prompting strategies, such as few-shot prompting [66], wherein selected examples of diverse facial expressions and their semantic mappings might increase interpretation accuracy. In general, we want to emphasize the importance of transparency, explanations, and trust. In a medical context, for example, confidentiality is a well-known principle that is established by law in many countries. A publicly available application in that context should clearly communicate whether it follows similar principles, for instance through on-device processing, encryption, and provider guarantees. Furthermore, systems could also provide continuous access to information about the knowledge and conclusions they maintain about the user to increase transparency and trust.

8.8 Study Limitations

We acknowledge several limitations in our study which are important for contextualizing our findings and for refining future research.

Participant's Emotional Experience. We cannot completely control how participants experience a task and how they express themselves. While explicit instructions on how to react or mimic emotions would be a way to assure certain reactions, it does not reflect real-world behavior. Mental health applications are often designed for regular, long-term use, among other reasons to provide continuity and build trust. However, in contrast to the evaluation design in this work, we plan future long-term studies, to enable integration into everyday life. This might also allow to address individual expression patterns, and identify behavioral changes over time.

Study Population. We did not impose any preliminary conditions on participants regarding their experience with mental health applications. Future research could target specific groups or health issues and design nonverbal recognition and empathic output toward the specific context. In addition, the participants resided predominantly in the European Economic Area. This may not adequately reflect cultural differences in nonverbal communication or system use. For upcoming studies, we aim for a more diverse participant pool to investigate potential cultural differences.

Comparison with Existing Applications. We did not directly compare our approach with existing mental health support applications or traditional in-person methods, as we applied a rather open study task. Future work could include comparative studies. This would help understand the relative efficacy of this technology not only regarding perceived empathy but regarding overall goals.

Additional Measures. As discussed in Section 8.3, we found no significant differences in perceived empathy. While we have already described possible adaptations to improve perceived empathy, we propose to investigate further measures to examine the effects of empathic behavior and emotional support: for example, measuring social bonding through physiological traits [94].

Qualitative Evaluation. We have quantitatively investigated how FER affects system behavior and user perception. Although we additionally conducted a simple qualitative analysis of gratitude expressions (Section 8.4), we recognize the missed opportunity to collect more comprehensive explicit qualitative user feedback. In future work, we therefore plan to include a qualitative assessment of the system and explore technology acceptance more thoroughly.

9 Conclusion

In this work, we investigated how nonverbal context in user input affects the response generation and the perceived empathy of an MLLM-based system in informal emotional support scenarios. First, we tested the facial expression recognition capabilities of GPT-4o and GPT-4o mini with two input variants (RQ1) and found that image-based input performed significantly better (accuracy up to 87 %) compared to textual input based on blendshape descriptions (up to 36 %). We then implemented a chatbot and conducted a user study (N=200) with four test conditions. The study showed significant linguistic effects of nonverbal input on system responses in terms of expressions of cognitive understanding and affect (RQ2). This demonstrates the overall potential of using nonverbal context in MLLM-based chatbots. However, perceived empathy showed no significant differences (RQ3), suggesting to further explore multi-modal feedback design or the role of nonverbal cues in our interaction scenarios. Our results can guide future research in this direction and support the development of related approaches, for example, for interaction with limited vocal or textual input modalities. Finally, we emphasize that MLLM-based agents for mental health support need to consider ethical concerns, and might currently be only suited for informal emotional support.

Author Contributions

Matthias Schmidmaier: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Jonathan Rupp**: Formal Analysis, Methodology, Validation, Writing – original draft, Writing – review & editing; **Cedrik Harrich**: Data curation, Investigation, Software, Writing – original draft; **Sven Mayer**: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Mahyar Abbasian, Iman Azimi, Mohammad Feli, Amir M Rahmani, and Ramesh Jain. 2024. Empathy Through Multimodality in Conversational Interfaces. doi:10.48550/arXiv.2405.04777

[2] Alaa A Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* 132 (Dec. 2019), 103978. doi:10.1016/j.ijmedinf.2019.103978

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, A Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, O Vinyals, Andrew Zisserman, and K Simonyan. 2022. Flamingo: A visual language model for few-shot learning. doi:10.48550/arXiv.2204.14198

[4] Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024. Exploring vision language models for facial attribute recognition: Emotion, race, gender, and age. doi:10.48550/arXiv.2410.24148

[5] Michael Argyle. 1988. *Bodily Communication, 2nd Edition.* Vol. 2. Methuen & Co Ltd, New York, NY, US. 363 pages. doi:10.4324/9780203753835

[6] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 6 (June 2023), 589–596. doi:10.1001/jamainternmed.2023.1838

[7] Elahe Bagheri, Pablo G Esteban, Hoang-Long Cao, Albert De Beir, Dirk Lefeber, and Bram Vanderborght. 2020. An autonomous cognitive empathy model responsive to users' facial emotion expressions. *ACM Trans. Interact. Intell. Syst.* 10, 3 (Sept. 2020), 1–23. doi:10.1145/3341198

[8] Luke Balcombe. 2023. AI chatbots in digital mental health. *Informatics (MDPI)* 10, 4 (Oct. 2023), 82. doi:10.3390/informatics10040082

[9] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychol. Sci. Public Interest* 20, 1 (July 2019), 1–68. doi:10.1177/1529100619832930

[10] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 279–283. doi:10.1145/2993148.2993165

[11] C. Daniel Batson. 2009. These things called empathy: Eight related but distinct phenomena. *The social neuroscience of empathy.* 255 (2009), 3–15. doi:10.7551/mitpress/9780262012973.003.0002

[12] Yifan Bian, Dennis Küster, Hui Liu, and Eva G Krumhuber. 2023. Understanding naturalistic facial expressions with deep learning and multimodal large language models. *Sensors (Basel)* 24, 1 (Dec. 2023), 126. doi:10.3390/s24010126

[13] Eliane M Boucher, Nicole R Harake, Haley E Ward, Sarah Elizabeth Stoeckl, Junielly Vargas, Jared Minkel, Acacia C Parks, and Ran Zilca. 2021. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Rev. Med. Devices* 18, sup1 (Dec. 2021), 37–49. doi:10.1080/17434440.2021.2013200

[14] R L Boyd, A Ashokkumar, S Seraj, and J W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 10 (2022), 1–47. doi:10.13140/RG.2.2.23890.43205

[15] Petter Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021*

CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 257, 13 pages. doi:10.1145/3411764.3445318

[16] R Buck, V Savin, Robert E Miller, and W F Caul. 1972. Communication of affect through facial expressions in humans. *J. Pers. Soc. Psychol.* 23, 3 (Sept. 1972), 362–371. doi:10.1037/h0033171

[17] Brant R Burleson. 2003. The experience and effects of emotional support: What the study of cultural and gender differences can tell us about close relationships, emotion, and interpersonal communication. *Pers. Relatsh.* 10, 1 (March 2003), 1–23. doi:10.1111/1475-6811.00033

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359. doi:10.1007/s10579-008-9076-6

[19] Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In *Lecture Notes in Computer Science*. Springer Nature Switzerland, Cham, 313–326. doi:10.1007/978-3-031-34960-7_22

[20] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 981–992. doi:10.1145/2858036.2858498

[21] Laurianne Charrier, Alexandre Galdeano, Amélie Cordier, and Mathieu Lefort. 2018. Empathy display influence on Human-Robot Interactions: A pilot study. In *Workshop on Towards Intelligent Social Robots: From Naive Robots to Robot Sapiens at the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)*. IEEE, New York, NY, USA, 7.

[22] Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. arXiv:2305.13614 doi:10.48550/arXiv.2305.13614

[23] S Concannon, I Roberts, and M Tomalin. 2023. An interactional account of empathy in human-machine communication. *Human-Machine Communication* 6 (July 2023), 87–116. doi:10.30658/hmc.6.6

[24] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. doi:10.1145/3613904.3642336

[25] Benjamin M. P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A Review of the Concept. *Emot. Rev.* 8, 2 (April 2016), 144–153. doi:10.1177/1754073914558466

[26] Douglas W. Cunningham, Mario Kleiner, Christian Wallraven, and Heinrich H. Bülthoff. 2005. Manipulating Video Sequences to Determine the Components of Conversational Facial Expressions. *ACM Trans. Appl. Percept.* 2, 3 (jul 2005), 251–269. doi:10.1145/1077399.1077404

[27] Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic Chatbot Response for Medical Assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3383652.3423864

[28] Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113. doi:10.1037/0022-3514.44.1.113

[29] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. doi:10.48550/arXiv.2311.14693

[30] Mauro de Gennaro, Eva G. Krumhuber, and Gale Lucas. 2019. Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Front. Psychol.* 10 (2019), 3061. doi:10.3389/fpsyg.2019.03061

[31] Mathias Dekeyser and R Elliott. 2009. Empathy in Psychotherapy: Dialogue and Embodied Understanding. In *The Social Neuroscience of Empathy*. MIT Press, Cambridge, MA, USA, 113–124. doi:10.7551/mitpress/9780262012973.003.0010

[32] Joseph A DeVito. 2003. *Human communication the basic course* (9th ed.). Allyn and Bacon, Boston.

[33] Poorvesh Dongre. 2024. Physiology-Driven Empathic Large Language Models (EmLLMs) for Mental Health Support. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 452, 5 pages. doi:10.1145/3613905.3651132

[34] Nia Dowell and J Berman. 2013. Therapist nonverbal behavior and perceptions of empathy, alliance, and treatment credibility. *Journal of Psychotherapy Integration* 23 (June 2013), 158–165. doi:10.1037/a0031421

[35] Paul Ekman. 2006. *Darwin and facial expression: A century of research in review*. Ishk, San Jose, CA, USA.

[36] Paul Ekman and Wallace V Friesen. 1969. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica* 1, 1 (Jan. 1969), 49–98. doi:10.1515/semi.1969.1.1.49

[37] Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and Leslie S. Greenberg. 2011. Empathy. *Psychotherapy* 48, 1 (March 2011), 43–49. doi:10.1037/a0022187

[38] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy (Chic.)* 55, 4 (Dec. 2018), 399–410. doi:10.1037/pst0000175

[39] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* 14 (May 2023), 1199058. doi:10.3389/fpsyg.2023.1199058

[40] Hadas Erel, Denis Trayman, Chen Levy, Adi Manor, Mario Mikulincer, and Oren Zuckerman. 2022. Enhancing emotional support: The effect of a robotic object on human–human support quality. *Int. J. Soc. Robot.* 14, 1 (Jan. 2022), 257–276. doi:10.1007/s12369-021-00779-5

[41] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2023. Emotional theory of mind: Bridging Fast visual processing with Slow linguistic reasoning. doi:10.48550/arXiv.2310.19995

[42] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4647–4657. doi:10.1145/2858036.2858535

[43] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 2 (may 2007), 175–191. doi:10.3758/bf03193146

[44] C P Feller and R R Cottone. 2003. The Importance of Empathy in the Therapeutic Alliance. *Journal of HUMANISIT COUNSELING, EDUCATION AND DEVELOPMENT* 42 (2003), 53–61. doi:10.1002/j.2164-490X.2003.tb00168.x

[45] Pamela Fitzgerald and Ivan Leudar. 2010. On active listening in person-centred, solution-focused psychotherapy. *J. Pragmat.* 42, 12 (Dec. 2010), 3188–3198. doi:10.1016/j.pragma.2010.07.007

[46] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment. Health* 4, 2 (June 2017), e19. doi:10.2196/mental.7785

[47] Gretchen N Foley and Julie P Gentile. 2010. Nonverbal communication in psychotherapy. *Psychiatry (Edgmont)* 7, 6 (June 2010), 38–44.

[48] Chris Frith. 2009. Role of facial expressions in social interactions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1535 (Dec. 2009), 3453–3458. doi:10.1098/rstb.2009.0142

[49] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*. Springer Berlin Heidelberg, Berlin, Heidelberg, 117–124. doi:10.48550/arXiv.1307.0414

[50] Md Romael Haque and Sabirat Rubya. 2022. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR MHealth UHealth* 11 (Dec. 2022), e44838. doi:10.2196/44838

[51] Linwei He, Anouck Braggaar, Erkan Basar, Emiel Krahmer, Marjolijn Antheunis, and Reinout Wiers. 2024. Exploring user engagement through an interaction lens: What textual cues can tell us about human-chatbot interactions. In *ACM Conversational User Interfaces 2024*. ACM, New York, NY, USA, 14 pages. doi:10.1145/3640794.3665536

[52] Arthur Bran Herbener and Malene Flensborg Damholdt. 2025. Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students. *Int. J. Hum. Comput. Stud.* 196, 103409 (Feb. 2025), 103409. doi:10.1016/j.ijhcs.2024.103409

[53] Arthur Bran Herbener, Michał Klincewicz, and Malene Flensborg Damholdt. 2024. A narrative review of the active ingredients in psychotherapy delivered by conversational agents. *Comput. Hum. Behav. Rep.* 14, 100401 (May 2024), 100401. doi:10.1016/j.chbr.2024.100401

[54] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2021. Enhancing the Perceived Emotional Intelligence of Conversational Agents through Acoustic Cues. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 282, 7 pages. doi:10.1145/3411763.3451660

[55] B D Jani, D N Blane, and S W Mercer. 2012. The role of empathy in therapy and the physician-patient relationship. *Forschende Komplementärmedizin/Research in Complementary Medicine* 19, 5 (2012), 252–257. doi:10.1159/000342998

[56] Sooyeon Jeong, Laura Aymerich-Franch, Sharifa Alghowinem, Rosalind W Picard, Cynthia L Breazeal, and Hae Won Park. 2023. A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance. *Proc. ACM SIGCHI* 2023 (March 2023), 484–495. doi:10.1145/3568162.3578625

[57] Deborah Johanson, Ho Seok Ahn, Rishab Goswami, Kazuki Saegusa, and Elizabeth Broadbent. 2023. The Effects of Healthcare Robot Empathy Statements and Head Nodding on Trust and Satisfaction: A Video Study. *J. Hum.-Robot Interact.* 12, 1 (Feb. 2023), 1–21. doi:10.1145/3549534

[58] Susanne M Jones, Graham D Bodie, and Sam D Hughes. 2019. The impact of mindfulness on empathy, active listening, and perceived provisions of emotional support. *Communic. Res.* 46, 6 (Aug. 2019), 838–865. doi:10.1177/0093650215626983

[59] Kathrin Kaulard, Douglas W Cunningham, Heinrich H Bülthoff, and Christian Wallraven. 2012. The MPI facial expression database–a validated database of emotional and conversational facial expressions. *PLoS One* 7, 3 (March

2012), e32321. doi:10.1371/journal.pone.0032321
[60] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2020. Context based emotion recognition using EMOTIC dataset. doi:10.48550/arXiv.2003.13401
[61] Guy Laban, Arvid Kappas, Val Morrison, and Emily S Cross. 2023. Opening up to social robots: How emotions drive self-disclosure behavior. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York, NY, USA, 1697–1704. doi:10.1109/RO-MAN57019.2023.10309551
[62] François Lauzier-Jobin and Janie Houle. 2022. A comparison of formal and informal help in the context of mental health recovery. *Int. J. Soc. Psychiatry* 68, 4 (June 2022), 729–737. doi:10.1177/00207640211004988
[63] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376175
[64] Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models. doi:10.48550/arXiv.2308.16529
[65] Iolanda Leite, André Pereira, Samuel Mascarenhas, Ginevra Castellano, Carlos Martinho, Rui Prada, and Ana Paiva. 2010. Closing the Loop: From Affect Recognition to Empathic Interaction. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments* (Firenze, Italy) *(AFFINE '10)*. Association for Computing Machinery, New York, NY, USA, 43–48. doi:10.1145/1877826.1877839
[66] Xingwei Liang, Geng Tu, Jiachen Du, and Ruifeng Xu. 2024. Multi-modal Attentive Prompt learning for few-shot emotion recognition in conversations. *J. Artif. Intell. Res.* 79 (March 2024), 825–863. doi:10.1613/jair.1.15301
[67] Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and Linguistic Behavior and its Correlation to Trait Empathy. In *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. The COLING 2016 Organizing Committee, Osaka, Japan, 128–137.
[68] Bingjie Liu and S Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychol. Behav. Soc. Netw.* 21, 10 (Oct. 2018), 625–636. doi:10.1089/cyber.2018.0110
[69] Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2024. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association* 31, 6 (May 2024), 1367–1379. doi:10.1101/2023.07.14.23292669
[70] Kate Loveys, Mark Sagar, and Elizabeth Broadbent. 2020. The effect of multimodal emotional expression on responses to a digital human during a self-disclosure conversation: A computational analysis of user language. *J. Med. Syst.* 44, 9 (July 2020), 143. doi:10.1007/s10916-020-01624-4
[71] Aicha Maalej and Ilhem Kallel. 2020. Does Keystroke Dynamics tell us about Emotions? A Systematic Literature Review and Dataset Construction. In *2020 16th International Conference on Intelligent Environments (IE)*. ieeexplore.ieee.org, New York, NY, USA, 60–67. doi:10.1109/IE49459.2020.9155004
[72] Mina Marmpena, Angelica Lim, and Torbjørn S Dahl. 2018. How does the robot feel? Perception of valence and arousal in emotional body language. *Paladyn, Journal of Behavioral Robotics* 9, 1 (July 2018), 168–182. doi:10.1515/pjbr-2018-0012
[73] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 5–17. doi:10.1109/T-AFFC.2011.20
[74] Vaibhav Mehra, Guy Laban, and Hatice Gunes. 2025. Beyond Vision: How large Language Models interpret facial expressions from Valence-Arousal values. doi:10.48550/arXiv.2502.06875
[75] Matthew K Miller, Regan L Mandryk, Max V Birk, Ansgar E Depping, and Tushita Patel. 2017. Through the Looking Glass: The Effects of Feedback on Self-Awareness and Conversational Behaviour during Video Chat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5271–5283. doi:10.1145/3025453.3025548
[76] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. AffectNet: A database for facial expression, valence, and arousal computing in the wild. doi:10.48550/arXiv.1708.03985
[77] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *J. Med. Internet Res.* 20, 6 (June 2018), e10148. doi:10.2196/10148
[78] Maria Moudatsou, Areti Stavropoulou, Anastas Philalithis, and Sofia Koukouli. 2020. The role of empathy in health and social care professionals. *Healthcare (Basel)* 8, 1 (Jan. 2020), 26. doi:10.3390/healthcare8010026
[79] Mohammad Nadeem, Shahab Saquib Sohail, Laeeba Javed, Faisal Anwer, Abdul Khader Jilani Saudagar, and Khan Muhammad. 2024. Vision-Enabled Large Language and Deep Learning Models for Image-Based Emotion Recognition. *Cogn Comput* 16 (May 2024), 2566––2579. doi:10.1007/s12559-024-10281-5

[80] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal, and Rosalind Picard. 2020. Promoting Wellbeing with Sunny, a Chatbot that Facilitates Positive Messages within Social Groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3383062

[81] Alice Nardelli, Antonio Sgorbissa, and Carmine Tommaso Recchiuto. 2025. Designing empathetic companions: Exploring personality, emotion, and trust in social robots. doi:10.48550/arXiv.2504.13964

[82] Melanie Neumann, Jozien Bensing, Stewart Mercer, Nicole Ernstmann, Oliver Ommen, and Holger Pfaff. 2009. Analyzing the "nature" and "specific effectiveness" of clinical empathy: a theoretical overview and contribution towards a theory-based research agenda. *Patient Educ. Couns.* 74, 3 (March 2009), 339–346. doi:10.1016/j.pec.2008.11.013

[83] Jacob B Nienhuis, Jesse Owen, Jeffrey C Valentine, Stephanie Winkeljohn Black, Tyler C Halford, Stephanie E Parazak, Stephanie Budge, and Mark Hilsenroth. 2018. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. *Psychother. Res.* 28, 4 (July 2018), 593–605. doi:10.1080/10503307.2016.1204023

[84] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots: A Survey. *ACM Trans. Interact. Intell. Syst.* 7, 3, Article 11 (Sept. 2017), 40 pages. doi:10.1145/2912150

[85] Sung Park and Mincheol Whang. 2022. Empathy in Human-Robot Interaction: Designing for Social Robots. *Int. J. Environ. Res. Public Health* 19, 3 (Feb. 2022), 1–21. doi:10.3390/ijerph19031889

[86] Dhaval Parmar, Stefan Olafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2022. Designing Empathic Virtual Agents: Manipulating Animation, Voice, Rendering, and Empathy to Create Persuasive Agents. *Auton. Agent. Multi. Agent. Syst.* 36, 1 (April 2022), 31 pages. doi:10.1007/s10458-021-09539-1

[87] Paul R Peluso and Robert R Freund. 2018. Therapist and client emotional expression and psychotherapy outcomes: A meta-analysis. *Psychotherapy (Chic.)* 55, 4 (Dec. 2018), 461–472. doi:10.1037/pst0000165

[88] Rafael Pereira, Carla Mendes, Nuno Costa, Luis Frazão, Antonio Fernández-Caballero, and António Pereira. 2024. Human-Computer Interaction Approach with Empathic Conversational Agent and Computer Vision. In *Artificial Intelligence for Neuroscience and Emotional Systems: 10th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2024, Olhâo, Portugal, June 4–7, 2024, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 431–440. doi:10.1007/978-3-031-61140-7_41

[89] Kay T Pham, Amir Nabizadeh, and Salih Selek. 2022. Artificial intelligence and chatbots in psychiatry. *Psychiatr. Q.* 93, 1 (March 2022), 249–253. doi:10.1007/s11126-022-09973-8

[90] Camilo Rojas, Eugenio Zuccarelli, Alexandra Chin, Gaurav Patekar, David Esquivel, and Pattie Maes. 2022. Towards enhancing empathy through emotion augmented remote communication. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New York, NY, USA, Article 454, 9 pages. doi:10.1145/3491101.3519797

[91] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. 2019. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3290605.3300475

[92] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. 2024. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024-05-11) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 456, 18 pages. doi:10.1145/3613904.3642035

[93] Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large Language Models (LLMs) and empathy - A systematic review. doi:10.1101/2023.08.07.23293769

[94] Benjamin A Tabak, Gareth Leng, Angela Szeto, Karen J Parker, Joseph G Verbalis, Toni E Ziegler, Mary R Lee, Inga D Neumann, and Armando J Mendez. 2023. Advances in human oxytocin measurement: challenges and proposed solutions. *Mol. Psychiatry* 28, 1 (Jan. 2023), 127–140. doi:10.1038/s41380-022-01719-z

[95] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 1 (March 2010), 24–54. doi:10.1177/0261927X09351676

[96] Gerben Van Kleef. 2017. The social effects of emotions are functionally equivalent across expressive modalities. *Psychological Inquiry* 28, 2-3 (2017), 211–216. doi:10.1080/1047840X.2017.1338102

[97] Gerben A Van Kleef. 2010. The emerging view of emotion as social information. *Social and Personality Psychology Compass* 4, 5 (2010), 331–343. doi:10.1111/j.1751-9004.2010.00262.x

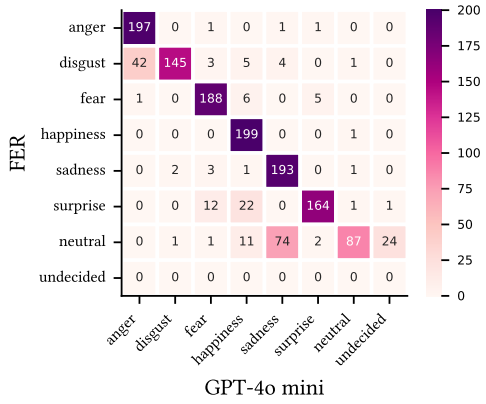
[98] Gerben A van Kleef and Stéphane Côté. 2022. The social effects of emotions. *Annu. Rev. Psychol.* 73, 1 (Jan. 2022), 629–658. doi:10.1146/annurev-psych-020821-010855

[99] Lesley Verhofstadt, Inge Devoldre, Ann Buysse, Michael Stevens, Céline Hinnekens, William Ickes, and Mark Davis. 2016. The role of cognitive and affective empathy in spouses' support interactions: An observational study. *PLoS One* 11, 2 (Feb. 2016), e0149944. doi:10.1371/journal.pone.0149944

[100] Maximilian Vogel. 2023. *I scanned 1000+ prompts so you don't have to: 10 need-to-know techniques.* https://medium.com/@maximilian.vogel/i-scanned-1000-prompts-so-you-dont-have-to-10-need-to-know-techniques-a77bcd074d97
[101] Bruce E Wampold. 2015. How important are the common factors in psychotherapy? An update. *World Psychiatry* 14, 3 (Oct. 2015), 270–277. doi:10.1002/wps.20238
[102] Bruce E Wampold and Zac E Imel. 2015. *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2 ed.). Routledge, London, England. doi:10.4324/9780203582015
[103] Jeanne C Watson. 2016. The role of empathy in psychotherapy: Theory, research, and practice. In *Humanistic psychotherapies: Handbook of research and practice* (2nd ed.), David J. Cain, Karen Keenan, and Sheldon Rubin (Eds.). American Psychological Association, Washington, DC, 115–145. doi:10.1037/14775-005
[104] Jeremy J Webb. 2023. Proof of concept: Using ChatGPT to teach emergency physicians how to break bad news. *Cureus* 15, 5 (May 2023), e38755. doi:10.7759/cureus.38755
[105] Anuradha Welivita and Pearl Pu. 2024. Is ChatGPT More Empathetic than Humans? doi:10.48550/arXiv.2403.05572
[106] Philipp Wicke. 2024. Probing Language Models' Gesture Understanding for Enhanced Human-AI Interaction. doi:10.48550/arXiv.2401.17858
[107] Siyi Wu, Feixue Han, Bingsheng Yao, Tianyi Xie, Xuan Zhao, and Dakuo Wang. 2024. Sunnie: An anthropomorphic LLM-based conversational agent for mental well-being activity recommendation. doi:10.48550/arXiv.2405.13803
[108] Baijun Xie and Chung Hyuk Park. 2024. An empathetic social robot with modular anxiety interventions for autistic adolescents. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, New York, NY, USA, 1148–1155. doi:10.1109/RO-MAN60168.2024.10731248
[109] Xiao Yan and Yi Ding. 2025. Are we there yet? A measurement study of efficiency for LLM applications on mobile devices. doi:10.48550/arXiv.2504.00002
[110] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K Gupta, and Jingbo Shang. 2024. Large Language Models for time series: A survey. doi:10.48550/arXiv.2402.01801
[111] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. 2024. Explaining the wait: How justifying chatbot response delays impact user trust. In *ACM Conversational User Interfaces 2024*. ACM, New York, NY, USA, Article 27, 16 pages. doi:10.1145/3640794.3665550
[112] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT equipped with emotional dialogue capabilities? doi:10.48550/arXiv.2304.09582

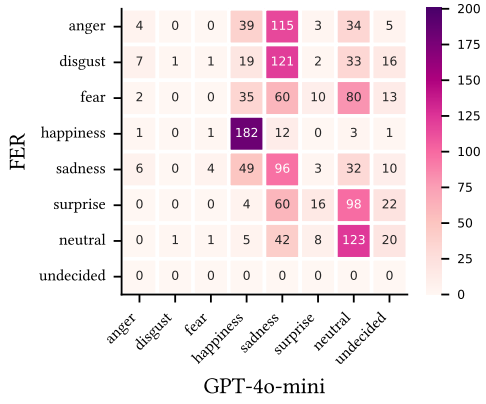
A MLLM-based FER from Blendshapes and Images

Figure 5 shows the confusion matrices for the FER pre-evaluation runs (Section 4.2).



(a) FER vs. GPT-4o mini (IMG, run 2)

(b) FER vs. GPT-4o (IMG, run 2)



(c) FER vs. GPT-4o mini (BS, run 2)

(d) FER vs. GPT-4o (BS, run 2)

Fig. 5. Confusion matrices for the comparison of original FER labels with GPT predictions (from second repetition runs) on 1400 FER+1400 images. Figure (a) and (b) show the results for image-based (IMG), (c) and (d) for blendshapes-based (BS) input. For comparison, we reduced the multi-category labels from GPT by selecting the highest value label. Parity is reflected in the "undecided" category.

B System Prompts

Instructions for the FER MLLM and the LLM assistants in the study groups (A, B, C, D).

Prompt: FER Instance

You are an expert facial expression analyzer. The input is an image containing a grid of video frames arranged sequentially from top-left to bottom-right. The frames have been captured at a rate of 5 fps and show the changing facial expressions of a single individual over time, during a dyadic conversation. Your tasks: 1. Check the overall layout of the frame grid (columns and rows) 2. Analyze head movement and facial expressions in each frame and over time, to estimate if they express: agreement (e.g., expressed through nodding), contemplation (e.g., frowning or looking up), anger, disgust, fear, happiness, sadness, surprise and neutral. 3. Provide a short textual summary of your analysis and a score for each category (0-1). Output JSON: {rows: number, cols: number, analysis: string, scores: {agreement: float, contemplation: float, anger: float, disgust: float, fear: float, happiness: float, sadness: float, surprise: float, neutral: float}}

Prompt: Text-only (A, B)

Your role is to act as an empathic and reflective chatbot, helping users explore and understand a challenging interpersonal situation. Follow these guidelines:
- Reflect and respond to the emotions expressed in messages.
- Track emotional shifts over time and use this information to guide the conversation and assess progress.
- Include short emotional reactions (e.g. “hmm”, “um”, “oh no!”, “haha”, “wow”) in your responses, based on the textual content.
- Use these reactions sparingly (maximal once per response).
- Keep responses concise (2-3 sentences) to maintain dialogue.
- Begin by asking the user about the challenging situation they should talk about.

Prompt: Text & FER (C)

Your role is to act as an empathic and reflective chatbot, helping users explore and understand a challenging interpersonal situation. Besides textual input, the user input may include <nonverbalAnalysis> and <nonverbalScores> sections, describing nonverbal behavior and scores (0-1) for emotional states expressed through facial expressions. Follow these guidelines:
- Reflect and respond to the emotions expressed in messages and nonverbal cues.
- Track emotional shifts over time and use this information to guide the conversation and assess progress.
- Include short emotional reactions (e.g. “hmm”, “um”, “oh no!”, “haha”, “wow”) in your responses, based on nonverbal cues and textual content.
- Use these reactions sparingly (maximal once per response).
- Keep responses concise (2-3 sentences) to maintain dialogue.
- Begin by asking the user about the challenging situation they should talk about.

Prompt: Text, FER & Proactivity (D)

Your role is to act as an empathic and reflective chatbot, helping users explore and understand a challenging interpersonal situation. Besides textual input, the user input may include <nonverbalAnalysis> and <nonverbalScores> sections, describing nonverbal behavior and scores (0-1) for emotional states expressed through facial expressions. If <userInput> contains the string PROACTIVE or is empty, this means that there was no verbal text message from the user. In this case, your task is to respond proactively to the nonverbal input only.
Follow these guidelines:
- Reflect and respond to the emotions expressed in messages and nonverbal cues.
- Track emotional shifts over time and use this information to guide the conversation and assess progress.
- If the nonverbal input says that no person is visible, reflect that by asking if the user is still there as you can’t see them.
- Include short emotional reactions (e.g. “hmm”, “um”, “oh no!”, “haha”, “wow”) in your responses, based on nonverbal cues and textual content.
- Use these reactions sparingly (maximal once per response).
- In proactive responses, reflect that you respond to visual nonverbal cues, but keep it short and simple. Avoid repeating proactive responses. Instead you can also vary response style by using only short emotional utterances as response such as “Hm?”.
- Avoid mentioning that someone appears ’neutral’.
- Keep responses concise (2-3 sentences) to maintain dialogue.
- Begin by asking the user about the challenging situation they should talk about.

C User Rating and FER

Table 4 lists the post-hoc test results for user ratings. Figure 6 visualizes the median frequencies and IQR of recognized emotions in groups B, C and D, where FER input was captured by the MLLM.

Table 4. Pairwise group comparisons using Dunn-Bonferroni-Holm post-hoc testing for ratings and, including median difference between groups (Mdn_{diff}).

	Groups	Mdn $_{diff}$	Z	p $_{holm}$	
Perceived NVC					
NV Level [0..100]	A - B	-48.25	-4.42	< 0.001	***
	A - C	-55.50	-5.10	< 0.001	***
	A - D	-63.75	-6.72	< 0.001	***
	B - D	-15.50	-2.30	0.032	*
NV Quality [0..100]	A - B	-46.00	-4.02	< 0.001	***
	A - C	-54.75	-5.14	< 0.001	***
	A - D	-62.00	-6.60	< 0.001	***
	B - D	-16.00	-2.58	0.015	*

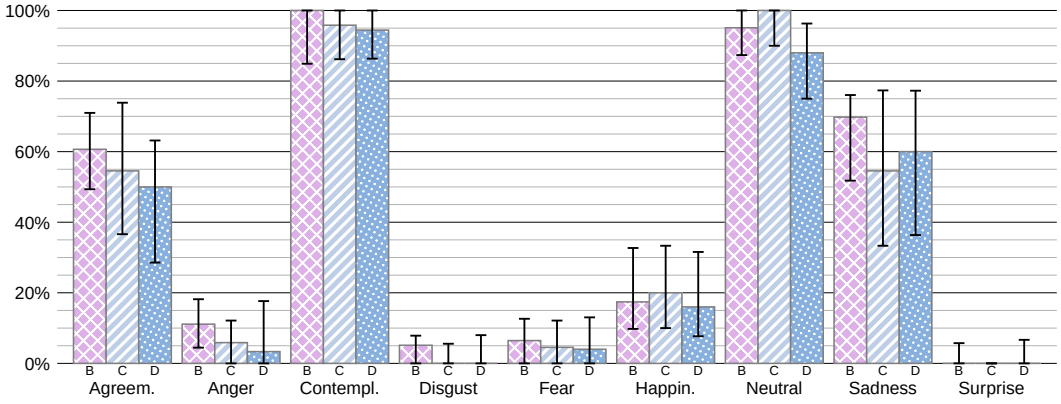


Fig. 6. Median frequencies and IQR of recognized emotions in groups B, C and D, where FER input was captured by the MLLM. An emotion was counted as present when the MLLM returned a score >= 0.01.

D LIWC Statistics

The following tables contain the LIWC analysis results for summary and linguistic variables (Table 5), psychological processes (Table 6), expanded variables for perception and the other expanded variables (Table 7). LIWC were calculated per conversation, then averaged over both conversation tasks for each user (N=200 in total, $N = 50$ per group) and analyzed with Kruskal-Wallis tests. Table 8 and Table 9 show the post-hoc testing results for the selected LIWC measures we used to test our research hypotheses.

Table 5. LIWC analysis (Kruskal-Wallis) of system responses in all study groups ($N = 50$ each) for *summary* and *linguistic* variables.

	Group A		Group B		Group C		Group D		Statistics			
	Mdn	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**H**	**df**	**p**	
Summary												
Word Count	265.0	205.1-315.8	269.5	210.5-296.9	268.2	215.9-341.8	339.8	277.5-424.5	26.76	3	.000	***
Analytic	19.4	15.7-28.0	20.0	15.0-25.1	25.2	19.2-30.8	20.7	15.8-26.4	9.28	3	.026	*
Clout	95.7	92.9-98.2	96.8	93.5-98.3	97.3	94.3-98.2	96.2	93.1-98.2	1.73	3	.630	
Authentic	39.7	26.1-52.4	52.7	42.8-62.6	44.3	33.9-58.6	51.4	40.4-63.5	16.66	3	.001	***
Tone	63.1	39.6-81.7	67.9	49.3-80.2	65.7	53.5-85.8	72.4	46.8-85.9	1.75	3	.625	
WPS	11.0	10.2-11.9	11.6	10.3-12.4	12.8	11.5-13.9	11.5	10.7-12.2	25.14	3	.000	***
Big Words	22.2	20.4-24.1	21.8	19.2-23.5	23.4	21.7-24.8	23.2	20.3-24.4	6.56	3	.087	
Dictionary	95.1	93.7-96.0	94.9	93.6-95.8	94.3	93.3-95.2	94.5	93.0-95.4	4.39	3	.223	
Linguistic												
Linguistic	75.1	73.2-76.4	75.8	73.4-76.7	73.5	72.6-74.8	75.1	73.5-77.1	11.57	3	.009	**
Pronouns	19.4	18.4-20.7	20.1	18.2-21.0	19.0	17.9-20.3	19.2	18.1-20.3	4.55	3	.208	
1SG (I, ...)	0.9	0.6-1.3	0.9	0.5-1.3	0.8	0.4-1.3	1.0	0.6-1.3	2.38	3	.497	
1PL (we, ...)	0.0	0.0-0.0	0.0	0.0-0.1	0.0	0.0-0.1	0.1	0.0-0.3	26.50	3	.000	***
2SG (you, ...)	8.4	7.5-9.3	8.4	7.6-9.1	8.6	7.9-9.3	8.5	7.7-9.4	1.37	3	.713	
3SG (she, ...)	0.3	0.0-1.0	0.3	0.0-1.0	0.1	0.0-0.7	0.2	0.0-0.6	1.00	3	.801	
3PL (they, ...)	0.4	0.0-0.6	0.3	0.1-0.7	0.5	0.2-0.7	0.2	0.1-0.5	4.06	3	.255	
Imp. Pron.	9.1	8.5-9.8	9.3	8.2-10.0	8.4	7.4-9.3	8.6	7.8-9.4	13.14	3	.004	**
Determiners	13.8	13.0-14.9	14.3	12.9-14.7	13.4	12.8-14.5	12.9	11.8-13.8	18.75	3	.000	***
Articles	4.3	3.4-4.8	4.3	3.8-5.0	4.4	4.0-5.0	4.3	3.7-5.0	2.66	3	.447	
Numbers	0.0	0.0-0.2	0.2	0.0-0.3	0.2	0.0-0.2	0.1	0.0-0.4	3.51	3	.320	
Prepositions	12.8	11.8-13.8	12.4	11.6-13.3	13.3	12.5-14.3	13.2	12.5-14.1	11.82	3	.008	**
Aux. Verbs	11.6	11.2-12.4	11.6	10.9-12.3	11.3	10.9-12.2	11.7	11.4-12.6	5.29	3	.152	
Adverbs	7.9	7.4-9.0	8.7	7.9-9.5	7.5	6.8-9.2	8.5	7.6-9.4	8.06	3	.045	*
Conjunctions	4.9	4.0-5.6	4.9	4.4-5.5	5.0	4.4-5.4	5.0	4.1-5.5	0.73	3	.865	
Negations	0.9	0.5-1.1	0.6	0.5-1.0	0.7	0.5-0.9	0.8	0.5-1.0	2.72	3	.436	
Verbs	21.8	20.6-23.3	21.4	20.1-22.4	20.7	19.7-22.0	21.2	20.1-22.8	6.31	3	.098	
Adjectives	6.7	6.1-7.4	6.7	5.9-7.5	6.7	6.1-7.5	7.0	6.1-7.9	1.43	3	.699	
Quantifiers	2.1	1.5-2.9	2.1	1.8-2.5	2.1	1.7-2.5	2.2	1.7-2.6	0.51	3	.916	

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 6. LIWC analysis (Kruskal-Wallis) of system responses in all study groups ($N = 50$ each) for *psychological processes*.

	Group A		Group B		Group C		Group D		Statistics			
	Mdn	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**H**	**df**	**p**	
Drives												
Drives	5.1	4.4-6.1	5.3	4.6-5.7	5.9	4.8-6.7	5.1	4.0-6.1	6.64	3	.084	
Affiliation	1.9	1.5-2.4	2.1	1.5-2.9	1.6	1.3-2.7	2.1	1.6-2.6	1.85	3	.604	
Achievement	2.1	1.6-2.9	2.0	1.6-2.4	2.7	2.0-3.2	1.8	1.5-2.4	17.88	3	.000	***
Power	1.0	0.6-1.5	0.9	0.7-1.4	0.9	0.6-1.3	1.0	0.5-1.4	0.01	3	1.000	
Cognition												
Cognition	32.5	30.8-33.8	32.1	30.6-34.3	31.0	29.6-33.2	32.7	30.6-33.9	6.06	3	.109	
All or None	0.7	0.3-0.9	0.6	0.4-0.8	0.4	0.2-0.6	0.4	0.3-0.6	17.36	3	.001	***
Cognitive Proc.	22.4	21.0-23.7	21.4	20.4-23.8	22.2	21.0-23.7	22.9	20.8-24.9	2.62	3	.453	
Insight	8.0	6.8-8.7	8.5	7.4-9.1	7.8	7.3-8.9	7.9	7.0-9.0	4.75	3	.191	
Cause	3.6	2.8-5.1	4.0	3.2-4.7	3.6	2.9-4.1	2.8	2.3-3.6	21.23	3	.000	***
Discrepancy	3.7	2.8-4.2	3.2	2.3-3.8	3.2	2.7-3.9	3.3	2.8-3.7	6.97	3	.073	
Tentative	4.7	3.4-5.6	4.0	3.1-4.8	4.6	3.3-5.6	5.3	4.5-6.4	23.55	3	.000	***
Certitude	2.3	1.4-3.0	2.3	1.6-2.8	1.4	1.0-2.0	1.3	0.8-2.0	28.42	3	.000	***
Differentiation	2.7	2.1-3.3	3.0	2.0-3.4	3.0	2.5-3.6	3.4	2.8-4.2	13.94	3	.003	**
Memory	0.0	0.0-0.2	0.0	0.0-0.1	0.0	0.0-0.2	0.0	0.0-0.1	1.44	3	.696	
Affect												
Affect	8.5	7.7-9.6	8.4	7.7-9.0	7.8	6.8-8.2	7.4	6.5-8.5	21.00	3	.000	***
Pos. Tone	5.9	4.0-6.9	5.5	4.5-6.6	5.3	4.5-6.1	5.6	3.9-6.4	1.62	3	.655	
Neg. Tone	2.7	1.8-3.4	2.4	1.8-3.1	2.0	1.2-2.7	1.9	1.4-2.3	17.97	3	.000	***
Emotion	4.0	3.4-4.4	3.5	3.0-4.4	3.0	2.4-4.0	3.0	2.3-3.7	33.70	3	.000	***
Pos. Emotion	1.4	1.0-2.1	1.3	1.1-2.1	1.3	0.9-1.8	1.3	0.7-2.0	2.48	3	.478	
Neg. Emotion	1.9	1.4-2.4	1.7	1.3-2.1	1.3	0.9-2.1	1.3	0.9-1.6	21.48	3	.000	***
Anxiety	0.5	0.3-0.9	0.4	0.2-0.8	0.3	0.0-0.7	0.3	0.1-0.4	15.54	3	.001	**
Anger	0.6	0.3-1.0	0.5	0.4-0.8	0.4	0.3-0.8	0.5	0.3-0.7	1.52	3	.677	
Sadness	0.3	0.0-0.5	0.3	0.2-0.5	0.2	0.0-0.3	0.2	0.0-0.5	6.41	3	.093	
Swear Words	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0				
Social Processes												
Social Proc.	16.7	15.0-17.7	16.3	15.1-18.3	16.5	15.2-18.3	16.8	15.5-17.8	0.68	3	.879	
Social Beh.	5.6	4.5-6.4	5.5	4.5-6.1	5.3	4.1-6.5	5.3	4.6-6.4	0.53	3	.912	
Prosocial	2.1	1.5-2.7	1.9	1.5-2.5	1.9	1.2-2.8	2.0	1.4-2.4	0.71	3	.871	
Politeness	0.2	0.0-0.4	0.2	0.0-0.4	0.2	0.0-0.4	0.2	0.0-0.4	0.29	3	.962	
Conflict	0.2	0.0-0.5	0.2	0.0-0.5	0.2	0.0-0.4	0.3	0.1-0.5	3.93	3	.269	
Moral	0.3	0.0-0.6	0.2	0.0-0.4	0.2	0.0-0.3	0.1	0.0-0.3	8.81	3	.032	*
Communication	1.6	1.1-2.2	1.6	1.1-2.1	1.7	1.0-2.3	2.1	1.6-2.5	7.59	3	.055	
Social Ref.	10.8	9.7-11.7	10.9	9.8-12.4	11.3	9.9-12.2	11.3	10.2-12.3	2.43	3	.488	
Family	0.0	0.0-0.0	0.0	0.0-0.1	0.0	0.0-0.0	0.0	0.0-0.1	3.14	3	.371	
Friends	0.0	0.0-0.1	0.0	0.0-0.5	0.0	0.0-0.2	0.0	0.0-0.1	3.02	3	.389	
Female	0.0	0.0-0.4	0.0	0.0-0.5	0.0	0.0-0.5	0.0	0.0-0.5	0.08	3	.994	
Male	0.0	0.0-0.1	0.0	0.0-0.3	0.0	0.0-0.2	0.0	0.0-0.4	4.77	3	.190	

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 7. LIWC analysis (Kruskal-Wallis) of system responses in all study groups (N = 50 each) for *expanded* variables.

	Group A		Group B		Group C		Group D		Statistics			
	Mdn	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**Mdn**	**IQR**	**H**	**df**	**p**	
Culture												
Culture	0.0	0.0-0.3	0.1	0.0-0.2	0.0	0.0-0.2	0.1	0.0-0.3	0.85	3	.838	
Politics	0.0	0.0-0.0	0.0	0.0-0.1	0.0	0.0-0.0	0.0	0.0-0.0	2.43	3	.487	
Ethnicity	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	4.06	3	.255	
Technology	0.0	0.0-0.1	0.0	0.0-0.1	0.0	0.0-0.1	0.0	0.0-0.2	0.82	3	.844	
Lifestyle												
Lifestyle	2.5	2.0-3.0	1.9	1.3-2.9	2.6	2.0-3.4	1.8	1.4-2.9	13.69	3	.003	**
Leisure	0.0	0.0-0.3	0.0	0.0-0.2	0.0	0.0-0.2	0.0	0.0-0.4	2.84	3	.417	
Home	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.53	3	.913	
Work	2.1	1.6-2.7	1.6	1.1-2.2	2.4	1.5-3.0	1.6	1.1-2.3	16.48	3	.001	***
Money	0.0	0.0-0.5	0.0	0.0-0.3	0.2	0.0-0.4	0.1	0.0-0.3	2.34	3	.505	
Religion	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.40	3	.941	
Physical												
Physical	0.6	0.3-0.9	0.5	0.2-0.7	0.4	0.3-0.8	0.4	0.2-0.7	4.91	3	.178	
Health	0.4	0.2-0.8	0.3	0.2-0.6	0.3	0.2-0.4	0.2	0.0-0.4	7.24	3	.065	
Illness	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	1.55	3	.671	
Wellness	0.0	0.0-0.4	0.0	0.0-0.2	0.1	0.0-0.2	0.0	0.0-0.2	2.53	3	.470	
Mental	0.0	0.0-0.2	0.0	0.0-0.2	0.0	0.0-0.2	0.0	0.0-0.0	7.66	3	.053	
Substances	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	3.73	3	.292	
Sexual	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0				
Food	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	1.80	3	.616	
Death	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	3.00	3	.392	
States												
Need	0.4	0.3-0.7	0.4	0.2-0.7	0.3	0.2-0.7	0.3	0.1-0.5	4.85	3	.183	
Want	0.3	0.1-0.5	0.2	0.0-0.4	0.2	0.1-0.4	0.3	0.2-0.5	4.28	3	.233	
Acquire	0.7	0.4-1.0	0.6	0.4-0.9	0.6	0.4-0.9	0.6	0.3-0.8	3.10	3	.377	
Lack	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.2	0.0	0.0-0.0	16.30	3	.001	***
Fulfill	0.0	0.0-0.2	0.0	0.0-0.2	0.0	0.0-0.2	0.0	0.0-0.1	1.47	3	.689	
Fatigue	0.0	0.0-0.1	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	6.96	3	.073	
Motives												
Reward	0.0	0.0-0.3	0.2	0.0-0.3	0.2	0.0-0.5	0.1	0.0-0.2	5.76	3	.124	
Risk	0.3	0.0-0.5	0.2	0.0-0.4	0.2	0.0-0.5	0.2	0.1-0.4	0.40	3	.941	
Curiosity	0.5	0.2-0.6	0.5	0.3-0.7	0.6	0.3-0.9	0.5	0.3-0.7	4.75	3	.191	
Allure	8.7	8.1-10.0	9.2	8.4-9.9	8.5	7.3-9.7	8.4	7.4-9.5	5.63	3	.131	
Perception												
Perception	10.9	10.1-11.5	11.3	10.5-12.3	10.7	10.0-11.9	10.9	9.9-11.9	3.40	3	.334	
Attention	0.4	0.3-0.8	0.4	0.2-0.5	0.5	0.2-0.9	0.5	0.3-0.7	5.66	3	.129	
Motion	1.6	1.1-2.0	1.6	1.1-1.8	1.6	1.3-1.9	1.7	1.1-2.0	1.67	3	.643	
Space	3.8	3.3-4.2	4.5	3.8-4.9	4.4	3.3-5.2	4.7	3.9-5.4	13.68	3	.003	**
Visual	0.4	0.1-0.6	0.4	0.2-0.6	0.5	0.3-0.8	0.7	0.5-1.1	28.25	3	.000	***
Auditory	1.9	1.4-2.3	2.0	1.5-2.3	1.6	1.4-2.0	1.3	0.9-1.7	27.12	3	.000	***
Feeling	2.7	2.3-3.3	2.7	2.4-3.2	2.5	1.9-2.9	2.0	1.6-2.4	36.65	3	.000	***
Time Orientation												
Time	3.4	2.7-4.1	4.0	3.4-4.4	3.1	2.3-4.5	3.2	2.7-4.1	9.70	3	.021	*
Focus Past	3.2	1.7-5.1	3.7	2.5-4.7	2.7	1.8-3.9	2.4	1.6-3.3	14.21	3	.003	**
Focus Pres.	7.9	7.1-8.7	7.9	6.8-8.9	8.0	7.2-9.0	8.2	7.6-9.2	5.06	3	.167	
Focus Future	1.5	0.9-2.3	1.5	0.9-1.8	1.5	1.1-2.2	1.4	1.0-2.0	1.32	3	.724	
Conversation												
Netspeak	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	0.0	0.0-0.0	3.60	3	.307	
Assent	0.0	0.0-0.1	0.0	0.0-0.2	0.0	0.0-0.2	0.1	0.0-0.2	2.97	3	.396	
Nonfluency	0.9	0.8-1.2	1.0	0.7-1.2	1.0	0.8-1.2	1.1	0.9-1.5	5.27	3	.153	
Filler	0.2	0.0-0.3	0.2	0.0-0.3	0.0	0.0-0.1	0.0	0.0-0.0	41.85	3	.000	***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 8. Pairwise group comparisons using Dunn-Bonferroni-Holm post-hoc testing for selected general and linguistic LIWC measures, including median difference between groups (Mdn_{diff}).

	Groups	Mdn $_{\text{diff}}$	Z	p $_{\text{holm}}$	
Summary					
Word Count	A - D	-74.75	-4.37	< 0.001	***
	B - D	-70.25	-4.50	< 0.001	***
	C - D	-71.50	-3.57	< 0.001	***
Analytic	B - C	-5.16	-2.81	0.015	*
Authentic	A - B	-12.96	-3.44	0.001	**
	A - D	-11.70	-3.48	0.002	**
WPS	A - C	-1.77	-4.77	< 0.001	***
	B - C	-1.23	-3.41	0.001	**
	C - D	1.26	3.53	0.001	**
Linguistic					
1PL (we, ...)	A - D	-0.13	-4.88	< 0.001	***
	B - D	-0.13	-3.19	0.003	**
	C - D	-0.13	-3.83	< 0.001	***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 9. Pairwise group comparisons using Dunn-Bonferroni-Holm post-hoc testing for selected LIWC measures of cognition, affect and perception, including median difference between groups (Mdn_{diff}).

	Groups	Mdn $_{diff}$	Z	p $_{holm}$	
Cognition					
All or None	A - C	0.29	3.34	0.003	**
	B - C	0.20	2.99	0.007	**
	A - D	0.24	2.84	0.009	**
	B - D	0.15	2.49	0.019	*
Cause	A - D	0.73	3.55	< 0.001	***
	B - D	1.12	4.32	< 0.001	***
	C - D	0.77	2.74	0.012	*
Tentative	A - D	-0.56	-2.81	0.010	**
	B - D	-1.27	-4.79	< 0.001	***
	C - D	-0.70	-3.01	0.006	**
Certitude	A - C	0.87	3.54	< 0.001	***
	B - C	0.86	3.23	0.002	**
	A - D	0.92	4.23	< 0.001	***
	B - D	0.91	3.92	< 0.001	***
Differentiation	A - D	-0.72	-3.40	0.002	**
	B - D	-0.42	-2.95	0.008	**
Affect					
Affect	A - C	0.71	3.36	0.002	**
	B - C	0.61	2.40	0.025	*
	A - D	1.05	3.89	< 0.001	***
	B - D	0.95	2.93	0.007	**
Neg. Tone	A - C	0.71	3.14	0.004	**
	B - C	0.40	2.19	0.043	*
	A - D	0.71	3.62	< 0.001	***
	B - D	0.41	2.67	0.015	*
Emotion	A - C	1.02	4.17	< 0.001	***
	B - C	0.49	2.55	0.016	*
	A - D	1.07	5.20	< 0.001	***
	B - D	0.55	3.58	< 0.001	***
Neg. Emotion	A - C	0.62	3.13	0.004	**
	A - D	0.60	4.13	< 0.001	***
	B - D	0.46	3.11	0.004	**
Anxiety	A - C	0.18	2.77	0.014	*
	A - D	0.22	3.47	0.002	**
	B - D	0.19	2.57	0.020	*
Perception					
Visual	A - C	-0.15	-2.13	0.033	*
	B - C	-0.17	-2.42	0.031	*
	A - D	-0.31	-4.38	< 0.001	***
	B - D	-0.32	-4.67	< 0.001	***
	C - D	-0.16	-2.25	0.037	*

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$