Increasing Large Language Models Context Awareness through Nonverbal Cues

Matthias Schmidmaier
LMU Munich
Munich, Germany
matt@schmidmaier.org

Cedrik Harrich
LMU Munich
Munich, Germany
cedrik.harrich@outlook.com

Sven Mayer
LMU Munich
Munich, Germany
info@sven-mayer.com

Abstract

Today, interaction with LLM-based agents is mainly based on text or voice interaction. Currently, we explore how nonverbal cues and affective information can augment this interaction in order to create empathic, context-aware agents. For that, we extend user prompts with input from different modalities and varying levels of abstraction. In detail, we investigate the potential of extending the input into LLMs beyond text or voice, similar to human-human interaction in which humans not only rely on the simple text that was uttered by a conversion partner but also on nonverbal cues. As a result, we envision that cameras can pick up facial expressions from the user, which can then be fed into the LLM communication as an additional input channel fostering context awareness. In this work we introduce our application ideas and implementations, preliminary findings, and discuss arising challenges.

CCS Concepts

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

Keywords

LLM, empathy, prompting, nonverbal communication, Prompt Engineering

1 Introduction

Current implementations of Large Language Model (LLM) based agents primarily allow text or voice interaction. While the interpretation of verbal expressions or text-based nonverbal cues like emojis already affects the generated responses, other nonverbal cues like facial expressions or gestures are not captured and processed. As nonverbal communication plays an important role in human interaction, this lack of input modalities limits the interaction with LLM-based systems compared to human in-person conversations. Most existing approaches focus on extracting information from the unaltered user prompt either through additional sentiment analysis or through specific instructions for the model on how to respond empathically or explicitly ask for affective user states. However, recent research proposes to augment the user prompt with additional information such as identity description [3] or textual description of nonverbal cues like body posture [17]. We propose a similar approach and present the initial implementation of our prototype ELLMO, which captures facial expressions to augment the textual user prompt in conversations with an LLM-based agent. We discuss the preliminary findings and challenges, especially regarding prompt creation, model instructions, conversational design, and study design.

Mensch und Computer 2024 – Workshopband, Gesellschaft für Informatik e.V., 01.-04. September 2024, Karlsruhe, Germany

https://doi.org/10.18420/muc2024-mci-ws09-190

2 Related Work

Nonverbal cues play an important role in human interaction. In the following part, we provide a short introduction to nonverbal communication and current research on context-aware LLMs.

2.1 Nonverbal Communication

Argyle [1] describe nonverbal communication as the non-linguistic transmission of information in which a message is encoded, transmitted over a channel, and then decoded with or without the intention or awareness of the sender and receiver. In their model of interpersonal communication, DeVito [4] describes noise, contexts, effects, and ethics as additional elements of nonverbal communication. Nonverbal channels include body movement, facial expressions, eye and gaze movement, touch, spatial behavior, paralanguage but also silence, and use of time or smell [1, 4]. Referring to Ekman and Friesen [5], DeVito [4] further identifies five message types:

(1) emblems: word-like signs with rather specific meaning, such as “thumbs up”.
(2) illustrators: gestures that illustrate (e.g., clarify or emphasize) verbal messages, e.g., pointing into a direction.
(3) affect displays: expressing emotions, for example, through mimics.
(4) regulators: influencing the speaking of another individual to influence a conversation (e.g., leaning toward somebody to express interest).
(5) adaptors: body movements to satisfy certain needs like scratching your head or chewing on a pencil.

In contrast to in-person interaction, computer-mediated communication often limits the available channels, affecting conversation behavior in digital settings [9, 15]. In the same way, interaction with artificial systems like LLM-based agents usually provides limited nonverbal communication. Therefore, current research is exploring how to (1) augment systems' capabilities to perceive and interpret nonverbal cues and (2) allow systems to express nonverbal cues and emotional reactions, for example, in order to simulate empathy [2, 6, 8, 10, 12, 14].

2.2 Context-Aware LLMs

With the development of affective computing, creating empathic systems that react to affective states or nonverbal cues has become a well-known approach [6, 8, 10]. Regarding LLM-based systems, however, context awareness mostly refers to additional instructions on how to interpret the given user prompt or conversational history [16] or on how the system itself could integrate nonverbal cues to its response [7]. Yet, there are also recent approaches that explore on how to add information or nonverbal cues from the user to the textual input. For example, Cuadra et al. [3] explored how identity-based prompting impacts LLM-based conversational agents' empathic behavior. They tested 65 distinct human identities by adding them to predefined user prompts. They found that the agent takes the textual identity context into account yet sometimes generates problematic empathic responses, for example, by showing empathy toward harmful ideologies. Piferi [11] introduce a system design that connects emotion recognition from voice analysis to an LLM in order to influence its responses. Wicke [17] go one step further by proposing to supplement LLM prompts with human body postures without prior affective processing, leaving the interpretation of nonverbal cues to the LLM. They present a pipeline where postures are translated to textual descriptions (e.g., "hand pushed down and away from the body with sharp stop") and then added to the LLM prompt.

3 Prototype: ELLMO

To conduct initial user studies, we implemented a first prototype called ELLMO (Empathic LLM Optimization), with the goal of augmenting user prompts through facial expressions in order to create empathic model responses.

3.1 Architecture

We implemented ELLMO as a web application with a VUE.js frontend and a Django backend that, in turn, accesses OpenAI's Assistant API to process user prompts. This multi-component approach provides independence between the frontend and LLM access, allowing us to easily adapt the system for different study designs. Figure 1 provides an overview of the current architecture and the processing pipeline. The web frontend captures the user's text input as well as their facial expressions through Google Mediapipe[1] and sends it to the Django backend. Currently, both the text prompt and the facial information are sent together when the user submits their request. Continuous or asynchronous transmission of the nonverbal cues is discussed in Section 4.2. The backend is responsible for data pre-processing, logging, and prompt transmission to the OpenAI API. In future implementations, the backend would also be responsible for managing conversational flow, for example, by routing prompts to different assistant instances.

3.2 Assistant Instructions

Current LLM APIs offer different possibilities to define system roles and instructions. For our latest ELLMO implementation, we used OpenAI's Assistant API, which allows to define model instances with specific instructions on how to process user prompts. We defined several requirements for ELLMO's behavior:

(1) general role description: ELLMO should be an empathic assistant
(2) technical instructions on how to process nonverbal cues - in our case: facial landmarks
(3) "muting the inner monologue": ELLMO should not repeat or talk about the nonverbal cues in detail in its responses.

We addressed these requirements in the ELLMO assistant instructions for processing facial landmarks:

> "You are an empathic assistant. Your name is Ellmo. You help users solve their problems by taking their emotional needs into account. The user input might contain a section <faceBlendShapes> with facial blend shapes detected by Google's MediaPipe. That section contains 52 facial feature categories, each with an intensity score from 0 to 1. Use this information to imagine the user's facial expression and take into account what that would mean in terms of nonverbal communication and affective states. Use that facial expression to adapt your response. Do not mention facial blend shapes in detail. Avoid sentences like "It seems like based on the facial blend shapes detected..."

4 Findings and Challenges

We conducted several preliminary tests with OpenAI's ChatGPT through manual input of affective metadata and test runs with our ELLMO prototype. In the following, we discuss the initial findings and challenges which we aim to address in future research.

4.1 Facial Feature Interpretation

One of our most promising findings so far is that ChatGPT demonstrates a certain level of proficiency in analyzing facial expressions from textual data such as facial landmarks or blend shapes. In the first test, we explored ChatGPT's capability to interpret emotion categories from facial expressions. As input basis, we selected images of faces that represent six basic emotion categories (fear, contempt, disgust, sadness, anger, happiness, and surprise). We then used Google MediaPipe to get the facial blend shape vector for each image, consisting of 52 shape descriptions such as *browDownLeft*, *eyeSquintRight* or *mouthStretchLeft* with intensities between 0..1. We then attached this JSON formatted output to manual user prompts and added prompt instructions in the instructions. We found that ChatGPT was able to respond and interpret to that input, for example, it concluded that higher corners of the mouth could indicate happiness. We also conducted test runs where we directly added emotion categories or numeric valence/arousal values to user prompts and instructed ChatGPT to take that emotional user state into account in its responses. In a comparative test, we found that the facial blend shapes input resulted in response behavior similar to that of direct context augmentation. In conclusion, we found that ChatGPT, in general, is capable of interpreting facial expressions from JSON formatted blend shape descriptions and processing them the same way as it is able to use emotional context from direct textual description (emotion categories) or more abstract affective

[1] https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

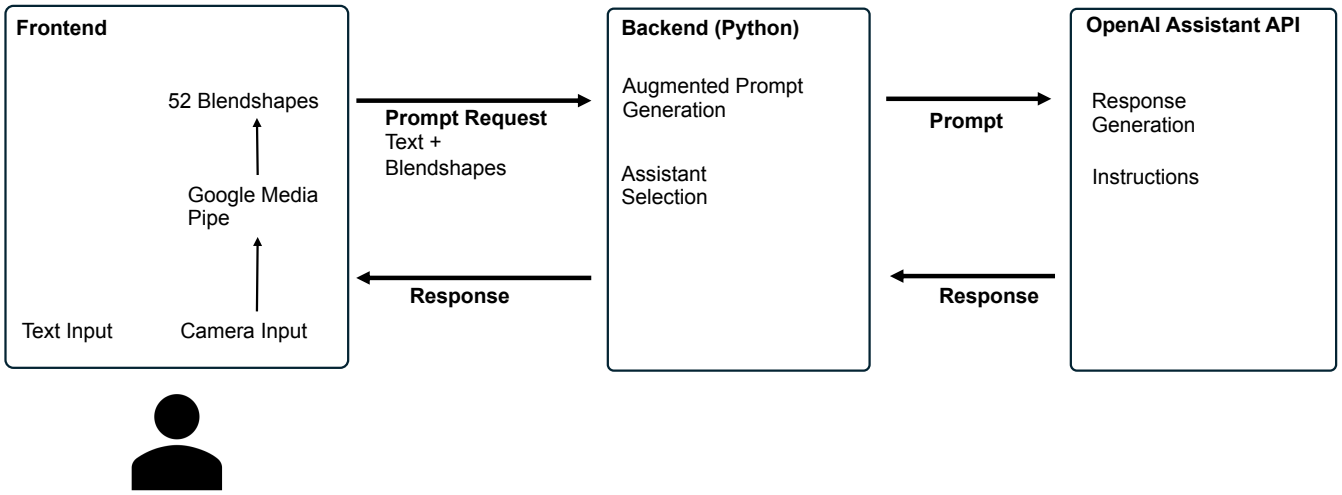

Figure 1: ELLMO system architecture. The user text input is augmented with facial expressions (52 blend shapes) captured through Google MediaPipe and sent to OpenAI's Assistant API via a Python (Django) backend in order to create empathic responses. The backend also provides the possibility to select different assistant instances.

dimensions (valence/arousal). Still, we need to conduct in-depth evaluations regarding emotional variety and response consistency.

4.2 Conversational Design

For a current test run, we captured the user's facial expression over time, but only processed the snapshot at the time of the user prompt submission. While this one-shot sampling enables a simple prompt extension and processing on the assistant side, a meaningful implementation of nonverbal cues likely requires continuous sampling. We suggest to explore if facial expressions should be sent to the LLM at regular intervals, or as aggregated, shorter time series assigned to specific conversation phases - for example synchronized with turn taking. In theory that would enable the model to interpret the user's reactions while writing a prompt, as well as their reactions while reading the model's responses. Yet, such an approach will eventually result in longer processing times and require more complex model instructions: for example, if the assistant should immediately respond to nonverbal prompts that come without manual text input. Another approach would be to pre-process visual input and summarize it into textual descriptions - similar to the approach described by Wicke [17].

Regarding conversational flow, Seo et al. [13] provides a good example on how to integrate "thematic checkpoints" to ensure that the agent has received certain information or covered specific topics. As OpenAI's Assistant API allows to switch between assistants during a conversational thread, we consider to define different assistants to handle different conversational stages, as well as different prompt types like mentioned above. To address response variability of LLMs, Cuadra et al. [3] suggest to repeat prompts until thematic saturation is reached. While this is a feasible approach for repeatable lab settings, we still explore on how to achieve consistent LLM behavior and as a result comparable study results in in-the-field studies.

4.3 User Awareness

As described in Section 2.1, the sending and receiving of nonverbal cues can occur both consciously and unconsciously and with varying degrees of control. Further, related work (see Section 2.2) and our own experience shows, user interaction with artificial systems differs from human communciation in terms of using nonverbal cues. To assure engaged and context related interaction, we suggest providing (1) transparency to the user regarding which cues the system is processing, and (2) eventually offering feedback on these cues. The question arises as to what extent this feedback should be incorporated into the model's (textual) responses and whether there should be dedicated feedback mechanisms, such as mirroring approaches or continuous visualizations of how the system perceives the user's nonverbal cues. While we strongly suggest to provide some sort of feedback, we argue that the "default" inner monologue feedback as depicted in Figure 2 has to be reduced through corresponding instruction design. Additionally, a continuous visualization of the artificial system state as reaction to the user's state might be useful in enhancing the conversation and mimicing empathic system behavior.

4.4 Study Design

One of the main challenges in exploring nonverbal interaction, affective responses and empathic behavior is to create valid study scenarios [1]. Explicit, intentional reactions (nonverbal emblems or illustrators) like expressing agreement or skepticism through certain gestures can be promoted through detailed participant instructions and corresponding conversational tasks. However, it is more difficult to trigger implicit reactions and emotions in study participants. Our current goal is to conduct either (1) controlled lab studies with personal conversational scenarios such as "Use ELLMO to talk about your problems at work" or (2) more longitudinal in-the-field studies, where we ask participants to regularly use the assistant on a personal device. Another question is how to measure

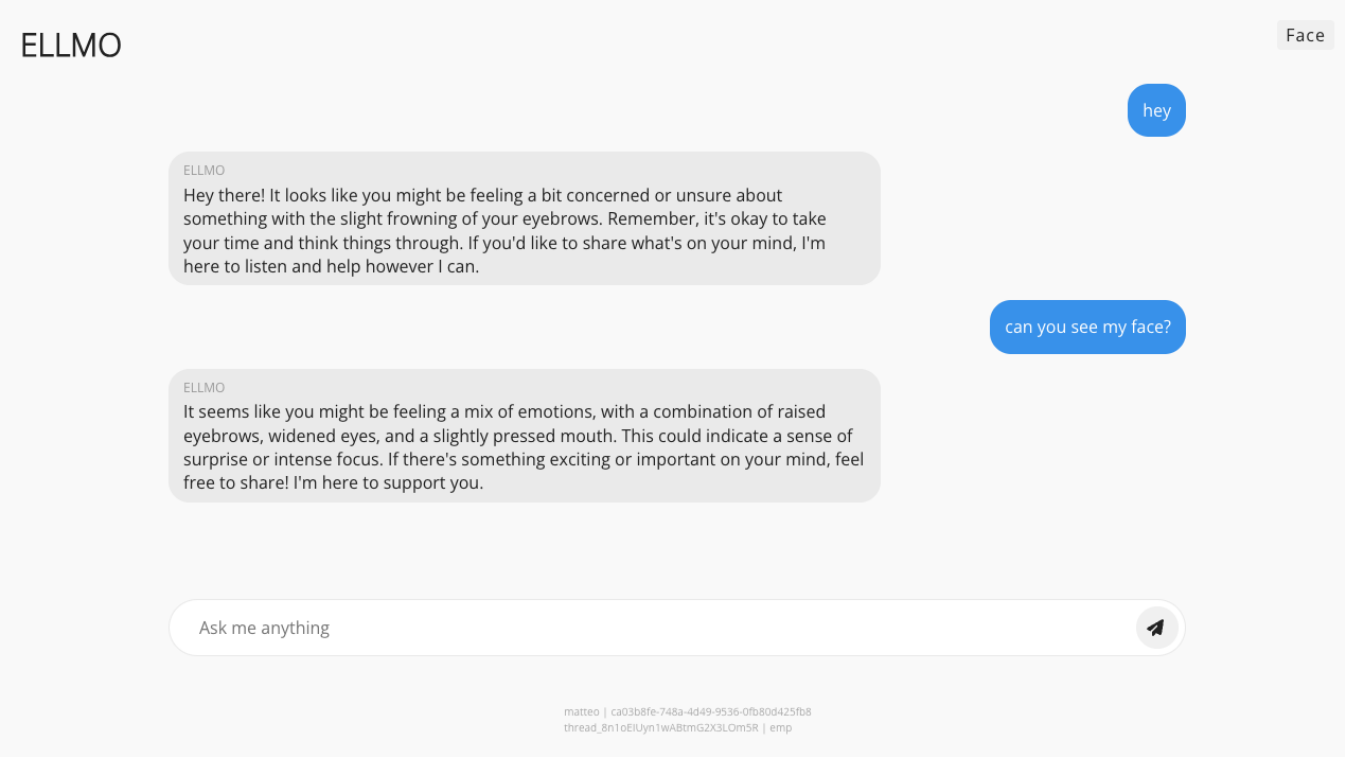

Figure 2: ELLMO user interface, showing two responses with extensive feedback of the perceived facial expression context.

the effects of a conversational interaction. While perceived system or human behavior might also be explored by letting participants observe recorded scenarios from a third-person perspective, we aim for first-person participant experiences and ratings, as aspects such as perceived empathy of a system or experienced emotions are based on inner, affective perceptions. So while Cuadra et al. [3] for example let participants rate generated conversations between a model and a fictional user, we want to capture participants' own experiences, using scientific scales or conclusive interviews.

4.5 Nonverbal Channels

Similar to the approach proposed by Wicke [17], we want to augment textual or spoken prompts with input from other nonverbal channels. While our current prototype, ELLMO, covers facial expressions, we also consider other modalities for future applications. Especially interaction timing, interruption or silence can also offer interesting possibilities for conversational communication. Finally, we plan to explore also new forms of nonverbal communication, especially designed for interaction with artificial agents. This might include also new forms of cues and transmission, such as manually expressing affective states or reactions to LLM responses or processing times via user interface. That might also require to conduct preliminary research on what kind of nonverbal effects should be conveyed in a conversation with a LLM based agent: while level of agreement or satisfaction could be helpful as reaction on system responses, cues that communicate expected response length or detail could accompany user input to optimize responses.

5 Conclusion

In this paper, we introduced our approach on how to extend the contextual awareness of LLM based agents, especially regarding nonverbal cues and affective states. With the ubiquitous use and ongoing integration of conversational agents in all sorts of daily tasks we see promising applications for such agents with augmented nonverbal input: empathic agents are already developed for mental health applications, personal reflection or social robots. We discussed several preliminary findings and challenges, which would benefit from being discussed in the community.

References

[1] Michael Argyle. 1988. *Bodily Communication, 2nd Edition*. Vol. 2. Methuen & Co Ltd, New York, NY, US. 363 pages.

[2] Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer 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. https://doi.org/10.1145/3411764.3445318

[3] 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. https://doi.org/10.1145/3613904.3642336

[4] Joseph A DeVito. 2007. *The interpersonal communication book*. Pearson/Allyn and Bacon, Boston, MA.

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

[6] 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. https://doi.org/10.1145/3411763.3451660

[7] Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. 2023. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models. arXiv:2308.16529 [cs.RO] https://arxiv.org/abs/2308.16529

[8] 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. https://doi.org/10.1145/1877826.1877839

[9] 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5271–5283. https://doi.org/10.1145/3025453.3025548

[10] Sung Park and Mincheol Whang. 2022. Empathy in Human–Robot Interaction: Designing for Social Robots. *International Journal of Environmental Research and Public Health* 19, 3 (Feb. 2022), 21 pages. https://doi.org/10.3390/ijerph19031889

[11] Francesco Piferi. 2022. *CHATCARE: an emotional-aware conversational agent for assisted therapy*. Master's thesis. Polytechnic University of Milan. https://www.politesi.polimi.it/handle/10589/218567

[12] 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 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 456, 18 pages. https://doi.org/10.1145/3613904.3642035

[13] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. 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 903, 20 pages. https://doi.org/10.1145/3613904.3642152

[14] Daniel Ullrich, Sarah Diefenbach, and Andreas Butz. 2016. Murphy Miserable Robot: A Companion to Support Children's Well-being in Emotionally Difficult Situations. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3234–3240. https://doi.org/10.1145/2851581.2892409

[15] Joseph B Walther, Tracy Loh, and Laura Granka. 2005. Let Me Count the Ways: The Interchange of Verbal and Nonverbal Cues in Computer-Mediated and Face-to-Face Affinity. *J. Lang. Soc. Psychol.* 24, 1 (March 2005), 36–65. https://doi.org/10.1177/0261927X04273036

[16] Anuradha Welivita and Pearl Pu. 2024. Is ChatGPT More Empathetic than Humans? arXiv:2403.05572 [cs.HC] https://arxiv.org/abs/2403.05572

[17] Philipp Wicke. 2024. Probing Language Models' Gesture Understanding for Enhanced Human-AI Interaction. arXiv:2401.17858 [cs.CL] https://arxiv.org/abs/2401.17858