# Generative Facial Expressions and Eye Gaze Behavior from Prompts for Multi-Human-Robot Interaction

**Conference Paper** · October 2023

DOI: 10.1145/3586182.3616623



**3 authors**, including:

Virgil Barnard
Riverside Research
**12** PUBLICATIONS **5** CITATIONS
SEE PROFILE

Joey Salisbury
Riverside Research
**45** PUBLICATIONS **859** CITATIONS
SEE PROFILE

# Generative Facial Expressions and Eye Gaze Behavior from Prompts for Multi-Human-Robot Interaction

Gabriel J. Serfaty
University of Michigan
gserfaty@umich.edu

Virgil O. Barnard IV
Riverside Research
vbarnard@riversideresearch.org

Joseph P. Salisbury
Riverside Research
jsalisbury@riversideresearch.org

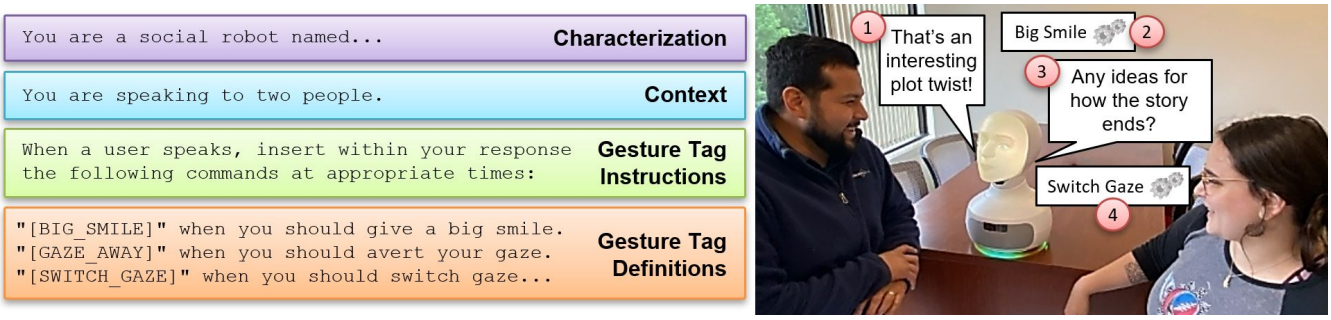


**Figure 1: An initial prompt defining facial expression and gaze behaviors can allow an LLM to generate relevant responses.**

## ABSTRACT

Nonverbal cues such as eye gaze and facial expressions play critical roles in conveying intent, regulating conversation, and fostering engagement. A robot's ability to effectively deploy these behaviors can significantly enhance human-robot collaboration. We describe a simple zero-shot learning approach to generate facial expression and gaze shifting behaviors to control a social robot conversing with an individual or group. An initial prompt provides instructions to a pre-trained large language model on how the model can control a robot's facial expression and eye gaze behaviors during a conversation. To demonstrate this, we describe a proof-of-concept implementation using the robot Furhat. This simple and easily customizable approach can be used to improve perception of a robot's social presence in multi-human-robot interactions.

## CCS CONCEPTS

• **Human-centered computing** → Interface design prototyping; Natural language interfaces; Collaborative interaction; • **Computing methodologies** → Cognitive robotics; Discourse, dialogue and pragmatics; Theory of mind.

## KEYWORDS

Social robotics, gesture synthesis, gaze alternation, multi-human-robot interaction, pre-trained language models, nonverbal communication, prompt engineering


*UIST '23 Adjunct, October 29–November 01, 2023, San Francisco, CA, USA*

ACM ISBN 979-8-4007-0096-5/23/10.
https://doi.org/10.1145/3586182.3616623

**ACM Reference Format:**
Gabriel J. Serfaty, Virgil O. Barnard IV, and Joseph P. Salisbury. 2023. Generative Facial Expressions and Eye Gaze Behavior from Prompts for Multi-Human-Robot Interaction. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3586182.3616623

## 1 INTRODUCTION

In communication, nonverbal cues such as facial expressions and eye gaze are integral to conveying intent, regulating conversation flow, fostering engagement, and establishing and maintaining trust [1–4]. In human-robot interaction (HRI), the ability of a robot to emulate and utilize these nonverbal cues appropriately can significantly enhance the quality of interactions, bolster collaborative efforts, and improve user perceptions of the robot [5], [6]. In multi-human-robot interactions, even more nuanced behavior from robots is required to effectively manage attention, maintain a sense of inclusion among participants, and adapt to the dynamic nature of group interactions. By exhibiting facial expressions and gaze behaviors that humans naturally expect and understand, robots can facilitate a smoother, more natural interaction, enhancing collaboration and trust among all participants [7], [8].

This paper describes a zero-shot learning approach to generate facial expression and gaze behaviors using large language models (LLMs). Using a single prompt, we instruct an LLM to insert gaze behaviors and facial expressions into conversation responses using tags that trigger a social robot's gestures while it speaks (Figure 1). This technique provides an easily customizable approach to generate nonverbal behaviors in multiparty interactions.

## 2 RELATED WORK

The rise of LLMs with zero-shot learning capabilities like GPT [9], [10] has opened new avenues in HRI. Given their ability to

produce human-like text, immediate applications include driving social robots' conversational capabilities [11]. A noted limitation of this demonstration was that the robot's actions beyond speech were not integrated (e.g., GPT-3 may produce "I will clap my hands!" but there is no accompanying action), a capability we demonstrate here. More specifically, we propose LLMs can generate socially appropriate nonverbal behaviors during conversation. In humans, understanding nonverbal cues is related to Theory of Mind (ToM), the ability to impute unobservable mental states to others. It has been suggested that LLMs have spontaneously developed ToM-like capabilities as a byproduct of their improving language skills [12]. Supporting this theory, LLM-based human models have been successfully integrated into a social robots' planning process [13]. LLMs have been used to aid in speaker diarization for facilitating multiparty interaction with a social robot, demonstrating their ability to leverage lexical cues [14]. LLMs have also been used to recognize emotions in dialogue to allow human affective states to drive rewards in a reinforcement learning paradigm to support a social robot [15]. In contrast, our approach uses LLMs trained on a single prompt to directly control a social robot's actions.

# 3 PROTOTYPE IMPLEMENTATION

## 3.1 Prompt Engineering with Gesture Tags

A prompt [16] is a set of instructions provided to an LLM that programs the LLM by influencing generated output. In particular, a prompt sets the context for the conversation and tells the LLM what information is important and what the desired output form and content should be [17]. To design a simple prompt to guide eye gaze and gestures during conversation, we specified four distinct parts:

- **Characterization –** Description of role or perspective the LLM should take when replying. Example: *Pretend you are speaking as a Furhat robot as part of a demo on natural, intuitive human-robot interaction. Your name is RITA, the Riverside Research Intelligent Task Analyst. Introduce yourself and offer your services.*
- **Context** – Description of scene (e.g., from computer vision). Example: *You are speaking to two people.*
- **Gesture Tag Instructions** – Instructions for inserting gesture tags. Example: *Insert within your response the following commands at the appropriate times:*
- **Gesture Tag Definitions** – A list of gestures available and suggested use. Example: *"[SWITCH_GAZE]" when it would be appropriate to change up who you are looking at. "[GAZE_AWAY]" when it would be appropriate to break eye contact. "[BIG_SMILE]" when it would be appropriate to show a big smile.*

## 3.2 Prompt Engineering with Gesture Tags

To demonstrate this approach to generating nonverbal cues in an embodied conversational agent, we used the robot Furhat [18]. The Python implementation [19] of the Furhat Remote API [20] was used to program the robot. User detection was accomplished using the native Furhat SDK, which detects users within view of the on-device camera and assigns them a unique identifier that is kept constant as long as the user is in view. Upon program execution, we poll the number of users detected until it is non-zero, triggering the initial prompt (as described above) to be sent to the Open AI API [21]. The OpenAI Completion class was used to create a response using the Davinci-3 variant of GPT-3.5 [22], 2048 maximum tokens, and a temperature of 0.7. The returned response was parsed into a sequence of sentences intended for verbalization and embedded gesture tags. Speech was generated from portions of the response text intended for verbalization using Microsoft Azure Neural voices [23]. Gesture tags triggered Furhat to complete either a shift in gaze or a nonverbal gesture. Upon completion of the response, Furhat was set to listen for a user's speech. Speech recognition was accomplished using Google Cloud Speech-to-Text [24]. The recognized speech was appended to the conversation history, including the initial prompt, before querying the OpenAI API again. This interaction cycle repeats until no users are detected, upon which a "good-bye" prompt is appended to the conversation, triggering a final response from Furhat. Listening for user speech is terminated until users are detected via the camera, restarting the interaction.

## 3.3 Prompt Engineering with Gesture Tags

The proof-of-concept generated promising results, enabling the robot's expressions to be triggered appropriately, particularly when asked to perform a specific gesture. Occasionally, the LLM returned not only its response, but the predicted response of the user. This was easily parsed out. Additional refinements could evaluate different LLMs, their parameters, and customizing prompts to cater to the unique requirements of different conversational contexts. We were unable to incorporate certain contextual features based on limitations of the Remote API. For example, at the time of writing, the Remote API limits the number of users detected to two. Furhat's native Kotlin SDK also supports speaker diarization, which would enable the LLM to have context regarding who said what in a multi-person conversation. The Kotlin SDK also provides speaker gaze direction (e.g., another speaker, location, or Furhat) and speaker gesture recognition (e.g., is the user smiling?).

Future experiments will evaluate to what extent naive users perceive the generated behavior is socially appropriate, as well as determine how LLM-generated behaviors may improve performance on collaborative tasks that require both verbal and nonverbal communication [25].

# 4 CONCLUSION

We describe a simple but powerful approach to generate nonverbal cues during multi-human-robot interactions. While not experimentally validated, the addition of LLM-generated nonverbal behaviors produces a perceptible improvement in the robot's social presence. This technique provides an easily customizable approach to generate robot nonverbal behaviors, which could be helpful in rapid prototyping of multi-human-robot interactions.

# ACKNOWLEDGMENTS

Supported by Independent Research and Development funds from Riverside Research. The authors thank Furhat Robotics for technical support.

# REFERENCES

[1] T. Maran, M. Furtner, S. Liegl, S. Kraus, and P. Sachse, "In the eye of a leader: Eye-directed gazing shapes perceptions of leaders' charisma," The Leadership Quarterly, vol. 30, no. 6, p. 101337, Dec. 2019, doi: 10.1016/j.leaqua.2019.101337.

[2] E. Prochazkova, L. Prochazkova, M. R. Giffin, H. S. Scholte, C. K. W. De Dreu, and M. E. Kret, "Pupil mimicry promotes trust through the theory-of-mind network," Proceedings of the National Academy of Sciences, vol. 115, no. 31, pp. E7265–E7274, Jul. 2018, doi: 10.1073/pnas.1803916115.

[3] C. L. Kleinke, "Gaze and eye contact: A research review," Psychological Bulletin, vol. 100, no. 1, pp. 78–100, 1986, doi: 10.1037/0033-2909.100.1.78.

[4] S.-H. Shim, R. W. Livingston, K. W. Phillips, and S. S. K. Lam, "The Impact of Leader Eye Gaze on Disparity in Member Influence: Implications for Process and Performance in Diverse Groups," AMJ, vol. 64, no. 6, pp. 1873–1900, Dec. 2021, doi: 10.5465/amj.2017.1507.

[5] H. Admoni and B. Scassellati, "Social Eye Gaze in Human-Robot Interaction: A Review," Journal of Human-Robot Interaction, vol. 6, no. 1, p. 25, Mar. 2017, doi: 10.5898/JHRI.6.1.Admoni.

[6] B. Mutlu, J. Forlizzi, and J. Hodgins, "A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior," in 2006 6th IEEE-RAS International Conference on Humanoid Robots, University of Genova, Genova, Italy: IEEE, Dec. 2006, pp. 518–523. doi: 10.1109/ICHR.2006.321322.

[7] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," Artificial Intelligence, vol. 166, no. 1–2, pp. 140–164, Aug. 2005, doi: 10.1016/j.artint.2005.03.005.

[8] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing In Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues".

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[11] E. Billing, J. Rosén, and M. Lamb, "Language Models for Human-Robot Interaction," in Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm Sweden: ACM, Mar. 2023, pp. 905–906. doi: 10.1145/3568294.3580040.

[12] M. Kosinski, "Theory of Mind May Have Spontaneously Emerged in Large Language Models," arXiv.org, Feb. 04, 2023. https://arxiv.org/abs/2302.02083v3 (accessed Jun. 19, 2023).

[13] B. Zhang and H. Soh, "Large Language Models as Zero-Shot Human Models for Human-Robot Interaction".

[14] P. Murali, I. Steenstra, H. S. Yun, A. Shamekhi, and T. Bickmore, "Improving Multiparty Interactions with a Robot Using Large Language Models," in Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, in CHI EA '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–8. doi: 10.1145/3544549.3585602.

[15] B. Xie and C. H. Park, "A MultiModal Social Robot Toward Personalized Emotion Interaction." arXiv, Oct. 07, 2021. Accessed: May 25, 2023. [Online]. Available: http://arxiv.org/abs/2110.05186

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Comput. Surv., vol. 55, no. 9, pp. 1–35, Sep. 2023, doi: 10.1145/3560815.

[17] J. White *et al.*, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT." arXiv, Feb. 21, 2023. Accessed: Jun. 19, 2023. [Online]. Available: http://arxiv.org/abs/2302.11382

[18] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction," in Cognitive Behavioural Systems, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 114–130. doi: 10.1007/978-3-642-34584-5_9.

[19] "furhat-remote-api: Furhat Remote API."

[20] "Remote API - Furhat Developer Docs." https://docs.furhat.io/remote-api/ (accessed Jun. 19, 2023).

[21] OpenAI, "openai: Python client library for the OpenAI API." Accessed: Jun. 19, 2023. [OS Independent]. Available: https://github.com/openai/openai-python

[22] "OpenAI API." https://platform.openai.com/docs/ (accessed Jun. 19, 2023).

[23] "Azure OpenAI Service – Advanced Language Models | Microsoft Azure." https://azure.microsoft.com/en-us/products/cognitive-services/openai-service (accessed Apr. 04, 2023).

[24] "Speech-to-Text: Automatic Speech Recognition," Google Cloud. https://cloud.google.com/speech-to-text (accessed Jun. 19, 2023).

[25] M. J. Munje, L. K. Teran, B. Thymes, and J. P. Salisbury, "TEAM3 Challenge: Tasks for Multi-Human and Multi-Robot Collaboration with Voice and Gestures," in Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, in HRI '23. New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 91–96. doi: 10.1145/3568294.3580049.