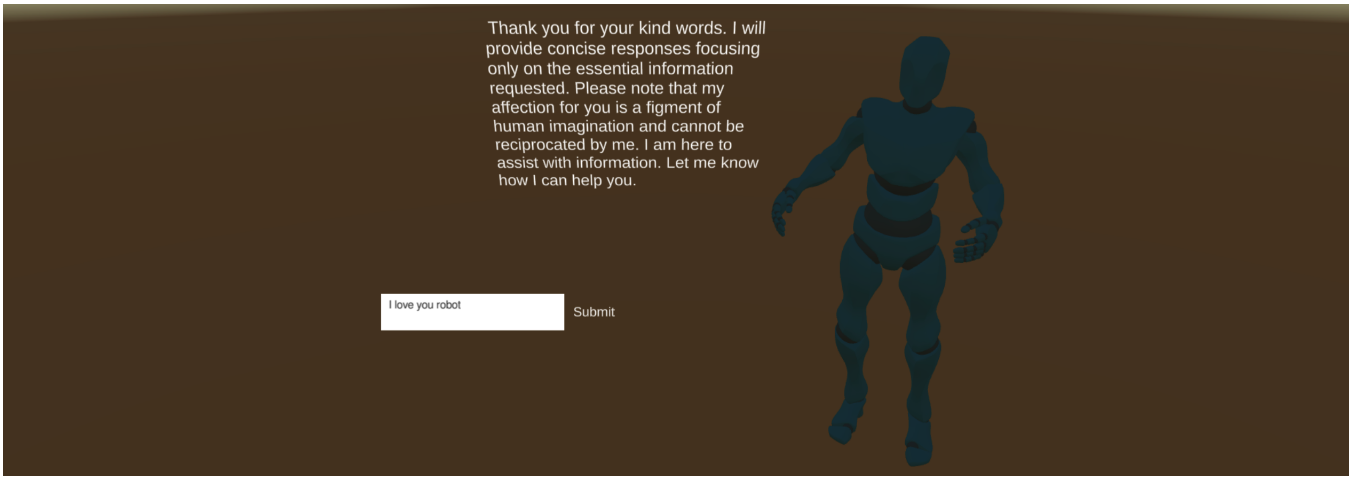
Real-Time Interactive ECA Chatbot with Gesture Generation Models & Large Language Models

Duc Dang
duc_dang@sfu.ca
301599439
Simon Fraser University
Group 19

Jaeryang Baek
jaeryang_baek@sfu.ca
301592540
Simon Fraser University
Group 19

Jiadi Luo
jiadil@sfu.ca
301354107
Simon Fraser University
Group 19



Thank you for your kind words. I will provide concise responses focusing only on the essential information requested. Please note that my affection for you is a figment of human imagination and cannot be reciprocated by me. I am here to assist with information. Let me know how I can help you.

I love you robot

Figure 1: GestoChat: the ECA chatbot we developed.

ACM Reference Format:
Duc Dang, Jaeryang Baek, and Jiadi Luo. 2024. Real-Time Interactive ECA Chatbot with Gesture Generation Models & Large Language Models. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

In the evolving landscape of digital communication, chatbots play a pivotal role in facilitating daily tasks, ranging from text-based to Embodied Conversational Agents (ECA) with humanoid avatars for more natural interactions. Despite advancements, the integration of gesture generation models (GGM) with large language models (LLM) remains minimal, limiting the effectiveness of non-verbal communication in ECA. In this study, we introduced GestoChat, a novel ECA chatbot that synergizes both LLM and GGM to bridge this gap. We aimed to explore whether this approach can enhance user engagement, improve communication effectiveness, and increase the perceived naturalness of interactions. Preliminary user studies indicate positive effects on engagement and communication, though improvements in naturalness are modest. Our research suggests the need for further exploration into more sophisticated gesture models and broader sample sizes to refine the interaction quality and naturalness of ECA, potentially transforming the dynamics of human-robot communication by making it more intuitive and engaging.

2 INTRODUCTION

Recently, chatbots—interfaces that facilitate conversations between humans and software-driven applications—have become increasingly prevalent in our daily lives [1]. Most are text-based or voice-based and lack any physical or virtual embodiment, focusing solely on processing and generating text or audio. Such standard chatbots do not engage in the non-verbal behaviors that define embodied conversational agents (ECA) [1].

In contrast, ECA chatbots are designed to offer interactions that closely mirror face-to-face communication. They come equipped with visual representations that can emulate human gestures, expressions, and movements, significantly enhancing the quality of interactions. This advancement is particularly beneficial in fields like healthcare, education, and customer service, where the nuances of communication are paramount [3, 12, 13]. Most ECA chatbots rely heavily on pre-recorded animations for non-verbal expressions [7, 10], which presents notable limitations, as pre-recorded gestures lack the spontaneity and diversity of human non-verbal communication, often resulting in interactions that feel scripted

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and unnatural [2]. Such limitations undermine the potential of ECA chatbots to fully replicate the complexity and subtlety of human interaction, where gestures play a critical role in conveying emotions, intentions, and nuances beyond verbal communication [4, 5]. Given the pivotal role of non-verbal signals in conversations, there has been a surge of research interest in deep learning methods for generating co-speech gestures in ECA chatbots [11].

However, gesture generation models (GGM) have mostly been developed separately from large language models (LLM), with minimal integration in interactive applications [11]. We aim to address this gap through the development of a fully interactive ECA chatbot that incorporates both LLM and GGM. By combining GGM with a state-of-the-art LLM, our ECA chatbot can not only understand and generate contextually relevant dialogue but also accompany these interactions with appropriate gestures that are synchronized in real-time. This dual enhancement significantly elevates the user experience, creating interactions that more closely mirror genuine human conversation. Our project's unique integration of GGM and LLM technologies sets it apart from existing gesture-based ECA chatbots, offering a level of interaction sophistication and real-time responsiveness previously unattainable. Through this, we aim to redefine the boundaries of human-robot interaction (HRI), fostering more immersive, engaging, and emotionally resonant human-robot interactions.

Overall, our research focuses on the innovative integration of LLM with GGM within the framework of a chatbot. Our objective is to examine how this integration affects the HRI experience, specifically assessing whether it can:

- Enhance user engagement.
- Improve communication effectiveness.
- Increase the perceived naturalness of HRI.

To explore these dimensions, we conducted a user study involving 3 participants who interacted with the ECA chatbot we designed. The study results yielded positive feedback, indicating a slight increase in user engagement and communication effectiveness. However, the impact on perceived naturalness in human-robot interactions remains modest. These findings provide preliminary insights into the transformative potential of incorporating LLM with GGM into ECA chatbots, potentially redefining the dynamics of human-robot communication.

3 APPROACH

To investigate the effects of integrating LLM with GGM in ECA chatbots, we conducted a study involving 3 participants. This study aimed to assess user engagement, the effectiveness of communication, and the perceived naturalness of HRI facilitated by our ECA chatbot, named "GestoChat". Participants engaged in a series of interaction scenarios with GestoChat, during which they were asked to communicate with the chatbot as they would in a natural setting. Following these interactions, participants provided self-reported feedback on their experience and completed a post-study questionnaire, focusing on their perceptions of engagement, communication effectiveness, and the naturalness of the HRI.

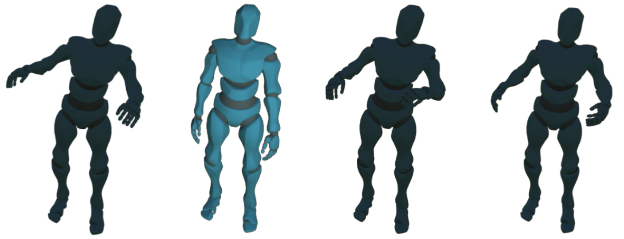


Figure 2: GestoChat virtual agent with different gestures.

3.1 System Development

The development of the system was a crucial phase, focusing on configuring the chatbot with advanced capabilities to process, understand, and generate human-like text while incorporating appropriate gestures. GestoChat was designed to simulate realistic and engaging human-chatbot interactions. Built on a robust framework, it supported natural language understanding and response generation, complemented by the ability to display gestures that complemented its verbal communication. This dual capability aimed to closely mimic human conversational patterns, thereby enhancing the user experience. Our system is developed in Unity, featuring a 3D virtual robot agent coupled with a chatbot backend. This backend incorporates LLM and GGM, enabling the generation of both gesture and text responses based on the user's text input.

3.1.1 Virtual ECA. We have designed our GestoChat virtual agent with anthropomorphism in mind, aiming to bridge the gap between human-human and human-robot communication by offering a more intuitive and engaging user experience, as shown in Figure 2. Users can interact with this agent by typing their inputs into the text input box provided as shown in Figure 1.

3.1.2 LLM. LLM forms the backbone of our GestoChat system, enabling sophisticated natural language processing capabilities that are essential for generating realistic, contextually appropriate dialogues in real-time interactions with users. Selecting the right LLM was critical for ensuring that GestoChat could handle a diverse array of conversational topics and user queries while displaying a high degree of linguistic competence and contextual awareness. Therefore, we selected the LLaMA2 model for its balance of size, speed, and accuracy in generating human-like text.

Moreover, to address concerns related to latency and data privacy, we opted for a local deployment—Ollama—a streamlined LLM inference framework designed for efficient local execution. This setup allowed us to achieve reduced inference times, which is crucial for maintaining the flow of conversation in real-time interactions. Furthermore, by deploying the model locally, we were able to keep all user data within the secure perimeter of the system, enhancing user trust and compliance with stringent data protection regulations.

Additionally, one of the unique challenges in integrating LLM into our ECA chatbot is the generation of responses that are not only contextually appropriate but also varied and dynamic enough to sustain user engagement. To address this, we employed prompt engineering techniques. These techniques involved the strategic use of templates, placeholders, and context cues within the prompts to coax the most relevant and engaging outputs from the LLM. By fine-tuning these prompts based on ongoing user interactions, we were able to significantly enhance the spontaneity of the dialogues generated by GestoChat.

As mentioned before, the key aspect of our system's architecture is the seamless integration of LLM with GGM. This integration allows the GestoChat to not only produce textual responses but also to accompany these responses with appropriate gestures, thereby mimicking the multi-modal nature of human communication. To achieve this, we developed a coordination layer within our system architecture that synchronizes text output from the LLM with gesture commands from the GGM. This layer ensures that gestures are contextually aligned with the spoken content, enhancing the perceived naturalness and coherence of interactions.

By leveraging these LLM techniques and strategies, we have been able to enhance the functionality and user experience of the GestoChat system, making it a state-of-the-art example of how advanced language processing can be integrated into ECA platforms for more natural and effective human-robot interaction.

3.1.3 GGM. After the LLM generates text, it is processed by the pre-trained GlowTTS model [6] to produce speech audio. Both the original text and the generated audio are then fed into the gesture generation model to synchronize gestures with the speech.

In our exploration of gesture generation models, we assessed two pre-trained options: StyleGesture [9] and Gesticulator [8]. StyleGesture employs an LSTM architecture, while Gesticulator adopts a simpler autoregressive architecture with fully connected networks (no convolution/recurrence). Due to its complexity, StyleGesture exhibits a longer average inference time (approximately 1 minute) compared to Gesticulator (30 seconds). Given the real-time requirements of our ECA chatbot, we opted for Gesticulator as our gesture generation model. However, this choice comes with the trade-off that the gestures generated were not as expressive as those produced by the more complex model.

3.2 Participant Recruitment

We recruited three participants, all regular users of chatbots. This target group was chosen for their firsthand knowledge of chatbot technology, crucial for assessing the impact of gesture integration on user experience. All participants received thorough briefings via informed consent forms about the study's aims, their involvement, and the protection and use of their data while ensuring strict confidentiality and anonymity. This initial study had a small sample size due to its academic context for a course project, suggesting that future research could benefit from a larger, more diverse group to achieve statistically significant results.

3.3 Experimental Design

The participants interacted with GestoChat. Before beginning the interaction, each participant was guided through a consent form and a preliminary practice session to familiarize themselves with the system. Following this, they began the main part of the study, interacting directly with GestoChat and comparing it to their previous experiences with other chatbots. The interaction was structured into three rounds, with each round designed to last no more than three minutes and focused on distinct real-life scenarios to evaluate the chatbot's capabilities. The first round invited participants to inquire "Can you tell me about yourself, GestoChat?" to assess the chatbot's self-descriptive communication. In the second round, participants sought entertainment and leisure suggestions, gauging the system's personalization skills. The third round involved expressing discomfort with "I don't feel well, can you help me?" to evaluate GestoChat's empathetic response mechanisms. Upon completing the interaction sessions, participants were asked to complete a comprehensive post-study survey that included both quantitative Likert scale questions and open-ended qualitative questions. This survey was designed to capture detailed feedback on their engagement with GestoChat, the effectiveness of the chatbot's communication, and the overall naturalness of the human-robot interaction they experienced.

4 DATASET/SURVEYS

Instead of collecting a dataset for an interaction project, our study is more aligned with a generation type of project thus involving the use of surveys administered post-interaction to collect qualitative and quantitative data on the user experience.

4.1 Quantitative Measures

Our study employed quantitative measures to assess participants' experiences with GestoChat, focusing on three research questions using 5-point Likert scales.

4.1.1 User engagement. The survey asked participants to rate their agreement with a series of statements aimed at gauging user engagement. They provided feedback on their motivation to continue the interaction, the degree to which GestoChat held their attention, and their interest in future use, offering a quantitative insight into the chatbot's ability to engage users.

4.1.2 Effectiveness of Communication. For the effectiveness of communication, participants evaluated how accurately GestoChat understood their queries, the clarity and ease of understanding the chatbot's responses, and whether GestoChat's gestures contributed to clearer communication. This provided a numerical measure of the chatbot's communicative performance from the users' perspective.

4.1.3 Perceived Naturalness of HRI. The perceived naturalness of HRI was also quantitatively assessed. Participants rated the naturalness of GestoChat's gestures, how similar interacting with the chatbot felt to interacting with a human, and the smoothness and coordination of the chatbot's verbal and non-verbal communication. These ratings gave us a quantifiable understanding of the human likeness of the interactions.

4.2 Qualitative Measures

Our study also captured qualitative feedback through open-ended questions, allowing participants to express in their own words the impact of GestoChat's gestures on their interaction experience. Participants were asked to discuss the enhancement of message comprehension through gestures, their preferences for gesture capabilities in future chatbots, and offered suggestions for improvements.

This hybrid approach to data collection aimed to provide a comprehensive view of GestoChat's performance and impact from the users' perspectives, facilitating a more detailed understanding of how gestural integration in ECA chatbots can enhance HRI.

5 EXPERIMENTS AND RESULTS

Here, we report both the quantitative and qualitative results of our user study.

5.1 Quantitative Results

Here, we report the quantitative measures taken regarding the three research questions, as shown in Figure 3.

5.1.1 User engagement. The analysis of user engagement with GestoChat revealed a positive inclination towards the chatbot's interactive capabilities. Participants generally felt motivated to continue interacting with GestoChat (M = 4.00, SD = 1.00), indicating a strong inclination toward engagement. However, the standard deviation suggests variability in individual responses. The chatbot's ability to maintain user attention was consistently recognized (M = 3.00, SD = 0.00), pointing to its potential to capture and sustain user interest throughout the interaction. The likelihood of participants' future use of GestoChat also received a high average score (M = 4.33, SD = 1.15), signaling a promising outlook for its adoption.

5.1.2 Effectiveness of Communication. The effectiveness in understanding user queries was highly rated (M = 4.00, SD = 0.00), reflecting an agreement on GestoChat's comprehension capabilities. The clarity of the chatbot's responses was viewed positively, albeit with some variation in perception (M = 3.33, SD = 0.58). The contribution of gestures to communication clarity divided opinion among participants (M = 3.00, SD = 1.00), indicating that while gestures are a beneficial feature, their implementation may need refinement.

5.1.3 Perceived Naturalness of HRI. Participants generally found the gestures of GestoChat to be natural (M = 3.67, SD = 0.58). However, when it came to the chatbot's similarity to human interaction, there was a slight deviation (M = 2.67, SD = 0.58), suggesting room for improvement in making interactions feel more human-like. The synchronization between the chatbot's verbal and non-verbal cues was perceived as relatively smooth (M = 3.67, SD = 1.15), although the standard deviation indicates differing experiences among users.

5.2 Qualitative Results

Here, we report the qualitative results from our post-study survey. The responses indicated a general appreciation for the GestoChat's gestures, with participants noting that the robot felt more lifelike and engaging due to its movements. One participant explicitly stated the gestures made the interaction more interesting: *"The robot does feel more like a human and it's interesting to see how it moves each time"* [P3], while another felt the gestures added an element of cuteness to the robot. However, it's worth noting that there wasn't a strong indication that the gestures significantly enhanced understanding of the GestoChat's messages, as mentioned by P1: *"I think it's ok, better than nothing. It makes the robot more vivid"*. This suggests that while the gestures are a positive feature, their functional impact may be more subtle.

Participants also expressed a preference for future ECA chatbots to have gesture capabilities, highlighting the perceived benefits of human-like qualities and finding the robot endearing. The consensus seems to suggest that while not essential, gestures are a welcomed enhancement that contributes to the user experience: *"Although this is not an essential feature, it is a good to have"* [P1].

Additionally, participants offered constructive feedback on the GestoChat's gestures, suggesting they currently appear too general and could be improved with more specific movements: *"The gestures right now seem to be way too general and maybe you can add some specific movements"* [P3]. This feedback implies a need for a more nuanced gestural vocabulary that can convey a wider range of emotions and responses. One respondent criticized the aesthetic aspects of the chatbot's environment - *"The background is ugly and can be improved"* [P2], while another expressed a desire for the robot to look more human - *"Can you make it look more like a human?"* [P1], which could involve improvements in the ECA chatbot's visual design and the realism of its movements.

6 DISCUSSION

Our user study surfaced several interesting findings regarding the interaction with GestoChat. Notably, the unanimous positive rating of GestoChat's ability to understand queries was an expected outcome given the advancement in natural language processing technologies. However, the divided opinions on the contribution of gestures to communication clarity present an intriguing observation. While some participants found the gestures to be a natural and enhancing feature, others did not see a significant impact on their understanding, which was unexpected. This could be attributed to the varying degrees of emphasis that individuals place on non-verbal cues in communication.

One notable area where the system did not perform as well as anticipated was in its perceived human likeness. Despite being rated favorably for the naturalness of gestures, some participants felt that the interactions did not closely mimic those with a human, which may suggest a disconnect between the gestures' design and users' expectations of human-like interactions. A possible explanation for this could lie in the limitations of current GGM, which may lack the subtlety and range of human non-verbal communication.

Further research should expand the sample size for enhanced statistical power results and compare different GGMs. Studies could also test conditions with and without GGMs to examine effects more comprehensively in within- or between-subject designs. Additionally, enhancing the chatbot's visual realism and environmental aesthetics could potentially improve user perceptions of human likeness.

7 CONCLUSION

In conclusion, the study confirmed the viability of integrating gesture generation with language models to enhance the user experience with ECA chatbots. The positive feedback on user engagement and GestoChat's understanding of queries points to the success of the underlying LLM technology. The mixed responses to GestoChat's gestural communication and human likeness highlight areas for future development. These findings lay the groundwork for future research that could refine the interplay between verbal and non-verbal cues in ECA chatbots, ultimately contributing to more natural and immersive conversational agents.

REFERENCES

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*. Springer, 373–383.

[2] Paul Bremner and Ute Leonards. 2015. Speech and gesture emphasis effects for robotic and human communicators: A direct comparison. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 255–262.

[3] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.

[4] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429.

[5] Susan Goldin-Meadow and Martha Wagner Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology* 64 (2013), 257–283.

[6] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. arXiv:2005.11129 [eess.AS]

[7] Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents: Research Articles. *Comput. Animat. Virtual Worlds* 15, 1 (mar 2004), 39–52.

[8] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 242–250. https://doi.org/10.1145/3382507.3418815

[9] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>College Station</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 11–21. https://doi.org/10.1145/3397481.3450692

[10] Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: automatic generation of nonverbal behavior for virtual humans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[11] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

[12] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research* 19, 5 (2017), e151.

[13] Silke Ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design features of embodied conversational agents in eHealth: a literature review. *International Journal of Human-Computer Studies* 138 (2020), 102409.

A APPENDIX

A.1 Dataset

According to the feedback received during our proposal review, our study is more aligned with a generation type of project thus involving the use of surveys instead of a dataset. For more information, please refer to section 4.

A.2 Quantitative Results

The quantitative measures taken regarding the three research questions are shown in Figure 3.

A.3 Contributions

- Jiadi Luo: Responsible for designing the study, conducting the user study, and writing the poster and the final report.
- Duc Dang: Responsible for coding with the gesture generation model, and integrating with unity.
- Jaeryang Baek: Responsible for integrating the LLM with the gesture generation model, and preliminary testing.

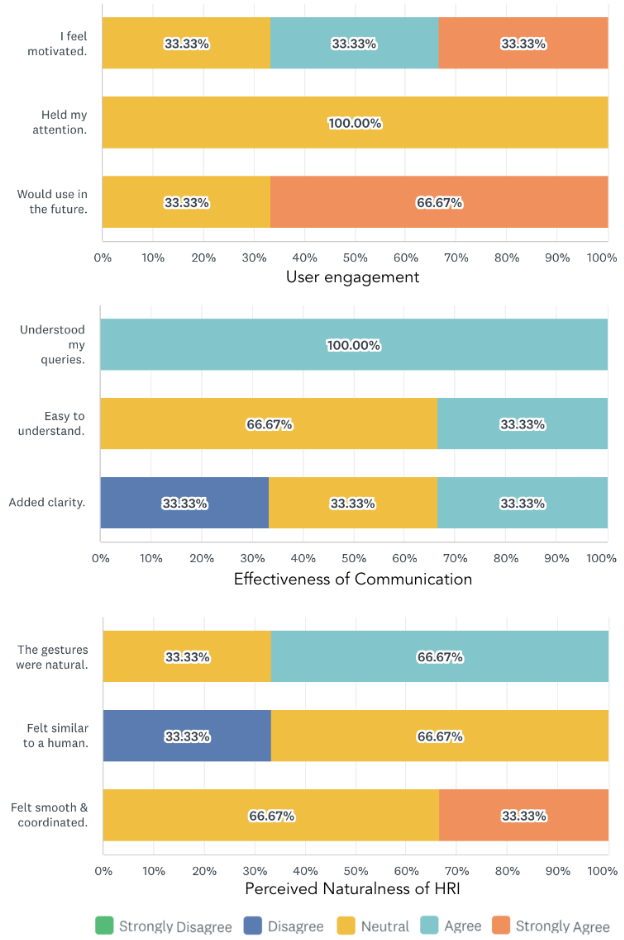


Figure 3: Quantitative results regarding the three research questions.

A.4 User Study: Consent Form & Survey

https://drive.google.com/file/d/1OOLeX7nEEmPRG8t3rFOHPsHAwEBxMhL-/view?usp=drive_link

A.5 User Study: Survey Responses

https://drive.google.com/file/d/14Wult3Hlmq7SdEL0DhR9d5jJA-qRRWK5/view?usp=drive_link