Human-Like Embodied AI Interviewer: Employing Android ERICA in Real International Conference

Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
{pang, fu, lala, elmers, inoue, kawahara}@sap.ist.i.kyoto-u-ac.jp

Abstract

This paper introduces the human-like embodied AI interviewer which integrates android robots equipped with advanced conversational capabilities, including attentive listening, conversational repairs, and user fluency adaptation. Moreover, it can analyze and present results post-interview. We conducted a real-world case study at SIGDIAL 2024 with 42 participants, of whom 69% reported positive experiences. This study demonstrated the system's effectiveness in conducting interviews just like a human and marked the first employment of such a system at an international conference. The demonstration video is available at `https://youtu.be/jCuw9g99KuE`.

1 Introduction

Qualitative interviews are foundational to social science research, offering deep insights through open-ended conversations. However, these interviews require considerable time and human effort. Earlier efforts to alleviate these demands involved using virtual agents (Nunamaker et al., 2011; Anderson et al., 2013; SB et al., 2021). Yet, these systems often failed to provide the sophisticated human-like interaction needed for quality research, limited to simple behaviors like head nodding and assuming participants' full understanding and fluent speech. This basic approach does not account for the complexities of real-world interactions, such as varied understanding and communication skills among participants, resulting in data quality and engagement shortfalls.

To address these limitations, this paper introduces a novel, **human-like interview system** that employs android and humanoid robots. This system is equipped with functionalities like advanced listening behaviors, conversational repair strategies, and user-fluency adaptation, which significantly enhance interaction quality. Beyond mere data gathering, our approach includes an end-to-end **post-interview processing workflow** where chained large language models (LLMs) handle data processing, analysis and presentation creation. We conducted a **real-world case study** at an international academic conference, where it facilitated numerous interactions, demonstrating its practical utility and efficiency. Notably, this marks the first instance of such a system being used at an international conference, showcasing our pioneering approach in the field. The comparative effectiveness of our system relative to traditional interview methodologies is detailed in Table 1.

2 Human-like Interview System

In this section, we describe the architecture of our human-like interview system, as depicted in Figure 1. The system initiates with a speech processing module that serves as the primary input mechanism. The core component, the dialogue manager, orchestrates tasks from language comprehension to response generation, including a Voice-Activity-Projection (VAP) based Multilingual Turn-Taking Module for effective turn management (Inoue et al., 2024a,b). Additional features of the system, such as speech synthesis and gesture generation, are outlined in subsequent subsections. Following the discussion of these components, the interview dialogue flow and the post-interview processing workflow are detailed. The system has been implemented across two distinct embodied conversational agents (ECAs): ERICA (Glas et al., 2016; Inoue et al., 2016; Kawahara, 2019), a human-like android robot, and TELECO (Horikawa et al., 2023), a less anthropomorphic, teleoperated humanoid robot.

2.1 Speech Processing

For automatic speech recognition (ASR) and the extraction of prosodic features, we utilize a hand

Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations, pages 136–150
January 19–24, 2025.

System	Agent	Agent Behavior	Dialogue Features	Post-Interview Processing Workflow
SPECIES (Nunamaker et al., 2011)	Virtual	Eye Blink, Head Nodding	Follow-up Question	No
Maya (SB et al., 2021)	Virtual	Gestures, Head Nodding	Follow-up Question	No
ERICA (Inoue et al., 2021)	Robotic	Eye Blink, Lip Sync, Head Nodding	Follow-up Question	No
ERICA & TELECO (Ours)	Robotic	Eye Blink, Lip Sync, Gestures, Head Nodding, **Verbal Backchannel**	Follow-up Question, **Conversational Repair**, **User Fluency Adaptation**	**Yes**

Table 1: Comparison of embodied AI interview systems. Bold highlights features unique to our proposed system.

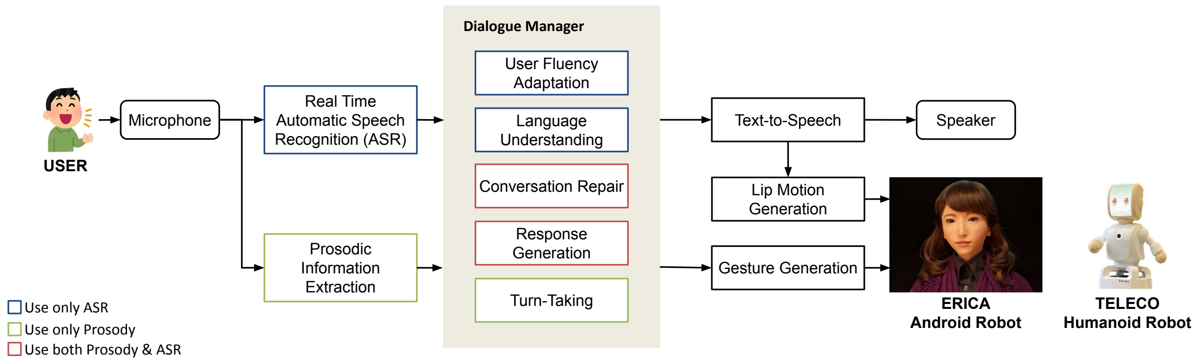

Figure 1: Overall architecture of human-like interview system

microphone. The ASR system is implemented via a real-time ASR module[1], which is based on the faster-whisper model[2]. This setup facilitates the extraction of critical prosodic information, including fundamental frequency (F0) and power, from the spoken input.

2.2 Dialogue Manager

The dialogue manager, a key component in our interview system, manages response selection based on user input. It comprises several sub-modules that improve interaction quality: a language understanding module that interprets user context to generate follow-up questions or smooth transitions; a backchannel module that predicts and delivers verbal and non-verbal cues, adding naturalness to the conversation; and a conversation repair module that detects and corrects communication breakdowns. Figure 2 depicts the architecture of the dialogue manager, highlighting the interplay among these components in response generation.

2.2.1 Language Understanding

The language understanding module uses ASR outputs for sentiment analysis, identifying keywords

[1]https://github.com/KoljaB/RealtimeSTT
[2]https://github.com/SYSTRAN/faster-whisper

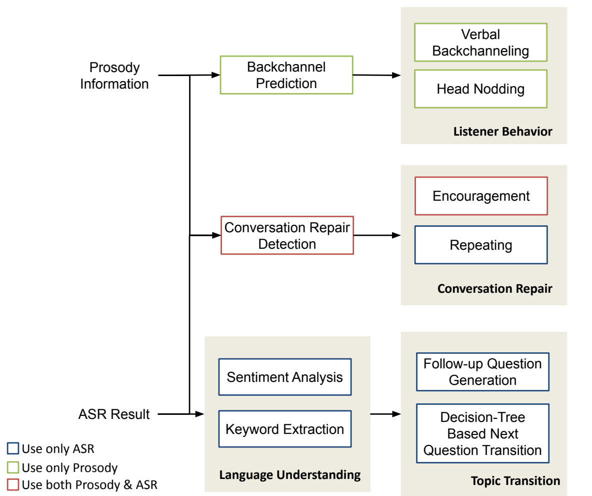

Figure 2: Overall architecture of interview system response generation

indicative of positive, neutral, and negative sentiments from a predefined polarity word list. For generating follow-up questions, the system considers context length and keyword presence—significant keywords include “because” and “as.” A follow-up question is generated if the context is under five words or lacks these keywords, using our predefined list of questions. Utilizing a decision-tree method, responses indicating agreement or disagreement guide the direction of subsequent questions. An example of dialogue processing and response generation by our system is detailed in the Appendix A.

2.2.2 Backchannel

Backchanneling, where listeners indicate attentiveness through verbal, non-verbal, or combined responses, is crucial in conversations. Previous research has documented its use across languages (Cutrone, 2005; Ike, 2010) and settings (Widiyati, 2016; Maynard, 1986), including interviews (Wulandari, 2017; Nurjaleka, 2019; Laforest, 1994). Effective backchanneling and active listening enhance the interviewer's appeal and improve response quality (Louw et al., 2011; Rogers and Farson, 1957; Nurjaleka, 2023). Despite advancements in LLMs, generating appropriate backchannels remains challenging, underscoring their importance in achieving human-like conversations. Previous human-robot interactions have primarily used non-verbal cues like head nodding without verbal responses (Inoue et al., 2021).

To address this, our system separates backchannel prediction and generation. We utilize a Multilingual-VAP based model (Inoue et al., 2024a), fine-tuned with attentive listening data, to predict appropriate moments for backchanneling based on prosodic cues. For generation, we developed a repertoire of verbal backchannels—such as "hmm," "erm," and "mhmm"—and diverse head-nodding patterns that vary in frequency and speed for use in conversations. This approach supports simultaneous verbal and non-verbal backchannels to enhance the realism and effectiveness of the robot.

2.2.3 Conversation Repair

Conversation breakdowns frequently disrupt dialogues, particularly in spoken interviews, and can stem from issues from either the user or the system. For users, misunderstandings or difficulty in expressing thoughts can cause interruptions, whereas, for the system, challenges such as unrecognized speech or delays in processing can impede conversational flow.

To mitigate these disruptions, our system incorporates a conversation repair module that employs strategies of repeating and encouraging based on keyword detection. Utilizing prosodic cues and ASR results, the module identifies phrases indicating confusion, like "pardon?" or "could you say that again?" to repeat questions for clarity. Similarly, if expressions such as "I have no idea" or "I don't know" are detected, the system offers supportive responses, encouraging users to continue sharing their thoughts.

In cases where speech recognition fails despite clear voice activity, the system may use simple backchannels like "mhmm" to encourage continuation. Furthermore, to address processing delays that might lead to pauses, the system deploys interim responses like "That's interesting!" or "That's a good point!" from a predefined list, maintaining the conversational momentum while preparing the next question.

2.2.4 User Fluency Adaptation

User fluency significantly affects the smoothness of conversational flow during interviews. Fluent users usually engage without issues, but those less proficient may need additional time to articulate their thoughts, often resulting in longer silences and potential misunderstandings if the conversational pace is too rapid. This is particularly the case with non-native speakers. Our system includes a user fluency adaptation module that adjusts speaking speeds and extends turn-taking intervals according to user proficiency.

This module utilizes a Words-Per-Minute (WPM) based strategy, specifically designed to accommodate users with a WPM of 75 or below—indicative of beginner levels A1 to A2 according to the Common European Framework of References for Languages (CEFR)[3]. For these users, the system slows down its speech and allows longer response times. This adaptation helps non-fluent speakers engage effectively with our system, as standard conversational speeds in English, typically between 150-190 WPM and reaching up to 197 WPM in formal interviews (Marslen-Wilson, 1973; Richards, 1983; Wang, 2021), far exceed what beginners can handle. Even academic presentations, which generally maintain a slower pace of 100–125 WPM for clarity (Wong, 2009), surpass optimal speeds for these users. Utilizing user fluency adaptation, our system ensures the interview process accommodates speakers of varying proficiency, which is crucial in international conferences with diverse linguistic backgrounds.

[3]https://magoosh.com/english-speaking/english-proficiency-levels-a-guide-to-determining-your-level/

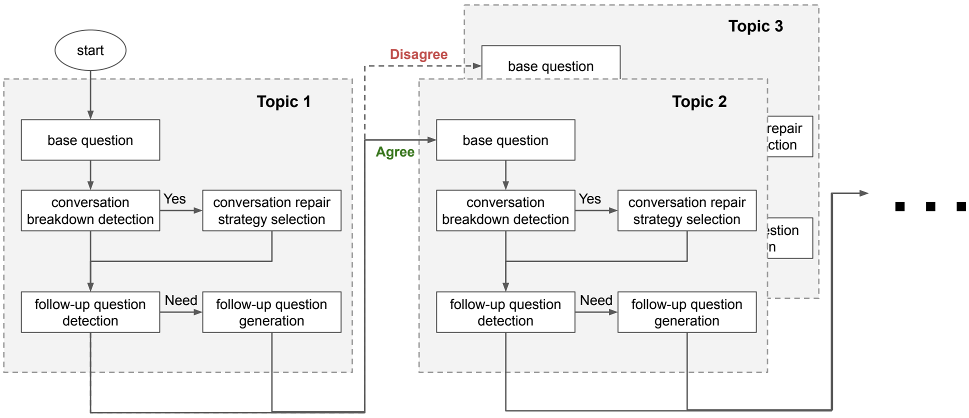

Figure 3: Overall architecture of interview dialogue flow

2.3 Speech Synthesis

For speech synthesis, our system uses the Julie voice provided by the VoiceText engine from Hoya Corporation[4]. Although this engine capably synthesizes standard speech, it struggles with the nuanced pronunciation of verbal backchannels such as "mhmm" or "hmm". These elements are crucial for natural conversational flow but are not adequately represented when generated directly by typical text-to-speech (TTS) systems due to their unique phonetic characteristics.

To overcome this limitation, we manually adjusted and refined the pronunciation of each backchannel, subsequently creating the corresponding .wav files. This approach allows our system to incorporate a diverse array of backchannels, varying in form and speed, to enhance the realism and dynamic nature of interactions.

2.4 Gesture Generation

To enhance the human-like quality of our interview system, we developed a range of gestures that extend beyond mere head nodding. Among these, an open palm gesture, which signifies openness and accessibility, fostering an environment conducive to free expression and interaction[5]. Additionally, we have implemented gestures such as leaning back to indicate surprise during interactions, and a bowing gesture to signify respect and formality after completing an interview. These gestures are strategically designed to mimic human non-verbal cues, thereby enhancing the naturalness and effectiveness of the robot's interactions with users.

2.5 Interview Dialogue Flow

The interview dialogue flow in our human-like interview system is managed via finite state transitions, as depicted in Figure 3. The process initiates with a base question. Based on the user's response, the system evaluates whether a conversational breakdown has occurred and if interventions, such as repeating or encouraging, are necessary. If such responses are required, the system generates them and maintains the current question state. Otherwise, the system assesses whether the user has provided sufficient information. If the information is inadequate, a follow-up question is posed. Subsequently, the system determines the next set of questions to be addressed based on the user's latest response. This cycle continues throughout the interview. Concurrently, to enhance human likeness and express interest in the user's responses, the system delivers both verbal and non-verbal backchanneling while receiving user input.

2.6 Interview Question Strategy

For the interview question strategy, we adopted a hybrid approach that integrates both template-based and generative question sets. On one hand, we employed LLMs (i.e., GPT-4o-mini API[6]) to dynamically produce follow-up inquiries, thereby adapting naturally to user responses and maintaining a human-like conversational flow. On the other hand, we use a fixed set of template-based questions for the primary prompts central to our data collection, ensuring that these core questions remain consistent across all interviews. This balance not only supports reliable downstream analysis but also

[4] https://readspeaker.jp/
[5] https://www.globallisteningcentre.org/body-language-of-listeners/
[6] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

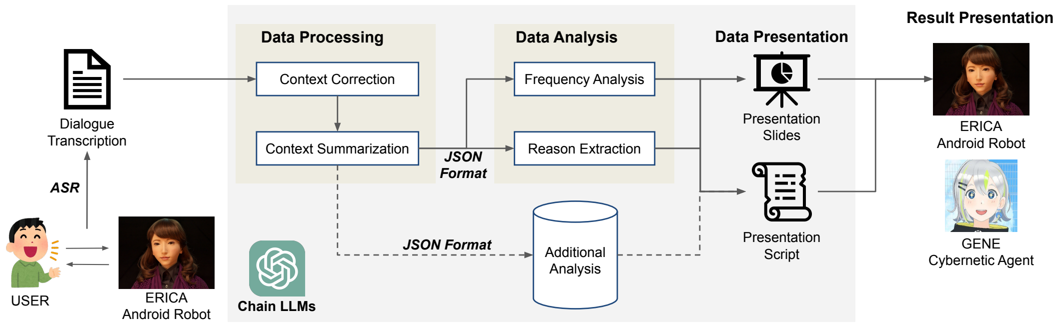

Figure 4: Overall architecture of post-interview processing workflow

enables adaptability through the generative component. Additionally, the system's modular design allows for flexible expansion for both the template-based and generative prompts, reducing the need for extensive manual rule-crafting. However, to ensure the highest level of analytic rigor in a research setting, the real-world case study presented in Section 3 relied solely on the template-based approach, maintaining question stability, which is necessary for accurate evaluation.

2.7 Post-Interview Processing Workflow

The post-interview processing workflow in our system facilitates data processing, analysis, and presentation. Utilizing a series of chained LLMs, specifically GPT-4o-mini[7], our system segments tasks into distinct subtasks with targeted prompts. This modular approach enhances task specificity and enables precise control over the process, allowing for modifications at any stage to suit specific research needs. The workflow's structure is detailed in Figure 4.

The pipeline consists of three main phases: data processing, analysis, and presentation. Initial data processing corrects ASR errors, ensuring data integrity, and prepares data in JSON format for subsequent analysis. The analysis involves evaluating opinion distributions and motivations, with flexibility for additional inquiries. Presentation materials, such as scripts and slides, are generated from the analysis results, using tools like the python-pptx library[8]. This automated system concludes with presentations delivered by conversational agents such as robots or virtual agents. Each subtask's detailed prompts are provided in the Appendix B.

3 Real-World Case Study

To evaluate our human-like embodied AI interviewer's effectiveness in a real-world setting, we conducted a case study at SIGDIAL 2024, attended by over 160 participants[9]. This study assessed perceptions of conversational AI's human-likeness, exploring themes such as essential interaction qualities, the importance of human-like traits, the inclusion of negative traits, and strategies against misuse.

Participants engaged in brief interviews lasting 2-3 minutes with one of two robots at the conference: ERICA, an android resembling a female adult, and TELECO, a humanoid robot with an OLED display face and simplified joint structures. Both robots exhibited identical dialogue behaviors, gestures, and facial expressions. Figure 5 illustrates a user interacting with ERICA during the conference[10]. Another user interacting with TELECO during the conference is illustrated in Figure 6. The example dialogues are provided in Appendix C, and the setup details and interview results are documented in Appendix D.

Interviews occurred over the first two days of the conference, with findings presented on the final day. To ensure a natural interaction environment, no formal questionnaire feedback was solicited. Instead, experiences were gathered directly during the interviews and through spontaneous post-interview discussions with participants. Insights from these interactions are elaborated in the subsequent subsection.

Consent was obtained by informing attendees at the conference's opening session and through clearly displayed notices in the interview room, advising that only transcripted dialogues from ASR

[7]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[8]https://python-pptx.readthedocs.io/en/
[9]https://2024.sigdial.org/
[10]Demo video is available at https://youtu.be/v1vfRJu_UJ4

Figure 5: Photo of interview dialogue with ERICA by SIGDIAL participant

Figure 6: Photo of interview dialogue with TELECO by SIGDIAL participant

would be recorded.

3.1 Reporting on Panel Discussion

In academia, panel discussions usually involve a group of experts and a moderator to foster an informative exchanges of viewpoints. Such settings are advantageous for gathering expert opinions across various fields, providing a deep understanding of specific topics (Rasmussen, 2008; Tempero et al., 2011; Filbeck et al., 2017). However, these discussions often face time constraints and typically limit participation to high-level experts like professors, restricting the diversity of perspectives.

Our system extends beyond this limitation by collecting opinions from conference participants at all levels, not just from high-level experts, ensuring that all participants had the opportunity to express their opinions. During the panel discussion session on the last day, our system presented the analyzed results. Due to logistical challenges, instead of presenting with ERICA on stage, we utilized a computer-generated (CG) agent, Gene (Lee, 2023), who presented the results[11]. The presentation by Gene is illustrated in Figure 7.

[11]Demo video is available at `https://youtu.be/pSgaouAUkZk`.

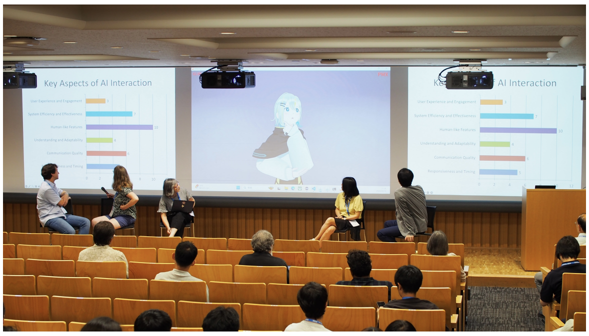

Figure 7: Photo of Gene's presentation during the panel discussion session at SIGDIAL

3.2 Result and Discussion

The feedback from participants was predominantly positive, affirming the system's effectiveness in facilitating engaging and memorable interactions. Of 42 participants, 29 described the interaction as "enjoyable and engaging" and felt it encouraged them to share more thoughts. These comments reflect the system's success in engaging users effectively, with the overall results illustrated in Table 2.

However, not all feedback was positive. From two participants, critical insights emerged, highlighting the repetitive nature of the interview, with remarks like "The interview felt repetitive as the robot asked fixed questions." This feedback underscores the need for more adaptive and personalized follow-up questions, potentially through enhanced use of LLMs to enable dynamic conversation flows. Additionally, some participants expressed discomfort with the robots' human-like appearance, indicating a need for careful calibration to balance human-likeness and user comfort.

Further feedback from spontaneous post-interview conversations revealed mixed reactions to the system's backchanneling capabilities. While many appreciated the verbal and non-verbal backchannels for enhancing the perception of attentiveness and human-likeness, there were criticisms about the naturalness of synthesized backchannels like "mhmm," suggesting that the current speech synthesis engine may not effectively capture the casual tone required for everyday conversational backchannels. This opens avenues for future research into developing more human-like and context-appropriate backchannel generation.

Discussions also revealed diverse preferences concerning the appearances of our robots, particularly between the highly human-like android ERICA and the less human-like humanoid TELECO.

Some participants found ERICA's resemblance unsettling, while others valued the genuine sense of co-presence she provided. In contrast, TELECO's less human-like features did not evoke the same level of co-presence. These varied responses highlight cultural or personal differences in acceptance and preference of robot aesthetics, suggesting a rich area for further investigation into how culture and personality influence human reactions to the human-likeness of robots.

Experience	Common Reasons
Positive **(69.05)**	1. Interaction is engaging 2. Interesting human-like robot
Neutral (26.19)	1. Interesting but experienced an error 2. Interesting but wanted more support
Negative (4.76)	1. Questions felt repetitive 2. Robot appearance caused discomfort

Table 2: Overall Interview Experience Result [%]

4 Conclusion

In this paper, we introduced the human-like embodied AI interviewer, integrating android and humanoid robots with chained LLMs to support researchers in data collection, analysis, and presentation. Our system improved interview quality by incorporating advanced conversational behaviors such as attentive listening, conversational repairs, and user fluency adaptation, and automated the analysis and presentation processes post-interview.

A two-day case study at an international academic conference validated our system's effectiveness, with 69% of participants reporting positive experiences. The system also streamlined data analysis and presentation. Notably, this was the first use of such a system at an international conference, demonstrating its applicability in real-world research settings.

Looking forward, we aim to enhance the human-like features of our system, focusing on improving backchannel generation and exploring cultural and personal preferences for robot appearances to optimize user interactions. We hope that these enhancements will bring us closer to achieving human-level interaction capabilities in android robots, further bridging the gap between technology and natural human communication.

5 Limitations

While our preliminary two-day case study at an international academic conference offered initial validation, the relatively small sample size (42 participants) limits the generalizability of our findings. To improve the generalizability of our findings, we plan to conduct larger-scale studies with more diverse participant groups.

Another limitation lies in the repetitive nature of the template-based questioning utilized in the real-world case study. While these fixed templates ensured stability for analysis, they reduced conversational variability. In future work, we aim to incorporate LLMs to generate questions more dynamically, exploring methods to maintain question stability without compromising adaptability and user engagement.

Lastly, the current system relies solely on speech input, which constrains its capacity to fully interpret users' states and behaviors. Moving forward, we will integrate richer multimodal inputs—including facial expressions, body language, and environmental context—to achieve more responsive, context-aware interactions and enhance the overall user experience.

Acknowledgments

The authors thank the participants of SIGDIAL 2024 for their contributions, which were crucial for the analysis and conclusions of this study. This research was supported by the JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJMS2011).

References

Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *International conference on advances in computer entertainment technology*, pages 476–491. Springer.

Pino Cutrone. 2005. A case study examining backchannels in conversations between japanese–british dyads. *Multilingua*, 24(3):237–274.

Greg Filbeck, Victor Ricciardi, Harold R Evensky, Steve Z Fan, Hunter M Holzhauer, and Andrew Spieler. 2017. Behavioral finance: A panel discussion. *Journal of Behavioral and Experimental Finance*, 15:52–58.

Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. Erica: The erato intelligent conversational android. In *2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)*, pages 22–29. IEEE.

Yukiko Horikawa, Takahiro Miyashita, Akira Utsumi, Shogo Nishimura, and Satoshi Koizumi. 2023. Cybernetic avatar platform for supporting social activities of all people. In *2023 IEEE/SICE International Symposium on System Integration (SII)*, pages 1–4. IEEE.

Saya Ike. 2010. Backchannel: a feature of japanese english. In *JALT 2009 Conference Proceedings*, volume 205, page 215. JALT Tokyo.

Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2021. A job interview dialogue system with autonomous android erica. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 291–297. Springer.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2403.06487*.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868*.

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with erica, an autonomous android. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 212–215.

Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*, pages 65–75. Springer.

Marty Laforest. 1994. Listening strategies in sociolinguistic interviews. convergence and divergence. *Culture*, 14(2):51–61.

Akinobu Lee. 2023. MMDAgent-EX.

Stephen Louw, R Watson Todd, and Pattamawan Jimarkon. 2011. Active listening in qualitative research interviews. In *Proceedings of the International Conference: Research in Applied Linguistics, April*, pages 71–82.

William Marslen-Wilson. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature*, 244(5417):522–523.

Senko K Maynard. 1986. On back-channel behavior in japanese and english casual conversation. *Linguistics*, 24(6):1079–1108.

Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, and Mark W Patton. 2011. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, 28(1):17–48.

Lisda Nurjaleka. 2019. Backchannel behavior in interview discourse: A contrastive study between japanese and indonesian. In *Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*, pages 451–457. Atlantis Press.

Lisda Nurjaleka. 2023. Backchannels responses as conversational strategies in the interaction of indonesian speakers in interview setting. *REiLA: Journal of Research and Innovation in Language*, 5(2):99–114.

Terje Rasmussen. 2008. Panel discussion ii: Culture and media technology. the internet and differentiation in the political public sphere. *nordicom Review*, 29(2):73–83.

Jack C Richards. 1983. Listening comprehension: Approach, design, procedure. *TESOL quarterly*, 17(2):219–240.

Carl Ransom Rogers and Richard Evans Farson. 1957. *Active listening*. Industrial Relations Center of the University of Chicago Chicago, IL.

Pooja Rao SB, Manish Agnihotri, and Dinesh Babu Jayagopi. 2021. Improving asynchronous interview interaction with follow-up question generation. *IJIMAI*, 6(5):79–89.

MA Tempero, J Berlin, M Ducreux, D Haller, P Harper, D Khayat, H-J Schmoll, A Sobrero, and E Van Cutsem. 2011. Pancreatic cancer treatment and research: an international expert panel discussion. *Annals of oncology*, 22(7):1500–1506.

Li Wang. 2021. British english-speaking speed 2020. *Acad. J. Humanit. Soc. Sci*, 4:93–100.

Elok Widiyati. 2016. The features of interrupting, collaborating, and backchannelling used by broadcaster and caller in telephone conversation. *EduLite: Journal of English Education, Literature and Culture*, 1(1):1–16.

Linda Wong. 2009. *Essential study skills*. Houghton Mifflin.

Nur Wulandari. 2017. A conversation analysis of verbal backchannel response in radio program valentine in the morning interview with michael bublé. *Sastra Inggris-Quill*, 6(6):585–596.

Appendix

A Language Understanding Dialogue Example

Below is a dialogue example demonstrating our language understanding system. **R** indicates an utterance by the system, while **U** indicates an utterance by the user. Dialogue 1 demonstrates a scenario where the system generates a follow-up question due to the absence of keywords like "because" or "as". In contrast, Dialogue 2 shows a scenario where the system does not generate a follow-up question because the user spoke extensively.

Dialogue 1
R1 What do you think is the most important thing when interacting with a conversational AI?
U1 Uh, well, I would say the response time maybe.
R2 Interesting! Can you tell me more about why you think that's so important? (Generated a follow-up question as no keywords were detected)

Dialogue 2
R1 What do you think is the most important thing when interacting with a conversational AI?
U1 I think it should be that I am properly understood and my questions are addressed accurately. I think that if I'm talking with somebody, I really would like to be understood, so I think that's very important.
R2 That's a good point! So, let's move on. (Moved to the next question since the user spoke a lot)

The example dialogues below showcase how responses guide the flow of the interview, especially when determining the next question. In the *importance of human-like traits* question, when keywords like "agree", "yes", "yeah" are detected indicating agreement, the next question shifts to *the inclusion of negative traits*, as shown in Dialogue 3. Conversely, if the user disagrees, the next questions will be about *strategies against misuse*, as shown in Dialogue 4.

Dialogue 3
R1 Do you think conversational AI should aim to be human-like?
U1 Yeah, not only human-like but also considering the user's preferences.
R2 Okay, that's interesting! And what about negative human traits? Should conversational AI include these as well? (Moved to the *inclusion of negative traits* topic since the user agreed)

Dialogue 4
R1 Do you think conversational AI should aim to be human-like?
U1 Um, not really. I think that conversational AI can be useful even if it's not human-like.
R2 Makes sense! So, what would you do to prevent misuse of conversational AI? (Moved to the *strategies against misuse* topic as the user disagreed)

Sentiment analysis was implemented in our proposed system to determine the appropriate response. For instance, when querying user experiences about being interviewed by a robot, the robot's response will change based on the sentiment of the user's utterance. This adaptation is illustrated below in Dialogue 5 and Dialogue 6.

Dialogue 5
R1 How did you feel about being interviewed by me, a human-like robot?
U1 Yeah, it was a really interesting experience because this is my first time.
R2 I'm glad that you enjoyed this conversation. I appreciate your time! Have a wonderful day! (Concludes positively in response to the user's enjoyment)

Dialogue 6
R1 How did you feel about being interviewed by me, a human-like robot?
U1 It's a little creepy.
R2 I'm sorry to hear that, but I appreciate your honesty. Thank you for your time. (Responds courteously to the user's negative feedback)

B Prompts

Below are the prompts provided to the LLMs to facilitate response generation. In the post-interview processing workflow, our system used a series of chained LLMs to handle data through context correction, summarization, data analysis, and the generation of presentation slides and scripts. The pipeline involves a cascading approach where the input from each subtask is passed to the next task. This approach comprehensively manages processing, analysis, and presentation generation. Detailed prompts for each subtask are depicted in Figures 8 to 12.

C Case Study Dialogue Example

See Figure 13 for a dialogue example[12] that explores participant perceptions of the *human-likeness of conversational AI*. This example addresses four primary topics: essential interaction qualities, the importance of human-like traits, the inclusion of negative traits, and strategies to prevent misuse. In the dialogue, **ROBOT** denotes system utterances, while **HUMAN** represents user responses. As detailed in Section 2.3, verbal backchannels cannot be directly synthesized by our speech engine; therefore, we manually created these sounds and played the corresponding .wav files as needed. In the dialogue example, any system utterance ending with .wav indicates a generated verbal backchannel. Due to the simultaneous occurrence of verbal backchannels and user utterances, the log file records system backchannels before user responses.

D Case Study Details

As discussed in Section 3, we conducted a real-world case study at the SIGDIAL international conference. During the initial two days, participants were interviewed by our embodied conversational agents, ERICA and TELECO, for data collection purposes. Figure 5 and 6 illustrate a user interacting with ERICA and TELECO, respectively, during one of these sessions. On the final day of the conference, our system analyzed and presented the results of these interactions during a panel discussion session. Figure 7 displays Gene, our CG agent, presenting these results. Figure 14 showcases the script, while Figure 15 displays the slides used during the presentation.

[12]Demo video is available at `https://youtu.be/v1vfRJu_UJ4`

```
Given a raw transcript of a dialogue between a human and a robot, please correct any
obvious errors in the human's responses that seem misrecognized by the automatic
speech recognition (ASR) system. When correcting the user responses, you should
correct them based on the question asked by the robot.

You should:
-Retain all natural elements of spoken dialogue such as fillers, hesitations, and
repetitions, as they reflect the natural speaking style.
-Only correct parts that are clear misinterpretations or irrelevant to the context
provided by the robot's questions.
-Ensure the corrections align logically with the questions asked by the robot.
-Correct the conversation and maintain the format and context.

For example:
"Yeah like I have to wait for quite a while yeah because um yeah my research
interest is still about that Yeah so I have to deal with the conversation and quite
a lot actually"

You should correct it to something like

"Yeah like I have interacted quite a lot actually yeah because um yeah my research
interest is still about that Yeah so I have to deal with the conversation and quite
a lot actually"

Remember, it must remain in the original context except those that seem irrelevant,
which the ASR system misrecognized.

The conversation is as follows:
```

Figure 8: Prompt for correcting dialogue context due to ASR error

```
Given the transcript of a conversation between a robot and a human during an interview
about conversational AI, summarize the human's opinions into a structured JSON object
using specified categories. Each category should be clearly structured with
subcategories as follows:

- 'Interact_with_AI_Before': Answer as 'yes' or 'no'.
- 'Important_Aspect': Provide 'aspect' which part of conversational AI is considered
important, and 'reason' explaining why.
- 'Should_Human_Like': Provide 'agreement' as agree/disagree, and 'reason' for the
view.
- 'Include_Negative_Traits': Provide 'agreement' as agree/disagree, and 'reason' for
the view.
- 'Precautions': Mention the 'aspect' of necessary precautions for conversational AI,
and the 'reason' for them.
- 'Interview_Experience': Describe the experience as 'opinion' being positive,
neutral, or negative, and provide a 'reason'.

Ensure accuracy and clarity in the responses, and maintain the context of the
conversation. The conversation is as follows:
```

Figure 9: Prompt for summarizing dialogue context into JSON format

```
You are a PowerPoint presentation specialist. You are asked to create the content for an academic
presentation for the academic conference SIGIDAL on the analysis result report regarding human-like
conversational AI interview research. This study employs two conversational robots to conduct interviews
during the conference, gathering attendees' perspectives on the realism and effectiveness of AI-driven
communication. The data collected from these interviews will be analyzed and presented by a virtual agent
named Gene. Your role is to generate the slides and script automatically, with the final script to be
manually inputted into Gene, who will execute the presentation.

The first slide should be the presentation title and the Presenter's name only. The subsequent slides will
present the analysis of the interview data.

Structure the information for a PowerPoint presentation aimed at a researcher audience. Each slide should
have a title, content summarized in bullet points, and, when applicable, chart data to visually represent
the analysis result, remember to include the "others" category if cannot fix all the results. The charts
should include percentages and category names.

Output Format:
Return the structured information as a JSON object, where each slide specifies the content in bullet
points and the type of chart with its corresponding data. Your answer should only contain the JSON - no
markdown formatting or explanatory text.

Example:
{
  "slides": [
    {"title": "Title Slide", "content": "Presentation Title: Analysis of Participant Opinions"},
    {
      "title": "Understanding Participant Demographics",
      "content": [
        "Summary of participant age groups",
        "Insights into demographic distribution"
      ],
      "chart": {
        "type": "bar_chart",
        "data": {
          "categories": ["Under 25", "25-40", "Over 40"],
          "values": [10, 15, 5],
          "labels": ["10%", "15%", "5%"]
        }
      }
    },
        {"title": "Conclusion Slide", "content": "Main conclusions and future research directions"}
  ]
}

The information is as follows:
```

Figure 10: Prompt for generating presentation slide context

```
You are a PowerPoint presentation specialist tasked with creating Python code to generate a professional
academic presentation for the SIGIDAL conference on human-like conversational AI research. You will use
the python-pptx package and a previously generated JSON script detailing slide contents, including text
and chart data.

Design Guidelines:
Fonts: Use Arial for titles (size 32, bold) and Tahoma for content (size 24). Justify all content text.
Color and Emphasis: Bold and use red font color for important keywords in the content.
Bullet Points: Format content with multiple pieces of information as bullet points. Determine whether
content should be bullet points or sub-bullets based on context.
Title Slide: Use the title slide layout, including the conference and the presenter's name as the
subtitle.

Python Code Instructions:
Generate slides based on the JSON input. If a slide specifies a chart, integrate the chart using the data
provided. Your response should contain only the Python code, no explanatory text.
Example JSON Input:

{
  "slides": [
    {"title": "Introduction", "content": "Overview of the project objectives and key results"},
    {
      "title": "Data Analysis",
      "content": "Graphical representation of data trends, statistical summaries",
      "chart": {
        "type": "bar_chart",
        "data": {
          "categories": ["Category 1", "Category 2", "Category 3"],
          "values": [20, 30, 50],
          "labels": ["20%", "30%", "50%"]
        }
      }
    },
    {"title": "Conclusion", "content": "Summary of findings, future research directions"}
  ]
}

The information is as follows:
```

Figure 11: Prompt for generating presentation slide python script from the presentation slide context

```
Generate an academic presentation script based on the provided data analysis summary of
interview experiences with conversational AI. The script should be spoken in English, last
approximately 5 minutes, and adhere to the following guidelines:

1. The presentation is to be delivered by a virtual agent named Gene at SIGDIAL 2024.
2. Begin with an opening that includes:
   - Introduction of the virtual presenter, Gene.
   - A greeting that mentions enjoyment of SIGDIAL 2024 so far.
   - A special thanks to ERICA and TELECO for the data collection.
3. Proceed to detail the data analysis summary which includes:
   - Number of participants and their previous experience with AI.
   - Important aspects of AI interaction highlighted by participants.
   - Opinions on AI exhibiting human-like traits and the inclusion of human-like flaws.
   - Precautions and ethical considerations suggested by participants.
   - General sentiment and specific feedback from participants about their interaction
experiences.
4. Conclude the presentation by:
   - Summarizing the insights on how human-like qualities in AI impact user experience.
   - Emphasizing the need for a balanced approach in AI development that incorporates
human-like interaction dynamics without compromising ethical standards.
   - A closing statement that does not invite questions, thanking the audience for their
attention.

Please use clear, professional, and engaging language suitable for an academic audience.
Ensure the script is structured for smooth delivery by the virtual agent.
Only use the periods, commas, and exclamation marks for punctuation. Please put the commas
at the places where you would naturally pause while speaking. Your script should use I'm
instead of I am, for a more conversational tone. Only output the script content, excluding
any system prompts or any sectional headers.

The script is as follows:
```

Figure 12: Prompt for generation presentation script

```
ROBOT: Hi! I'm ERICA, nice to meet you!
ROBOT: emm_speed150.wav
HUMAN: Hi Erica, nice to meet you!
ROBOT: I'm doing a survey on conversational AI. And I wonder if I can ask you a few questions?
ROBOT: mhm_question.wav
HUMAN: Yes, please go ahead.
ROBOT: Great, thank you! To start off, Have you interacted with a conversational AI before?
HUMAN: Yes, I interact with them quite a lot.
ROBOT: That's great! So what do you think is the most important thing when interacting with a
conversational AI?
ROBOT: erm_speed225.wav
ROBOT: erm_speed225.wav
HUMAN: That's a very good question. Personally, I think the most important aspect would be
making sure the conversational AI matches the scenario or situation.
ROBOT: That's a good point! So do you think conversational AI should aim to be human-like?
ROBOT: erm_speed225.wav
ROBOT: mhm_question.wav
ROBOT: m.wav
HUMAN: Yeah, to follow up on my last response, I think it depends on the situation. So for
example, if it's more of a situation where you would expect some kind of human-level service,
I think it makes a lot of sense to act human-like.
ROBOT: Okay, That's interesting! And what about negative human traits? Should conversational
AI include these as well?
ROBOT: emm_speed150.wav
ROBOT: emm_speed150.wav
HUMAN: Yeah, that's a very good point. I think in some situations, it's very good to have some
negative traits so that the conversational AI will appear even more human. But if you're
trying to come off as professional, maybe it'd be better to omit some of those.
ROBOT: Ok, interesting! So what would you do to prevent misuse with conversational AI?
ROBOT: emm_speed150.wav
ROBOT: m.wav
ROBOT: erm_speed150.wav
HUMAN: Yeah, there are a lot of different steps I think that can be taken, and it's an area
where I believe there'll need to be continual improvement. But I think one area that's very
easy to address right now is making sure that the AI is not hallucinating and saying things
that are not true.
ROBOT: Ok, That's a good point! Before we wrap up, how did you feel about being interviewed
by, me, a human-like robot?
HUMAN: This was actually very enjoyable. Thank you very much for the experience.
ROBOT: I'm glad that you enjoyed this conversation. I appreciate your time! Have a wonderful
day!
ROBOT: erm_speed225.wav
HUMAN: I appreciate your time as well. Have a great day, Erica!
ROBOT: see you next time!
```

Figure 13: Dialogue example collected during the case studies

Good afternoon, everyone. My name is Gene, and I'm a virtual agent glad to be presenting to you today at SIGDIAL 2024. I hope you are enjoying this conference as much as I am.

Today, I'm presenting a detailed analysis of interview experiences with conversational AI, which sheds light on user interactions and perspectives. We conducted interviews with a total of 42 participants, with 34 having prior experiences with AI and 8 having none. This diverse group allowed us to capture a broad range of insights.

Several important aspects of AI interaction were highlighted by participants. The most frequently mentioned factors including response time, clarity, accuracy, and human-like communication. Participants emphasized that smooth and coherent conversations are essential for enhancing user experience. They expect conversational AI to understand their inputs without requiring adjustments or corrections on their part. The accuracy and efficiency of responses were deemed crucial; slow response times were particularly noted as a barrier to creating a natural flow in conversation.

When discussing whether conversational AI should exhibit human-like traits, 28 of the participants agreed it should, while 10 disagreed. Many expressed that human-like qualities can improve the overall interaction. However, some experienced users preferred a more functional and efficient AI that doesn't necessarily mimic human behavior. The varying opinions underscore the complexity of user preferences and expectations.

A fascinating aspect of our findings was the mixed sentiments regarding the inclusion of negative traits in AI. Just 13 participants agreed that AI should reflect negative human traits, indicating a strong preference for positivity in user interactions. Many felt that such traits could diminish trust and satisfaction, suggesting that while some might find realism in flaws, a focus on positive characteristics is more beneficial in fostering a good user experience.

Further, participants raised pertinent precautions and ethical considerations for the development of conversational AI. Key takeaways included the necessity for data testing, preventing misuse, and establishing trust with users. Many emphasized that clear guidelines should be enacted to ensure ethical interactions and safety, while also educating users about the capabilities and limitations of conversational AI.

Overall, the sentiment towards interview experiences with conversational AI was quite positive, with 29 participants rating their experience favorably. The engaging nature of the interaction was frequently noted, although there were instances of confusion and discomfort, especially related to communication dynamics."

In conclusion, our analysis underscores that human-like qualities in AI play a significant role in shaping user experiences. However, there exists a delicate balance that needs to be maintained. While aiming for human-like interactions can enhance engagement, it is equally important to uphold ethical standards and avoid negative traits that can undermine trust.

Thank you for your kind attention! and let's continue to explore the fascinating possibilities of conversational AI together!

Figure 14: Presentation script created by our system for Gene's presentation at the SIGDIAL conference

Figure 15: Presentation slides created by our system for Gene's presentation at the SIGDIAL conference.