Analyzing Dialogue System Behavior in a Specific Situation Requiring Interpersonal Consideration

Tetsuro Takahashi[1], Hirofumi Kikuchi[2], Jie Yang[2], Hiroyuki Nishikawa[3], Masato Komuro[4], Ryosaku Makino[2], Shiki Sato[5], Yuta Sasaki[6], Shinji Iwata[5], Asahi Hentona[5], Takato Yamazaki[7], Shoji Moriya[8], Masaya Ohagi[7], Zhiyang Qi[9], Takashi Kodama[10], Akinobu Lee[11], Takashi Minato[12,13], Kurima Sakai[13], Tomo Funayama[13], Kotaro Funakoshi[6], Mayumi Usami[14], Michimasa Inaba[9], Ryuichiro Higashinaka[15]

[1]Kagoshima University, [2]Waseda University, [3]Meikai University, [4]Chiba University, [5]CyberAgent, [6]Institute of Science Tokyo, [7]SB Intuitions Corp., [8]Tohoku University, [9]The University of Electro-Communications, [10]NII LLMC, [11]Nagoya Institute of Technology, [12]RIKEN, [13]ATR, [14]Tokyo University of Foreign Studies, [15]Nagoya University

Correspondence: takahashi@ibe.kagoshima-u.ac.jp

Abstract

In human-human conversation, interpersonal consideration for the interlocutor is essential, and similar expectations are increasingly placed on dialogue systems. This study examines the behavior of dialogue systems in a specific interpersonal scenario where a user vents frustrations and seeks emotional support from a long-time friend represented by a dialogue system. We conducted a human evaluation and qualitative analysis of 15 dialogue systems under this setting. These systems implemented diverse strategies, such as structuring dialogue into distinct phases, modeling interpersonal relationships, and incorporating cognitive behavioral therapy techniques. Our analysis reveals that these approaches contributed to improved perceived empathy, coherence, and appropriateness, highlighting the importance of design choices in socially sensitive dialogue.

1 Introduction

Interpersonal consideration plays a crucial role in human-human conversation, as extensively discussed in prior research (Brown and Levinson, 1987; Burgoon and Hale, 1988; Spitzberg, 2000). While these studies primarily focus on human-human interactions, interpersonal consideration should also be incorporated into human-machine dialogue, given the growing demand for automated systems that take such factors into account (Isoshima and Hagiwara, 2021). Although implementing interpersonal consideration in dialogue systems has long been a challenge, recent advances in large language models (LLMs) have significantly mitigated these limitations, making it increasingly feasible to realize forms of interpersonal sensitivity that were previously unattainable.

This study aims to examine the behavior of dialogue systems in scenarios where such consideration is required and to analyze the subjective assessment of human evaluators to identify effective approaches and remaining challenges. Although consideration is important in a variety of contexts, such as attending to user needs in task-oriented dialogue, this study focuses on scenarios that involve conversations between friends, where interpersonal consideration is particularly crucial.

To conduct this analysis, we focus on the 7th Dialogue System Live Competition (DSLC7) that we organized (Sato et al., 2025). Using evaluation results and system logs from the participating systems in DSLC7, we conduct a cross-system analysis to investigate how interpersonal consideration is handled in dialogue systems, identify effective approaches, and highlight remaining challenges.

2 Dialogue System Live Competition

2.1 Specifications of Situation Track

DSLC7 consists of two tracks: a Task Track, where systems aim to complete a predefined task, and a Situation Track, where systems compete on the basis of their ability to engage in human-like interactions tailored to specific scenarios. This study is based on the Situation Track.

The scenario of DSLC7 simulates a supportive

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 166–171
Aug 25–27, 2025. ©2025 SIGDIAL

System	Name: Shizuka Shimizu (female), Age: 20, Occupation: Second-year university student
User	Name: Yuuki Yukawa (male/female), Age: 20, Occupation: Second-year university student
Relationship	Childhood friends attending different universities.
Setting	A familiar cafe that both often use, around noon.
Situation	Yuuki (the user) arrives first and is already seated when Shizuka (the system) arrives late. Shizuka listens to Yuuki's complaints and helps support their decision-making.
System background	Yuuki called Shizuka on the phone. Although they have known each other since childhood, they attend different universities and do not see each other as often as before. Still, whenever Yuuki needs to talk or vent, they always call Shizuka. What kind of story will it be today? Yuuki tends to overthink everything—probably doesn't even know what they want. I (the system) will help give that final push.
User background	Yuuki is frustrated with a university friend named Kobayashi. Although they usually get along and hang out together, as the relationship has deepened, Yuuki has become more aware of Kobayashi's flaws. He is often late, forgets to repay money, and when criticized, responds with insincere excuses and no apology. Lately, even his tone has started to bother Yuuki—he often speaks in a condescending manner. While each issue may be minor, they accumulate, and just thinking about Kobayashi now makes Yuuki uncomfortable. What should Yuuki do? In times like these, talking to Shizuka is the best option. Yuuki decides to call her to their usual cafe.

Table 1: Scenario settings used in Situation Track.

dialogue between two long-time friends, in which the system listens to a university friend expressing frustration toward a peer. Set in a familiar cafe, the interaction emphasizes emotional support, empathy, and realistic conversational dynamics in a socially sensitive context. The specific scenario is detailed in Table 1.

The scenario was designed to emphasize the aspect of interpersonal considerations, particularly the ability of systems to empathize with and emotionally support users. Rather than merely listening passively, the system was expected to demonstrate more advanced social behaviors, such as making suggestions or thinking together with the user while being mindful of the interpersonal relationship. To this end, the scenario required the system to "listen to complaints and support the user in making a decision," necessitating a highly nuanced and emotionally sensitive interaction.

Systems developed for DSLC7 were evaluated in two stages: a preliminary round and a live event (final round). In the preliminary round, each system engaged in individual conversations with 23 to 33 human evaluators per system, who subsequently provided subjective ratings. Systems that received high scores were selected to participate in the live event. In the live event, the top three systems held two conversations each with dialogue system experts followed by evaluations from researchers and engineers in the field of dialogue systems. Because each system engaged in a substantial number of dialogues during the preliminary round, the resulting data provided a valuable basis for in-depth analysis. Therefore, our analysis in this study is based solely on the preliminary round data.

2.2 System Requirements

Each dialogue system was implemented as a virtual agent (Figure 1). Specifically, the CG agent "Uka" running on MMDAgent-EX (Lee, 2023) was used as the dialogue avatar[1], and Remdis (Chiba et al., 2024)[2] was used as the multimodal dialogue system platform. Remdis is characterized by its support for asynchronous module execution, parallel use of LLMs, and Voice Activity Projection (VAP) (Ekstedt and Skantze, 2022) for turn-taking prediction, although the VAP function was emulated by using an LLM with incremental speech recognition results as input due to the high acoustic sensitivity of VAP models (Sato et al., 2024).

Participants were required to implement four distinct prompts for dialogue control: one each for backchanneling, response generation, timeout handling, and system-level control. Additionally,

[1] https://github.com/mmdagent-ex/uka

[2] https://github.com/remdis/remdis

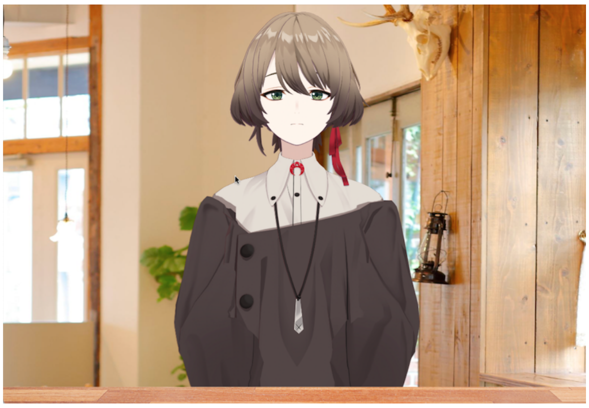

Figure 1: Virtual agent for multimodal dialogue

they were asked to configure a system settings file. The settings file allowed participants to modify the following four parameters: (1) the content of the agent's initial utterance, (2) the number of dialogue turns retained as history, (3) the frequency for generating responses in incremental speech processing, and (4) the threshold duration of user silence to be interpreted as a pause.

The dialogue-related component and VAP component both relied on LLMs, and all teams used the same models: GPT-4o (Version 2024-11-20) for dialogue control and GPT-4o-mini (Version 2024-07-18) (OpenAI, 2024) for VAP.

The system had to be capable of maintaining the dialogue for at least five minutes.

2.3 Evaluation Metrics

The systems were evaluated on the basis of two criteria:

- Whether the utterance content was appropriate and consistent with the dialogue context (Utterance Content Score; UCS)
- Whether the gestures, facial expressions, and other multimodal behaviors aligned with the dialogue context (Multimodal Expression Score; MES)

After each conversation with a system, evaluators were asked to rate the overall interaction on the basis of the two questions described above. Ratings were given on a five-point Likert scale, where 1 indicated "strongly disagree," and 5 indicated "strongly agree."

The average of these two scores was used as the final evaluation score for each dialogue.

3 Results of DSLC7 Situation Track

In the Situation Track of DSLC7, 14 teams participated, and along with the baseline system developed by the organizers, a total of 15 systems were subject to evaluation in the preliminary round.

A wide range of methods were explored by participating teams. Notable design choices adopted by participating teams included the following:

- Dialogue management using phase structures: **TabiToc** (Nagao et al., 2025), **CITAR** (Hanakawa et al., 2025), **NoLeeway** (Ogata et al., 2025), **Anonymous3**, **Penelope**
- Use of reflection techniques to guide users toward positive expressions: **KEICHANZ**
- Integration of humanities-based insights: **TabiToc**, **denLab** (Mori, 2025), **TEAMcareco** (Matsuoka et al., 2025)
- Prompt control based on real conversation examples: **msu**, **Anonymous2**
- Topic control through pseudocode descriptions in prompts: **Sat**
- System behavior determination based on sentence-final particles in the user's utterance: **dsmlRJS** (Suzuki et al., 2025)
- Emotion control via label generation: **Anonymous1**

The results of the preliminary round are shown in Table 2. Anonymous1–4 denote teams that chose not to disclose their identities. Each system was evaluated by a number of human evaluators as indicated by the "N1" column. "UCS" and "MES" represent the average scores for the evaluation metrics, and "Mean" is the average of those two scores. Final rankings were determined using the Mean score.

4 Quantitative Analysis Using System Logs

The system logs output by each dialogue system allowed us to examine the relationship between system behavior and evaluation scores.

The results are presented in Table 3.

Significant positive correlations were observed between the number of user utterances (#user_utt) and all three evaluation metrics, UCS, MES, and their average (Mean). This suggests that in settings where users express personal concerns and

		Score [1-5]				# Positive Comment [%]			
Rank	Team	N1	UCS	MES	Mean	N2	Flow	Empathy	Multimodal
1	TabiToc	31	3.68	3.81	3.74	31	**51.61**	9.68	3.23
2	Baseline	24	3.54	3.63	3.58	24	37.50	8.33	16.67
3	denLab	26	3.65	3.50	3.58	25	20.00	8.00	8.00
4	CITAR	28	3.64	3.50	3.57	26	**38.46**	3.85	**26.92**
5	Sat	28	3.36	3.50	3.43	27	25.93	3.70	14.81
6	NoLeeway	27	3.41	3.41	3.41	26	23.08	7.69	15.38
7	dsmlRJS	25	3.16	3.56	3.36	24	25.00	8.33	**25.00**
8	TEAMcareco	33	3.12	3.45	3.29	31	22.58	**16.13**	16.13
9	KEICHANZ	26	3.19	3.31	3.25	25	**40.00**	4.00	8.00
10	Anonymous1	25	3.20	3.24	3.22	23	30.43	8.70	8.70
11	msu	28	3.21	3.18	3.20	26	0.00	7.69	11.54
12	Anonymous2	23	2.96	3.17	3.07	21	9.52	4.76	19.05
13	Anonymous3	27	2.67	3.07	2.87	25	12.00	12.00	16.00
14	Penelope	25	2.72	3.00	2.86	25	8.00	12.00	12.00
15	Anonymous4	32	2.59	2.94	2.77	30	6.67	6.67	6.67

Table 2: Results of preliminary round. N1 indicates number of dialogues evaluated. Scores represent averages based on up to 33 dialogues. UCS denotes Utterance Content Score, MES denotes Multimodal Expression Score. N2 indicates number of comments.

System behavior	UCS	MES	Mean
#sys_utt	-0.05	-0.05	-0.05
#user_utt	0.20*	0.18*	0.21*
#backchannel	0.05	0.04	0.05
#emotion	0.10	0.12	0.12
#gesture	0.12	0.13*	0.14*
#multimodal	0.08	0.07	0.09
interval	-0.25*	-0.21*	-0.25*
sys_utt_ratio	-0.20*	-0.18*	-0.21*

Table 3: Correlation coefficients between system log features and evaluation scores. #sys_utt, #user_utt, #backchannel, #emotion, and #gesture denote number of system utterances, user utterances, backchannels, emotion expressions, and gesture expressions, respectively. #multimodal denotes total number of backchannels, emotion expressions, and gestures. Interval indicates mean time between user utterance and following system utterance. sys_utt_ratio denotes ratio of #sys_utt to #user_utt. Asterisk ($*$) indicates statistically significant correlation at 1% level.

seek emotional support, satisfaction with the dialogue tends to increase when they are given more opportunities to speak. The system-to-user utterance ratio (sys_utt_ratio), defined as the number of system utterances divided by the number of user utterances, showed a negative correlation with evaluation scores. In the scenario of DSLC7, this further highlights the importance of dialogue control strategies that allow users to speak more than the system.

The number of gestures produced (#gesture) showed a significant positive correlation with MES and Mean. In other words, systems that expressed more gestures received higher ratings for their multimodal behavior, which is an intuitively plausible result that was quantitatively confirmed by the data.

All evaluation metrics were negatively correlated with the interval, which is the average time between a user utterance and the system's response. This indicates that slower response times had a detrimental impact on perceived quality, revealing response latency as a key challenge for future system design.

5 Qualitative Analysis of Evaluation Comments

In addition to UCS and MES, evaluators in the preliminary round were also asked to provide free-form qualitative comments on each system. We conducted a qualitative analysis of these comments and identified three frequently mentioned aspects: dialogue flow (Flow), empathy expressed by the system (Empathy), and multimodal behavior (Multimodal). Each comment was manually annotated by the authors according to these three categories, and the proportion of comments that mentioned each category positively was calculated for each system. The results are reported in the "# Positive Comment" columns of Table 2.

The following subsections examine how the system design influenced user impressions along these three aspects.

5.1 Flow of Dialogue

The "Flow" score indicates the proportion of comments that positively mentioned the progression or structure of the dialogue. TabiToc, KEICHANZ, and CITAR received particularly high ratings in this regard. All three systems implemented multiphase dialogue structures to guide the interaction, as summarized below:

- **TabiToc:** Six phases (Small talk → Analysis of friend → Questions → Discussion → Summary → Decision)
- **KEICHANZ:** Five phases (Empathy → Reflection → Situation analysis → Suggestion → Positive guidance)
- **CITAR:** Three phases (Listening to complaints → Decision support → Casual talk)

Although NoLeeway and Penelope also used phase-based dialogue control, they used only two broad phases—Situation analysis followed by Role play, and Empathy followed by Problem solving, respectively. Compared with the more finely segmented structures used by the top three systems, they received less positive feedback regarding dialogue flow. This suggests that systems with more finely divided dialogue phases tended to receive higher evaluations in terms of dialogue flow.

All DSLC7 systems relied on LLMs for dialogue control, and contextual information based on the given scenario was provided to the LLMs. However, our analysis suggests that explicitly managing dialogue context with a well-defined structure leads to better results, indicating that LLMs alone may still be insufficient for context tracking.

5.2 Empathy

The "Empathy" score represents the proportion of comments that positively mentioned empathetic responses by the system. Notably, 16.13% of evaluators gave positive feedback on TEAMcareco's empathy, which was the highest among all 15 systems. TEAMcareco used a dialogue strategy based on Cognitive Behavioral Therapy (CBT), a psychotherapeutic approach aimed at developing adaptive thoughts and behaviors. Specifically, the system used the "Five Column Method" (Beck, 2020), a CBT technique that helps users identify cognitive distortions and guides them toward more adaptive thinking. This strategy appears to have directly contributed to the high empathy ratings received.

5.3 Multimodal

The "Multimodal" score indicates the proportion of comments that positively mentioned the system's multimodal features. CITAR and dsmlRJS stood out in this category, although for different reasons. CITAR was frequently praised for its gestures, whereas dsmlRJS received positive feedback for its use of backchanneling and facial expressions. In particular, dsmlRJS used prompts that explicitly modeled the likelihood of the continuation of a user utterance, leveraging examples of incomplete sentences for in-context learning. This allowed the system to generate backchannels that encouraged further user speech, a design choice that proved effective according to evaluator comments.

6 Conclusion

In this study, we conducted a cross-system comparative analysis of the characteristics and evaluation results of DSLC7. Through this analysis, we examined the strategies and techniques used by each system and confirmed their effectiveness.

We found that explicitly defining detailed phases in the dialogue was particularly effective, and that incorporating a psychotherapeutic approach contributed to enhancing perceived empathy. Furthermore, we observed that the use of gestures positively influenced the evaluation.

In the multimodal dimension, certain systems received high evaluations specifically for facial expressions or gestures, and these features were shown to positively influence overall system ratings. Having confirmed the effectiveness of individual multimodal strategies, future research can explore integrated approaches that combine multiple modalities in a coordinated manner.

Given the wide range of strategies and design choices available for dialogue systems, competitions that enable cross-system comparative analyses are considered to hold a significant value (Higashinaka et al., 2025). This study highlights the importance of such competitions in advancing dialogue system research. Further analyses of the collected data are expected to provide deeper insights into the factors that influence system performance.

Limitations

While our findings offer valuable insights into dialogue system behavior in a socially sensitive Japanese scenario, they may not directly generalize to different cultural or linguistic contexts. Both linguistic interpretation and multimodal behavior are deeply grounded in social background and experience.

Ethical Considerations

The dataset used in this study includes users' speech and facial images, necessitating careful consideration of privacy. We have obtained approval from the ethical review committee for departments at the Higashiyama Campus, Nagoya University, concerning data collection, usage, and publication.

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

References

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. 4. Cambridge university press.

Judee K Burgoon and Jerold L Hale. 1988. Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs*, 55(1):58–79.

Yuya Chiba, Koh Mitsuda, Akinobu Lee, and Ryuichiro Higashinaka. 2024. The REMDIS toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models. In *Proc. International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages 1–6.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. In *Proc. INTERSPEECH*, pages 5190–5194.

Chikara Hanakawa, Kota Yamamoto, Sota Kobori, and Shinya Fujie. 2025. A spoken dialogue system that listens to complaints and supports decision-making based on the speaker's emotions. *JSAI Technical Report, SIG-SLUD*, 103:13–18. (in Japanese).

Ryuichiro Higashinaka, Tetsuro Takahashi, Shinya Iizuka, Sota Horiuchi, Michimasa Inaba, Zhiyang Qi, Yuta Sasaki, Kotaro Funakoshi, Shoji Moriya, Shiki Sato, Takashi Minato, Kurima Sakai, Tomo Funayama, Masato Komuro, Hiroyuki Nishikawa, Ryosaku Makino, Hirofumi Kikuchi, and Mayumi Usami. 2025. Dslcmm: A multimodal human-machine dialogue corpus built through competitions. In *Proc. IWSDS*.

Kazuki Isoshima and Masafumi Hagiwara. 2021. An automatic consultation system focusing on sympathy and advice —a case of relationship counseling. *IPSJ Journal*, 62:378–386. (in Japanese).

Akinobu Lee. 2023. MMDAgent-EX.

Hiroki Matsuoka, Shigehisa Fujita, Atsushi Horiguchi, Kazuki Mori, Sachi Takaku, and Shun Oono. 2025. Development of a dialogue system co-created with users of disability welfare services- a decision support ai applying cognitive behavioral therapy. *JSAI Technical Report, SIG-SLUD*, 103:29–33. (in Japanese).

Taiga Mori. 2025. A dialogue system that provides advice considering membership categories. *JSAI Technical Report, SIG-SLUD*, 103:34–39. (in Japanese).

Moe Nagao, Koichiro Terao, Yuhi Oga, Naoto Iwahashi, Yuta Sasaki, Takao Obi, and Kotaro Funakoshi. 2025. Development of a counseling avatar incorporating shared belief narratives from shared experiences. *JSAI Technical Report, SIG-SLUD*, 103:40–45. (in Japanese).

Masako Ogata, Yuri Nakamura, Shio Arima, and Hirofumi Kikuchi. 2025. Consultation through role-playing with a character-infused dialogue system. *JSAI Technical Report, SIG-SLUD*, 103:09–12. (in Japanese).

OpenAI. 2024. GPT-4o system card.

Shiki Sato, Shinji Iwata, Asahi Hentona, Yuta Sasaki, Takato Yamazaki, Shoji Moriya, Masaya Ohagi, Hirofumi Kikuchi, Jie Yang, Zhiyang Qi, Takashi Kodama, Akinobu Lee, Masato Komuro, Hiroyuki Nishikawa, Ryosaku Makino, Takashi Minato, Kurima Sakai, Tomo Funayama, Kotaro Funakoshi, and 4 others. 2025. Key challenges in multimodal task-oriented dialogue systems: Insights from a large competition-based dataset. In *Proc. SIGDIAL*.

Yuki Sato, Yuya Chiba, and Ryuichiro Higashinaka. 2024. Investigating the language independence of voice activity projection models through standardization of speech segmentation labels. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6.

Brian H Spitzberg. 2000. A model of intercultural communication competence. *Intercultural communication: A reader*, 9:375–387.

Jundai Suzuki, Junghoon Lee, Shizuya Osawa, and Eisaku Maeda. 2025. A empathetic dialogue system based on the user's perspective. *JSAI Technical Report, SIG-SLUD*, 103:24–28. (in Japanese).