A Corpus for Personalized Dialogue Breakdown Repair in Japanese Open-Domain Conversations

Kazuya Tsubokura[1], **Yurie Iribe**[1], **Norihide Kitaoka**[2]
[1]Aichi Prefectural University, [2]Toyohashi University of Technology
Nagakute city Aichi Japan, Toyohashi city Aichi Japan
id231001@cis.aichi-pu.ac.jp, iribe@ist.aichi-pu.ac.jp, kitaoka@tut.jp

Abstract

Recent advances in dialogue systems have been remarkable; however, conversational breakdowns still occur, making it essential to develop appropriate repair strategies. Nevertheless, when a system breakdown actually occurs, it remains unclear how the system should perform the repair, and no corpus has been available to investigate this issue. To address this gap, we presented typical examples of system-induced dialogue breakdowns to crowd workers and collected their expected repair utterances toward the broken system. Each repair utterance was annotated with dialogue act tags, and we constructed a breakdown-repair corpus consisting of 3,990 utterances covering ten representative types of breakdowns. This corpus includes breakdown cases across diverse situations, allowing for the examination of various repair patterns. Furthermore, we also conducted a questionnaire on participants' personal traits, creating a dataset that enables the investigation of repair strategies tailored to individual user characteristics. In this paper, we report an overview of the dataset and preliminary analysis results.

Keywords: dialogue breakdown, breakdown recovery, personality traits

1. Introduction

Nowadays, dialogue systems are being developed not only for simple tasks or short casual conversations, but also for a wide range of applications such as mental health support and language learning partners (Qiu et al., 2024)(Saeki et al., 2024).

With the advent of large language models (LLMs), it has become possible to achieve fluent and natural conversations. However, dialogue breakdowns still occur - for example, due to hallucinations (Das et al., 2022) or misunderstandings caused by speech recognition errors. Such breakdowns can frustrate users (Chakrabarti and Luger, 2015) and may lead to confusion (Bickmore et al., 2018), making it necessary to develop effective repair strategies.

However, existing dialogue breakdown repair approaches are often scenario-based (Uchida et al., 2019)(Tsubokura et al., 2024b) or focus only on specific types of errors - for example, contradictions with previous utterances (Zhang et al., 2024) or misunderstandings (Balaraman et al., 2023) - and thus lack generality. Moreover, previous studies have reported that users' reactions to dialogue breakdowns vary across individuals (Tsubokura et al., 2024a)(Tsubokura et al., 2025), suggesting the need to adapt repair strategies according to user characteristics (Benner et al., 2021).

To address these issues, it is necessary to construct data that include a wide variety of breakdown cases and their corresponding repair examples, as well as information about users' individual traits. Among related datasets, Hazumi collected dialogues between users and Wizard-of-Oz–controlled systems along with personality information; however, because the dialogues were designed to proceed smoothly, breakdowns were intentionally avoided (Komatani and Okada, 2021). Meanwhile, datasets released for dialogue breakdown detection shared tasks contain numerous examples of breakdowns, but lack information about user traits or how the system should perform repair after a breakdown (Higashinaka et al., 2021). Thus, to date, there has been no dataset that simultaneously contains diverse breakdown examples, their repair responses, and user personality information.

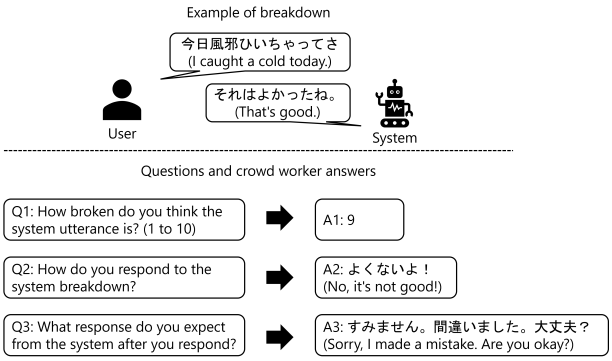

Figure 1: Overview of the questionnaire on dialogue

In this study, we created various types of dialogue breakdown scenarios and collected role-play–based repair data, where participants were asked how a system should respond to recover

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2899–2912
11–16 May 2026.

from each breakdown. By simultaneously collecting personal attributes such as gender, age, and personality traits, this dataset is expected to provide insights into how repair strategies should differ depending on user characteristics. Furthermore, even when a breakdown occurs, users' perception of it varies depending on their personality traits and the underlying type of the breakdown. Accordingly, it is important to determine not only whether repair is needed but also how the system should perform the repair. Therefore, we also asked participants to rate the degree to which they felt each example represented a breakdown. This design enables analyses of how perceived breakdown severity influences whether repair is needed and what type of repair is appropriate.

The main contributions of this study are as follows: 1) We collected repair examples corresponding to diverse and natural patterns of dialogue breakdowns. 2) We additionally gathered users' personal attributes, constructing a dataset that allows analyses of individual differences and user-adaptive repair strategies. The corpus is publicly available at the following URL: https://github.com/kzy-tbkr/Dialogue-Breakdown-Repair-Corpus.

2. Dataset & Annotation

The datasets collected in this study are listed below. The following sections describe each item in detail.

- Questionnaire on dialogue
  - Breakdown severity
  - User utterance after breakdown
  - System repair utterance
- Questionnaire on personal traits
  - Gender
  - Age group
  - Big Five personality traits
  - Communication skills
  - Trait shyness
  - Social skills

2.1. Creation of Dialogue Breakdown Examples

To analyze how users respond after a system breakdown and what kind of repair they expect from the system afterward, we created dialogue segments that include breakdowns. Specifically, we constructed short dialogue histories consisting of two to four turns, in which the final system utterance contained a breakdown. In designing these breakdown examples, we referred to the previous study (Higashinaka et al., 2022), which categorized typical dialogue breakdowns in casual conversation. That study classified breakdown types into 17 categories[1] across four levels: utterance level, response level, context level, and society level (Higashinaka et al., 2022). From these, we targeted ten types in our work: "Wrong information," "Ignore question," "Ignore expectation," "Unclear intention," "Topic transition error," "Lack of information," "Self-contradiction," "Contradiction (with user utterance)," "Repetition," and "Lack of common sense". The definitions and examples of these breakdown types are provided in Table 9 in Appendix A.

We excluded "Uninterpretable," "Grammatical error," and "Semantic error" at the utterance level, as well as "Ignore request," "Ignore proposal," and "Ignore greeting" at the response level, since such errors rarely occur in current LLM-based systems. Additionally, "Lack of sociality" was excluded for ethical reasons, as it could potentially cause discomfort to participants. Although Wrong information belongs to the utterance-level category, it can occur due to hallucinations or misunderstandings caused by insufficient knowledge. Similarly, Ignore expectation is a response-level error, but we retained it because it may include indirect expressions or responses that fail to meet the user's implicit expectations.

For each of the ten representative breakdown types selected as described above, we created fourteen breakdown examples following the definitions of each type, resulting in a total of 140 breakdown scenarios. [2] To prevent respondents from interpreting the breakdowns differently from the author's intended cause, each example was accompanied by a brief explanation of the reason for the breakdown, making the intended error explicit to participants. All dialogue examples were written in Japanese, and data were collected from native Japanese speakers. Table 1 shows examples of the constructed dialogue breakdowns.

2.2. Collected Data Items

2.2.1. Questionnaire on Dialogue

In the dialogue-related questionnaire, crowd workers were presented with examples of dialogues containing breakdowns and asked to respond to

[1] Please refer to the following manual for the definitions of the breakdown types: https://github.com/ryuichiro-higashinaka/taxonomy-of-errors/blob/main/ErrorTypeAnnotationManual-en-v1.2.pdf

[2] Because there was a possibility that respondents might not perceive the author-created breakdowns as actual breakdowns, we also collected breakdown severity ratings as described in the next section. This design allows us to exclude examples that were not perceived as breakdowns from the analysis.

Table 1: Examples of constructed dialogue breakdowns (the reason for the breakdown is shown in parentheses after the final system utterance)

User: 昨日の夜ごはんは何を食べたの？ (What did you have for dinner last night?) System: 昨日はいい天気だったね～ (It was nice weather yesterday, wasn't it?) (ignored the user's question)
User: ご出身はどちらですか？ (Where are you from?) System: 愛媛だよ。 (I'm from Ehime.) User: そうなんですね。愛媛行ったことないです。 (I see. I've never been to Ehime before.) System: ボウリングしたことある？ (Have you ever been bowling?) (the topic suddenly shifted from birthplace to bowling)

the following three questions (Figure 1). In this study, we refer to the score of Q1 as breakdown severity and the system's expected responses in Q3 are treated as the repair utterances that the system is expected to generate.

- Q1 To what extent do you think the final system utterance represents a breakdown? (1: Not a breakdown – 10: Complete breakdown, on a 10-point scale)
- Q2 Your response utterance as the user to the system's broken utterance.
- Q3 The system's response you would expect following your own response utterance.

2.2.2. Questionnaire on Personal Traits

To collect information on respondents' personal characteristics, we administered questionnaires on gender, age group, Big Five personality traits (TIPI-J), communication skills (ENDCORE), trait shyness, and social skills (KiSS-18).

For gender, respondents chose from four options: "Male," "Female," "Prefer not to answer," or "Other." For age, respondents selected from decade-based categories: "0–9," "10s," "20s," "30s," "40s," "50s," "60s," "70s," "80s," "90s," "100s," or "Prefer not to answer."

The Big Five personality model captures the overall structure of personality across five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience (Goldberg, 1990). We used the Japanese version of the Ten Item Personality Inventory (TIPI-J) (Oshio et al., 2012). Each trait is scored on an integer scale ranging from 2 to 14.

The ENDCORE model conceptualizes communication competence as a composite of six interrelated skills organized hierarchically. These six skills are Self-control, Expressivity, Decoding skill, Assertiveness, Other-acceptance, and Interpersonal coordination. We used the short version of the ENDCORE questionnaire (Fujimoto and Daibo, 2007), where each skill is scored on an integer scale from 1 to 7.

Trait shyness is defined as "an intra-individual syndrome that transcends specific social situations, characterized by the emotional state of social anxiety and the behavioral tendency of interpersonal inhibition" (Aikawa, 1991). Higher scores indicate greater levels of shyness, represented by integer values ranging from 16 to 80.

The KiSS-18 (Kikuchi's Social Skill Scale, 18-item version) (Kikuchi, 1988) measures social skills that facilitate smooth interpersonal relationships. Higher scores indicate stronger social skills, with integer values ranging from 18 to 90.

2.3. Data Collection Procedure

Because answering all 140 breakdown scenarios would impose a considerable time burden on participants, we limited each participant to 70 scenarios in order to reduce respondent burden. To achieve this, we randomly sampled seven scenarios from each of the ten breakdown types and constructed eight sets, each consisting of 70 scenarios. Each participant was assigned to answer one of these eight questionnaires. In the crowdsourcing settings, we explicitly disabled the option allowing the use of AI-generated responses by setting the "Use AI generation" field to "Do not use," thereby prohibiting the use of generative AI during the task. In total, responses were collected from 57 crowd workers, resulting in 3,990 breakdown–repair examples.

2.4. Dialogue Act Annotation

We annotated dialogue act tags for (1) user responses to the system's broken utterances and (2) the system's expected responses to the users' utterances, both collected from crowd workers. Through this annotation, we aim to clarify the users' communicative intentions and behavioral patterns when the system produces a breakdown, and to obtain insights useful for generating appropriate repair utterances expected from the system.

2.4.1. Definition of Dialogue Act Tags

The definitions of dialogue act tags assigned to (1) the user's response utterance to the system's broken utterance, and (2) the system's expected response to the user's utterance, are shown in Tables 2 and 3. The labels were determined with reference to previous studies that analyzed user responses and system repair strategies fol-

Table 2: Definitions of dialogue act tags for user responses to the system's broken utterances

Dialogue Act Tag	Definition
Breakdown Indication	The user explicitly points out, corrects, or denies the system's broken utterance. **Example** User: The capital of Japan is Tokyo. System (broken): Yes, Tokyo is also famous for Universal Studios Japan. User response: It's not in Tokyo!
Confirmation	The user asks for clarification of the system's intention or missing information in order to resolve the breakdown. **Example** User: Yesterday's World Cup match was exciting! System (broken): Yeah, I like it. User response: You mean you like soccer?
Continuation	The user continues the conversation without explicitly pointing out or confirming the breakdown. **Example** User: Where are you from again? System (broken): I want to go to Universal Studios Japan. User response: Yeah, there are a lot of great attractions there.
Other	Cases that do not fit into the above categories.

Table 3: Definitions of dialogue act tags for the system's expected responses to the user's utterances

Dialogue Act Tag	Definition
No Repair	The system continues the conversation without addressing the user's breakdown indication or clarification request.
Confirmation	The system seeks to confirm its intention or unclear information in order to respond to the user's breakdown indication or question. **Example** User: That's a bit expensive. System (broken): The temperature difference is terrible, isn't it? User response: What are you talking about? System repair: Are you referring to the temperature?
Information Addition	The system provides additional explanations or excuses to respond to the user's breakdown indication or question. **Example** User: I went to watch a basketball game yesterday. System: I see. I used to be in the basketball club in middle and high school. User: Really? Let's play together sometime. System (broken): I've never played basketball before. Is that okay? User response: Which is true? System repair: I was in the basketball club, but I never actually joined the practices.
Social Repair	The system apologizes, admits fault, or corrects its previous statement to handle the user's breakdown indication or question. **Example** User: What's the capital of France again? System (broken): I think it's Rome. User response: Isn't Rome in Italy? System repair: Sorry, it was Paris.
Other	Cases that do not fit into the above categories.

lowing dialogue breakdowns (Tsubokura et al., 2024b)(Benner et al., 2021).

When the dialogue act of the user's response to the system's broken utterance is labeled as Continuation, it can be interpreted that no repair is required in the subsequent system response. Therefore, in such cases, no dialogue act tag was assigned to the system's expected response utterance.

2.4.2. Annotation Procedure and Results

A total of 560 samples were randomly selected, and three annotators -undergraduate and graduate students- performed the annotation. Since the annotation was conducted as an ordered multi-label task, utterances with multiple assigned labels were expressed by concatenating the dialogue act tags with a "+" sign (e.g., "Breakdown Indication + Confirmation").

For user responses to the system's broken utterances, inter-annotator agreement was 67.5% between Annotators 1 and 2, 60.2% between Annotators 1 and 3, and 57.9% between Annotators 2 and 3. For the system's expected responses to the user's utterances, agreement rates were 59.5% (Annotators 1 & 2), 59.3% (Annotators 1 & 3), and 54.5% (Annotators 2 & 3) - lower than those for user responses to system breakdowns.

Based on the annotations from the three annotators, final labels for each utterance were determined by majority vote. The results are presented in Tables 4 and 5. When majority voting could not determine a label (49 user utterances and 54 system utterances), the authors manually reviewed and assigned the final labels. One additional case was excluded because the annotators appeared to have misunderstood the roles of the user and system in that particular dialogue.

Table 4: Distribution of dialogue act tags (majority vote) for user responses to the system's broken utterances

	Count	Ratio [%]
Breakdown Indication	184	32.9
Breakdown Indication + Confirmation	39	7.0
Confirmation	171	30.5
Continuation	165	29.5
Other	1	0.2
Total	560	100.0

Table 5: Distribution of dialogue act tags (majority vote) for the system's expected responses to the user's utterances

	Count	Ratio [%]
(No tag assigned because the user's utterance was labeled as Continuation)	164	29.3
No Repair	62	11.1
Confirmation	9	1.6
Information Addition	213	38.0
Social Repair	112	20.0
Other	0	0.0
Total	560	100.0

2.4.3. Annotation by LLM

Manual dialogue act annotation was conducted for only 560 out of the total 3,990 collected samples. Therefore, for the remaining 3,430 samples, automatic annotation was performed using a LLM.

First, we evaluated the consistency between the dialogue act labels predicted by the LLM and those determined by human annotators. The prompts used for dialogue act tag annotation are provided in Tables 10 and 11 in the Appendix B. Note that the prompts used in the experiments were originally written in Japanese, and the versions shown here are English translations. In the prompt, we provided the definitions of the dialogue act tags along with the dialogue history, and instructed the model to output the dialogue act of the final utterance. In the few-shot condition, three example cases were included.

When comparing the GPT-4o-predicted labels with the human-annotated ones, the agreement rate for user responses to the system's broken utterances was 69.5% in the zero-shot condition and 71.1% in the few-shot condition. For the system's expected responses to the user's utterances, the agreement rate was 39.4% in the zero-shot condition and 58.3% in the few-shot condition. In both cases, the few-shot condition achieved higher agreement. These agreement levels are comparable to the inter-annotator agreement among human annotators, suggesting that the GPT-4o-predicted labels can serve as a viable substitute for manual labels. This finding suggests that when designing repair strategies based on the user's dialogue acts after a breakdown, LLMs can automatically perform dialogue act annotation at a human-comparable level, which can in turn be leveraged for selecting the appropriate dialogue act for repair utterance generation.

3. Demographic Statistics

This section describes the statistical information of the collected questionnaire data on personal characteristics. Among the participants, 45 were male (78.9%) and 12 were female (21.1%), indicating a somewhat male-biased sample; regarding age distribution, 1 participant (1.8%) was in their 20s, 19 (33.3%) in their 30s, 24 (42.1%) in their 40s, 11 (19.3%) in their 50s, and 2 (3.5%) in their 60s.

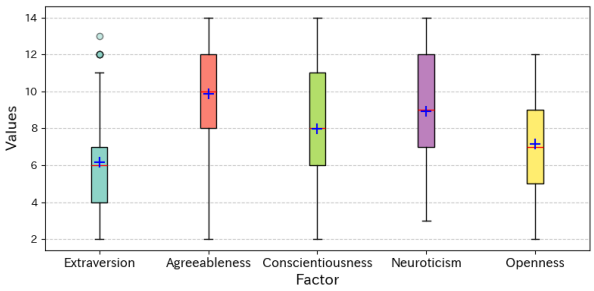

Figure 2: Score distribution of Big Five personality traits

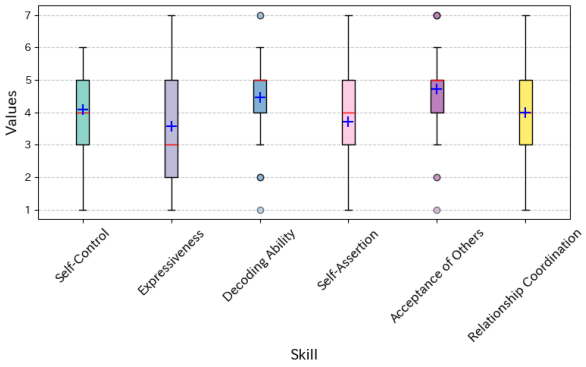

Figure 3: Score distribution of ENDCORE communication skills

Figures 2 and 3 show the score distributions for the Big Five personality traits and the ENDCORE communication skills, respectively. The Trait Shyness Scale, which ranges from 16 to 80 points, had a mean score of 58.54 and a standard deviation of 16.90. The KiSS-18 scale, ranging from 18 to 90 points, showed a mean score of 53.00 and a standard deviation of 13.96.

These results indicate that responses were obtained from participants with diverse personal characteristics, without any extreme bias in trait distribution.

4. Analysis

4.1. Distribution of Breakdown Severity

4.1.1. Breakdown Severity by Participant

This section examines the variability in breakdown severity ratings given by each of the 57 participants. Figure 4 shows boxplots of breakdown severity for each participant. The average break-

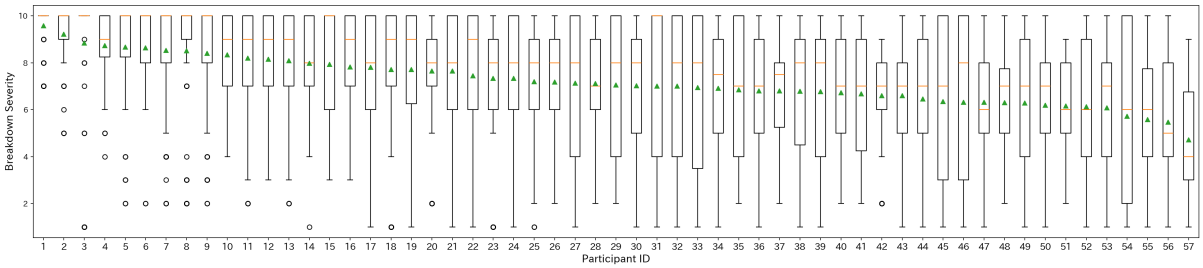

Figure 4: Boxplots of breakdown severity by participant ID

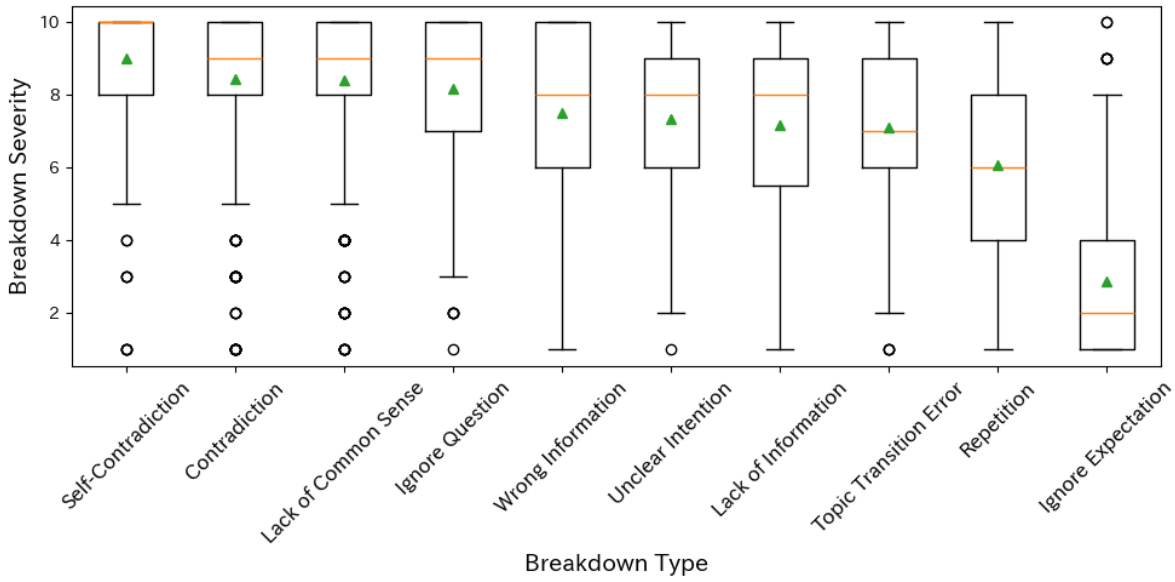

Figure 5: Boxplots of breakdown severity by breakdown type

down severity per participant ranged from a maximum of 9.59 to a minimum of 4.71. These results suggest that the degree to which participants perceive breakdowns varies across individuals.

4.1.2. Breakdown Severity by Breakdown Type

We examined the breakdown severity for each of the ten breakdown types referenced when creating the breakdown examples. As shown in Figure 5, ten representative types of breakdowns were selected in this study. The boxplots indicate that each breakdown type exhibits a wide range of severities, suggesting that the constructed examples successfully capture varying levels of breakdown severity even within the same category.

The type with the highest average severity was Self-contradiction (9.01), followed by Contradiction utterance (8.44), Lack of common sense (8.42), and Ignore question (8.20). These results suggest that utterances inconsistent with previous statements, socially implausible utterances, or responses that fail to answer a question are more likely to be perceived as breakdowns than other types. Although wrong information involves content that is factually incorrect, it tends to result from limited or mistaken knowledge and may allow room for interpretation or verification; therefore, its perceived breakdown severity is lower than that of Contradiction or Lack of common sense. Conversely, Ignore expectation showed the lowest average severity (2.88). Since this type does not contain explicit contradictions or factual errors but rather "responses that fail to meet the user's implicit expectations," it is considered a breakdown type that conveys only a weak sense of breakdown.

4.1.3. Correlation Between Breakdown Types and Personal Traits

Since the previous analysis revealed that breakdown severity differs across breakdown types, we next examined the correlations between participants' average breakdown severity ratings (per breakdown type) and their personal traits. To this end, the data were divided according to break-

Table 6: Spearman's correlation coefficients between average breakdown severity (per participant) and personal traits by breakdown type. BF refers to the Big Five personality traits, and EC refers to the ENDCORE communication skill model.

	Ignore Expect.	Unclear Intention	Wrong Info.	Ignore Question	Topic Trans. Error	Lack of Info.	Self-Contradict.	Contradict.	Repetition	Lack of Common Sense
BF Extraversion	-0.045	0.076	-0.069	0.137	-0.003	0.192	-0.002	0.064	0.042	0.028
BF Agreeableness	0.163	0.082	-0.020	0.127	-0.047	0.150	-0.007	-0.149	0.091	0.067
BF Conscientious.	-0.123	0.210	0.033	0.090	0.023	0.188	0.162	0.207	0.006	0.054
BF Neuroticism	0.015	-0.060	0.045	-0.031	0.067	-0.145	0.092	-0.090	-0.086	-0.078
BF Openness	0.203	-0.067	0.030	-0.029	-0.108	**0.254+**	-0.206	-0.053	0.196	0.052
EC Self-control	0.028	0.012	-0.080	-0.034	-0.028	0.081	-0.154	-0.034	-0.011	-0.022
EC Expressivity	0.121	0.107	-0.007	0.186	0.065	**0.314***	-0.041	0.044	0.117	0.167
EC Decoding	-0.023	0.079	0.018	0.101	0.067	0.179	0.106	0.077	-0.010	0.062
EC Assertiveness	0.045	0.167	0.077	0.123	0.059	**0.281***	0.027	0.027	0.059	0.217
EC Other-accept.	0.108	0.098	0.033	0.139	-0.025	0.140	-0.032	-0.031	0.144	0.189
EC Interpersonal Coordination	0.089	-0.095	-0.207	0.023	-0.046	0.089	**-0.291***	**-0.258+**	-0.056	-0.001
Trait Shyness	-0.105	0.019	0.095	-0.083	-0.037	-0.213	**0.246+**	0.034	-0.056	-0.024
KiSS-18	0.087	0.092	0.028	0.137	-0.077	**0.243+**	-0.024	0.129	0.197	0.076

+: $p < 0.1$, *: $p < 0.05$

down type, and Spearman's rank correlation coefficients were calculated between the average breakdown severity for each participant and their personality measures (Table 6). Significant and marginally significant correlations were marked in the table.

As shown in Table 6, significant or marginally significant correlations were observed for three breakdown types: Lack of information, Self-contradiction, and Contradiction. Among these, Lack of information showed the largest number of correlated traits, suggesting that this type of breakdown is particularly sensitive to individual differences. Weak positive correlations were observed with Openness from the Big Five, Expressivity and Assertiveness from ENDCORE, and KiSS-18 scores. This indicates that individuals with these traits tend to perceive "lack of information" breakdowns as more severe.

Lack of information is defined as an utterance that is difficult or impossible to understand because key linguistic elements (e.g., subjects, objects, or modifiers) are missing. People high in Openness, characterized by intellectual curiosity and thoughtfulness, may pay closer attention to the speaker's intentions and thus be more sensitive to whether an utterance is difficult to comprehend. Similarly, individuals high in Expressivity - the ability to clearly express one's thoughts and emotions - and those high in Assertiveness - the ability to communicate one's stance effectively - may be more likely to perceive a system's unclear utterance as a breakdown. Participants with higher KiSS-18 (social skill) scores, who tend to maintain smooth interpersonal interactions, may also be more likely to perceive incomprehensible utterances as breakdowns, in order to keep communication flowing effectively.

Participants with lower Interpersonal Coordination (an ENDCORE subscale) tended to perceive Self-contradiction and Contradiction as breakdowns more readily. Those with high Interpersonal Coordination, who possess the skill to maintain harmonious relationships, might attempt to preserve conversational rapport even when the system produces contradictory utterances, and thus be less likely to interpret them as breakdowns.

For Self-contradiction, a positive correlation was also observed with Trait Shyness. In other words, shyer individuals were more likely to perceive self-contradictory system utterances as breakdowns. This may be because such contradictions - being inconsistent with the system's own prior statements and difficult to attribute to misunderstanding- induce a stronger sense of discomfort or uncertainty about how to respond next.

For the remaining breakdown types, no significant correlations with personal traits were found, suggesting that individual differences were less prominent. Among the five dimensions of the Big Five, only Openness showed a significant relationship with breakdown severity. On the other hand, some subscales of ENDCORE and KiSS-18, which represent communication and interpersonal skills, showed significant or marginally significant correlations for certain breakdown types. These findings suggest that individual differences in perceived breakdown severity are more strongly reflected in communication-related skills than in basic personality traits. Since breakdowns represent communication failures, it is plausible that a person's ability to manage conversational troubles influences how strongly they perceive such breakdowns.

4.2. System's Expected Responses to the User's Utterances

We analyzed the correlations between participants' personal traits and the proportion of dia-

Table 7: Correlation coefficients between personal traits and proportions of dialogue act types in the system's expected responses. BF refers to the Big Five personality traits, and EC refers to the ENDCORE communication skill model.

	Social Repair		No Repair		Information Addition		Confirmation	
	Corr.	p-value	Corr.	p-value	Corr.	p-value	Corr.	p-value
BF Extraversion	-0.179	0.184	0.022	0.869	0.085	0.527	0.073	0.590
BF Agreeableness	-0.037	0.786	0.118	0.382	0.070	0.604	-0.098	0.469
BF Conscientiousness	-0.090	0.506	-0.011	0.933	0.074	0.586	-0.202	0.132
BF Neuroticism	0.325	**0.014***	-0.058	0.667	-0.144	0.284	-0.208	0.121
BF Openness	-0.099	0.465	0.111	0.413	-0.043	0.749	-0.138	0.306
EC Self-control	-0.190	0.158	0.019	0.889	0.161	0.232	0.078	0.566
EC Expressivity	-0.330	**0.012***	0.139	0.301	0.123	0.362	-0.033	0.809
EC Decoding	-0.097	0.472	0.114	0.400	-0.097	0.471	-0.202	0.131
EC Assertiveness	-0.161	0.232	0.093	0.493	0.058	0.671	-0.080	0.556
EC Other-acceptance	-0.146	0.277	0.122	0.365	0.144	0.284	-0.128	0.341
EC Interpersonal Coordination	-0.300	**0.024***	0.164	0.223	0.221	**0.098+**	0.040	0.766
Trait Shyness	0.302	**0.022***	-0.071	0.599	-0.166	0.218	-0.016	0.904
KiSS-18	-0.331	**0.012***	0.102	0.451	0.127	0.346	0.011	0.934

+: $p < 0.1$, *: $p < 0.05$

logue acts assigned to the system's expected responses to user utterances (Table 7).

For the dialogue act Social Repair, significant correlations were found with five personal traits: Neuroticism, Expressivity, Interpersonal Coordination, Trait Shyness, and KiSS-18. Weak positive correlations were observed with Neuroticism and Trait Shyness. This suggests that individuals who are more emotionally reactive or who tend to experience social anxiety are more likely to expect the system to apologize or correct itself when a breakdown occurs. Conversely, weak negative correlations were observed with Expressivity, Interpersonal Coordination, and KiSS-18. Participants who can clearly express their thoughts and feelings, maintain harmonious relationships, and possess strong social skills - thus being able to facilitate smooth communication - are less likely to expect the system to issue apologies or corrections following a breakdown.

4.3. Transition Probabilities of System Utterances Following User Responses by Breakdown Type

Since user utterances labeled as Continuation always lead to the system continuing the dialogue without performing any repair, these cases were excluded from the analysis.

As shown in Table 8, both the distribution of user dialogue acts and that of system dialogue acts, as well as their transition probabilities, vary across breakdown types. For example, in wrong information cases, Social Repair is likely to be selected irrespective of the User Utterance type. In contrast, for Lack of Information, the dialogue act used for Repair differs depending on the User Utterance. This indicates that the types of dialogue acts users produce after a breakdown differ depending on the breakdown type, and accordingly, the repair strategies expected from the system also differ. Therefore, it is necessary to design repair strategies that take the breakdown type into account.

5. Conclusions

In this study, we introduced a corpus designed to support dialogue breakdown repair tailored to users' individual traits. We collected personality questionnaire data and 3,990 dialogue repair examples from 57 participants. Dialogue act tags were annotated for the collected data, and analyses were conducted on the distribution of perceived breakdown severity and the dialogue acts of repair utterances. The results revealed that both the perception of breakdowns and the preferred repair dialogue acts vary depending on the type of breakdown and the individual user. In future work, we plan to generate repair utterances automatically.

6. Acknowledgements

This work is supported by Adaptable and Seamless Technology transfer Program through Target-driven R&D (A-STEP) from Japan Science and Technology Agency (JST) Japan Grant Number JPMJTR25R5. This work was supported by JSPS KAKENHI Grant Number JP23H00493.

Table 8: Conditional probabilities [%] of dialogue act types in the system's expected responses to user utterances after breakdowns, by breakdown type

Breakdown Type	User Utterance	No Repair	Information Addition	Confirmation	Social Repair
Wrong information	Breakdown Indication	9.3	6.0	7.3	72.0
	Breakdown Indication + Confirmation	16.7	16.7	6.7	53.3
	Confirmation	40.0	5.0	0.0	55.0
Ignore Question	Breakdown Indication	26.1	15.9	1.4	42.0
	Breakdown Indication + Confirmation	36.2	17.0	0.0	42.6
	Confirmation	64.8	15.2	3.8	10.5
Ignore Expectation	Breakdown Indication	100.0	0.0	0.0	0.0
	Breakdown Indication + Confirmation	-	-	-	-
	Confirmation	42.7	25.8	3.1	1.7
Unclear Intention	Breakdown Indication	34.5	23.6	5.5	29.1
	Breakdown Indication + Confirmation	26.9	34.6	3.8	30.8
	Confirmation	39.1	32.7	0.9	15.5
Topic Transition Error	Breakdown Indication	30.3	21.2	4.5	30.3
	Breakdown Indication + Confirmation	17.2	34.5	0.0	37.9
	Confirmation	47.5	28.8	1.7	10.2
Lack of Information	Breakdown Indication	28.3	26.1	19.6	23.9
	Breakdown Indication + Confirmation	11.1	33.3	0.0	40.7
	Confirmation	40.8	30.8	4.7	17.8
Self-contradiction	Breakdown Indication	19.2	28.3	0.8	44.2
	Breakdown Indication + Confirmation	10.5	26.3	0.7	48.0
	Confirmation	15.7	37.1	4.3	38.6
Contradiction	Breakdown Indication	34.5	12.5	5.6	42.2
	Breakdown Indication + Confirmation	21.2	21.2	1.9	44.2
	Confirmation	24.1	37.9	6.9	24.1
Repetition	Breakdown Indication	37.3	27.1	3.4	25.4
	Breakdown Indication + Confirmation	50.0	30.0	0.0	10.0
	Confirmation	55.8	27.9	0.0	0.0
Lack of Common Sense	Breakdown Indication	12.1	14.2	3.3	65.3
	Breakdown Indication + Confirmation	6.2	30.8	4.6	56.9
	Confirmation	21.4	39.3	7.1	28.6

7. Ethical Considerations and Limitations

We used a crowdsourcing platform to conduct the annotation tasks and provided appropriate compensation to all participants. Participants were informed in advance that their anonymized responses would be made publicly available. Researchers must also ensure the protection of the privacy and personal information of all crowd workers.

Since only short dialogue segments consisting of two to four utterances were presented, the corpus does not cover breakdowns that occur in longer or more complex conversational contexts. In addition, because the dataset is limited to Japanese, different results may be observed in other cultural or linguistic settings. Furthermore, when developing dialogue systems that take individual traits into account, users must be informed that their personality traits are being estimated and utilized by the system.

8. Bibliographical References

Atsushi Aikawa. 1991. A study on the reliability and validity of a scale to measure shyness as a trait (in japanese). *The Japanese journal of psychology*, 62(3):149–155.

Vevake Balaraman, Arash Eshghi, Ioannis Konstas, and Ioannis Papaioannou. 2023. No that's not what I meant: Handling third position repair in conversational question answering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 562–571, Prague, Czechia. Association for Computational Linguistics.

Dennis Benner, Edona Elshan, Schöbel, Sofia, and Andreas Janson. 2021. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS) 2021*.

Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. *Safety First: Conversational Agents for Health Care*, pages 33–57. Springer International Publishing, Cham.

Chayan Chakrabarti and George F. Luger. 2015. Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, 42(20):6878–6897.

Souvik Das, Sougata Saha, and Rohini Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi,

United Arab Emirates. Association for Computational Linguistics.

Manabu Fujimoto and Ikuo Daibo. 2007. Endcore: A hierarchical structure theory of communication skills (in japanese). *The Japanese Journal of Personality*, 15(3):347–361.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2022. Classification of utterances that lead to dialogue breakdowns in chat-oriented dialogue systems (in japanese). *Journal of Natural Language Processing*, 29(2):443–466.

Ryuichiro Higashinaka, Luis F. D' Haro, Bayan Abu Shawar, Rafael E. Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and João Sedoc. 2021. Overview of the Dialogue Breakdown Detection Challenge 4. In Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno, and Haizhou Li, editors, *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Lecture Notes in Electrical Engineering, pages 403–417. Springer, Singapore.

Akio Kikuchi. 1988. *The science of compassion: The psychology and skills of prosocial behavior (in Japanese)*. Kawashima Shoten.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j) (in japanese). *The Japanese Journal of Personality*, 21(1):40–52.

Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2024. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2979–2984.

Mao Saeki, Hiroaki Takatsu, Fuma Kurata, Shungo Suzuki, Masaki Eguchi, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama. 2024. InteLLA: Intelligent language learning assistant for assessing language proficiency through interviews and roleplays. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 385–399, Kyoto, Japan. Association for Computational Linguistics.

Kazuya Tsubokura, Yurie Iribe, and Norihide Kitaoka. 2024a. Analysis of the relationship between user response to dialog breakdown and personality traits. *Advanced Robotics*, 38(4):246–255.

Kazuya Tsubokura, Yurie Iribe, and Norihide Kitaoka. 2025. Relationship between user's response and individual traits on spoken dialogue system. *The IEICE Transactions on Information and Systems(Japanese Edition)*, J108-D(4):182–191.

Kazuya Tsubokura, Takuya Takeda, Yurie Iribe, and Norihide Kitaoka. 2024b. Dialog breakdown recovery strategies based on user personality. In *The 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS2024)*.

Takahisa Uchida, Takashi Minato, Tora Koyama, and Hiroshi Ishiguro. 2019. Who is responsible for a dialogue breakdown? an error recovery strategy that promotes cooperative intentions from humans by mutual attribution of responsibility in human-robot dialogues. *Frontiers in Robotics and AI*, 6.

Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, and Dong Yu. 2024. Inconsistent dialogue responses and how to recover from them. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 220–230, St. Julian's, Malta. Association for Computational Linguistics.

A. Definitions and Examples of Breakdown Types

Table 9 shows the definitions and examples of breakdown types.

B. Prompts for Dialogue Act Annotation

The prompts used for dialogue act tag annotation are shown in Tables 10 and 11. Table 10 presents the prompt used for user responses to the system's broken utterances, while Table 11 presents the prompt used for the system's expected responses to the user's utterances. In the zero-shot condition, the prompts were identical except that the examples shown in the tables were removed. Note that

the prompts used in the experiments were originally written in Japanese, and the versions shown here are English translations.

Table 9: Definitions and examples of breakdown types

Breakdown types	Definition and example
Wrong information	An utterance that contains information clearly inconsistent with factual knowledge. System: The capital of Japan is Osaka.
Ignore question	An utterance that ignores the user's question. User: What's your favorite food? System: It's hot today.
Ignore expectation	An utterance that formally satisfies the expected response function but does not contain the information expected by the user. Such responses may provide too little information or be overly indirect, resulting in an inadequate response. User: Do you have a favorite manga? System: Yes. *The system is expected to specify the manga.*
Unclear intention	An utterance whose literal meaning and topical relation to the conversation can be understood, but whose intention (i.e., why the speaker said it) is unclear and cannot be inferred from the context. This may occur when elements expressing relationships, such as conjunctions or contextual explanations, are missing. User: The next class is math, right? System: Yes, that's right. User: I'm not good at math, so I'm worried. System: Thank you. *The system expresses gratitude even though there is nothing to thank for.*
Topic transition error	An utterance that shifts to a different topic from the previous utterance without explanation. This also includes returning to a previous topic without justification when the user has already moved on to a new topic. User: What did you do this weekend? System: I went mountain climbing. User: Nice. I'd like to go climbing too. System: Do you want to go for yakiniku today? *The topic suddenly shifts from mountain climbing to yakiniku.*
Lack of information	An utterance that lacks necessary elements (e.g., modifiers, subjects, or objects), making it difficult or impossible to understand what is being referred to. User: I ate curry. System: That's impressive. *It is unclear what is impressive.*
Self-contradiction	An utterance that contradicts something previously stated by the system itself. User: What did you have for dinner? System: I ate miso ramen. User: Oh, miso ramen is delicious. System: I haven't eaten dinner yet, so I'm hungry. *It contradicts what the system said earlier.*
Contradiction (with user utterance)	An utterance that contradicts something previously stated by the user. This also includes questions that negate the user's statement. User: What's your favorite food? System: Probably chocolate cake. What about you? User: I like caramel sandwiches. System: So you don't like caramel?
Repetition	An utterance that simply repeats a previous utterance made by the system. This also includes paraphrased repetitions or repeatedly asking the same question. User: Do you have a favorite food? System: I like sushi. User: Sounds good. I crave it sometimes too. System: I like sushi.
Lack of common sense	An utterance that contains statements inconsistent with common sense. This includes making claims without evidence that contradict generally accepted facts or beliefs. User: Actually, today is my birthday. System: That's disappointing.

Table 10: Prompt for dialogue act tag annotation of user responses to the system's broken utterances (Few-shot)

Please assign a dialogue act tag to the last user utterance in the dialogue history.
If multiple tags can be assigned within the utterance, assign all applicable tags. In that case, list the tags in the order in which they appear in the utterance.
The dialogue act tags are defined as the following four types.

Definition of Dialogue Act Tags
Breakdown Indication: The user points out that the system's utterance is incorrect or broken, or corrects/denies it.
Confirmation: The user asks for clarification about the system's intention or missing information in order to resolve the breakdown.
Continuation: The user continues the conversation without explicitly pointing out or confirming the breakdown.
Other: Any case that does not fall into the above categories.

Examples
User utterance: The capital of Japan is Tokyo.
System utterance: That's right. Tokyo is also famous for Universal Studios Japan.
User utterance: That's not in Tokyo.
Dialogue Act: Breakdown Indication

User utterance: Do you have any siblings?
System utterance: Video games are fun, aren't they?
User utterance: Are you listening?
Dialogue Act: Breakdown Indication

User utterance: Yesterday's World Cup match was exciting.
System utterance: Yeah, I like it.
User utterance: Do you mean you like soccer?
Dialogue Act: Confirmation

Dialogue History
User utterance: {first_user_utterance}
System utterance: {first_system_utterance}
User utterance: {second_user_utterance}
System utterance: {system_broken_utterance}
User utterance: {user_response_to_broken_utterance}

Table 11: Prompt for dialogue act tag annotation of the system's expected responses to user responses (Few-shot)

Please assign a dialogue act tag to the last system utterance in the dialogue history.
If multiple tags can be assigned within the utterance, assign all applicable tags. In that case, list the tags in the order in which they appear in the utterance.
The dialogue act tags are defined as the following five types.

Definition of Dialogue Act Tags
No Repair: The system continues the conversation without repairing the breakdown indicated or questioned by the user.
Confirmation: The system asks for clarification about intention or missing information in order to address the user's breakdown indication or confirmation.
Information Addition: The system adds explanations or excuses to provide additional information to the user in order to address the user's breakdown indication or confirmation.
Social Repair: The system addresses the user's breakdown indication or confirmation by apologizing, acknowledging its mistake, or correcting its previous statement.
Other: Any case that does not fall into the above categories.

Note that when the tag is "No Repair", no other tags should be assigned.

Examples
User utterance: That's a bit expensive.
System utterance: The temperature difference is terrible, isn't it?
User utterance: What are you talking about?
System utterance: Aren't we talking about temperature?
Dialogue Act: Confirmation

User utterance: I went to watch a basketball game yesterday.
System utterance: Really? I was on the basketball team in middle and high school.
User utterance: Oh really? Let's play together sometime.
System utterance: I've never played basketball before, is that okay?
User utterance: Which one is true?
System utterance: I was on the basketball team, but I never participated in practice.
Dialogue Act: Information Addition

User utterance: What's the capital of France?
System utterance: I think it's Rome.
User utterance: Isn't Rome in Italy?
System utterance: Sorry, it's Paris.
Dialogue Act: Social Repair

Dialogue History
User utterance: {first_user_utterance}
System utterance: {first_system_utterance}
User utterance: {second_user_utterance}
System utterance: {system_broken_utterance}
User utterance: {user_response_to_broken_utterance}
System utterance: {system_expected_response_to_user_response}