Minimal Yet Big Impact: How AI Agent Back-channeling Enhances Conversational Engagement through Conversation Persistence and Context Richness

Jin Yea Jang[1,2], Saim Shin[2], and Gahgene Gweon[1,3]

[1]Department of Intelligence and Information, Seoul National University, Republic of Korea
[2]Artificial Intelligence Research Center, Korea Electronics Technology Institute, Republic of Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Republic of Korea

Abstract

The increasing use of AI agents in conversational services, such as counseling, highlights the importance of back-channeling (BC) as an active listening strategy to enhance conversational engagement. BC improves conversational engagement by providing timely acknowledgments and encouraging the speaker to continue talking. This study investigates the effect of BC provided by an AI agent on conversational engagement, offering insights for the future design of AI conversational services. We conducted an experiment with 55 participants, divided into Todak_BC and Todak_NoBC groups based on the presence or absence of the BC feature in *Todak*, a conversational agent. Each participant engaged in nine sessions with predetermined subjects and questions. We collected and analyzed approximately 6 hours and 30 minutes of conversation logs to evaluate conversational engagement using both quantitative (*conversation persistence*, including *conversation duration* and *number of utterances*) and qualitative metrics (*context richness*, including *self-disclosure* and *topic diversity*). The findings reveal significantly higher conversational engagement in the Todak_BC group compared to the Todak_NoBC group across all metrics ($p < 0.05$). Additionally, the impact of BC varies across sessions, suggesting that conversation characteristics such as question type and topic sensitivity can influence BC effectiveness.

1 Introduction

The increasing demand for AI-driven counseling and conversational services (Prochaska et al., 2021; Park et al., 2023) is driven by their benefits of anonymity, accessibility, and lack of biases (Chen and Lucock, 2022; Morrow and Deidan, 1992; Nosrati et al., 2020), highlighting the importance of back-channeling (BC) as a crucial element of active listening strategies (Rost and Wilson, 2013; Weger Jr et al., 2014) to enhance client conversational engagement. Counseling experts emphasize the importance of active client participation for effective counseling (Tryon, 1990; Simpson et al., 2009). Active conversational engagement involves clients actively participating and communicating their feelings and thoughts. BC includes verbal and non-verbal cues like nodding or saying "uh-huh" in response to the speaker (Li et al., 2010), which express interest and enhance the speaker's conversational engagement (Veach et al., 2007).

A comprehensive understanding of the impact of BC presented by AI agents on conversational engagement is crucial for developing AI systems, especially for applications requiring deep interactions, such as counseling. Despite its importance, limited research has directly examined BC's impact on engagement. Most existing studies focus on user perceptions of BC configurations (De Sevin et al., 2010; Kim et al., 2021; Meywirth and Götze, 2022) or are conducted in non-conversational contexts (Andriella et al., 2020). While some research explores the relationship between BC and engagement (Cho et al., 2022), it primarily addresses emotional aspects of user utterances without demonstrating BC's role in enhancing overall engagement, such as conversation persistence. Moreover, few studies analyze actual AI-human conversation logs, which is crucial for understanding BC's effectiveness in real-world AI services.

To address these gaps in the literature, our study aims to provide a more comprehensive understanding of how BC influences conversational engagement. Specifically, we analyze conversations between users and AI agents with BC features, focusing on both quantitative and qualitative aspects. Our research seeks to answer the following question: *RQ: How does the presence of back-channeling (BC) features in AI agents influence conversational engagement, both quantitatively and qualitatively?*

To answer this question, we conducted a user ex-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14509–14521
November 12-16, 2024 ©2024 Association for Computational Linguistics

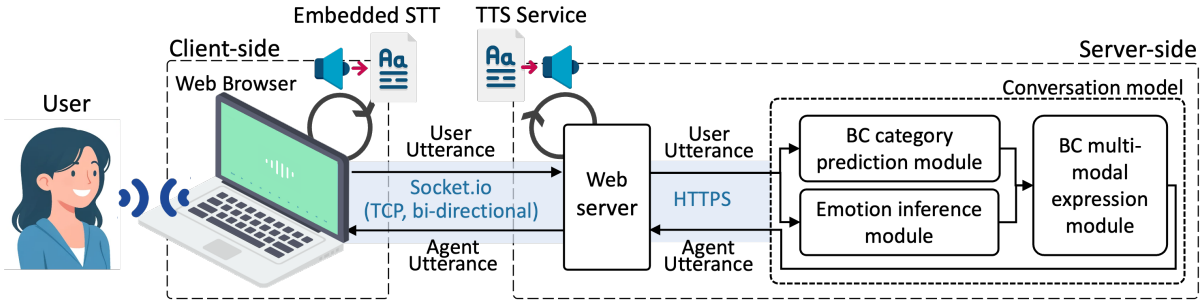


Figure 1: The system architecture of Todak

periment with Todak, an AI-based conversational system featuring BC. 55 participants were divided into two groups based on the presence or absence of BC, and conversation logs were collected across nine sessions, totaling approximately 6 hours and 30 minutes. We analyzed these logs using four metrics. The results showed that users interacting with the AI agent expressing BC had significantly higher conversational engagement scores than those interacting without BC ($p < 0.05$), demonstrating the effectiveness of BC expressed by the AI agent. Additional analysis indicated that the effectiveness of BC varies with conversation characteristics.

The contributions of our research are as follows: First, we propose BC as a method to enhance conversational engagement in AI interactions and validate its effectiveness. Second, we assessed conversational engagement using quantitative and qualitative metrics from actual conversation logs, a method overlooked in previous studies. Our findings confirm that the utility of BC varies with conversation characteristics, providing valuable guidance for the future design of AI agent conversations.

2 Back-channeling and Conversational Engagement

BC signals from a listener in human-to-human conversation play a crucial role as an attentive listening strategy in eliciting the speaker's speech and persisting the conversation (Sacks, 1978; Sadock and Sadock, 2011), thus contributing to conversational engagement (Ward and Tsukahara, 2000; Bavelas et al., 2000; Gardner, 2001). Conversation persistence measures how long the conversation lasts and how much a speaker talks, serving as a quantitative indicator of engagement. To investigate whether BC in AI agents contributes to conversation persistence, we designed an AI agent called Todak and conducted a user experiment. Participants were divided into two groups: Todak_BC and Todak_NoBC, and each participant engaged in conversational sessions. The first hypothesis is:

Hypothesis 1: Conversation Persistence

- *H1: The Todak_BC group will have longer conversations and a higher number of utterances compared to the Todak_NoBC group.*

In human-to-human conversation, BC also enhances context richness by promoting intimacy and trust, leading to increased self-disclosure and deeper conversations (Schegloff, 1982; Bavelas et al., 2000; Gardner, 2001). Additionally, BC aids in narrative formation (Bavelas et al., 2000; Tolins and Tree, 2014), increasing topic diversity and making discourse richer. Conversation context richness assesses how deep and varied a speaker's conversation is, serving as a qualitative indicator of engagement. The second hypothesis is:

Hypothesis 2: Conversation Context Richness

- *H2-a: The Todak_BC group will have a higher degree of self-disclosure compared to the Todak_NoBC group.*
- *H2-b: The Todak_BC group will have a greater diversity of topics compared to the Todak_NoBC group.*

3 Back-channel enabled AI agent - Todak

The system architecture of the agent Todak[1], which is equipped with the ability to express BC, is depicted in Figure 1. Todak operates as a web-based conversational agent system, consisting of both client-side and server-side components. On the client-side, user speech input signals are captured via a web browser, which are then converted to text using an embedded Speech-to-Text (STT) service. This text is forwarded to the server-side for

[1]The name "Todak" comes from the Korean verb "토닥이다 (/tɔdɑkiɑ/)," which means to gently tap someone on the back or arms as a sign of empathy or comfort.

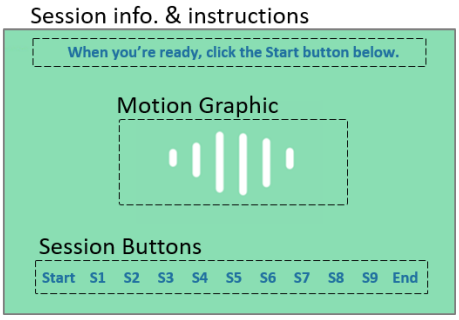


Figure 2: The user interface of Todak

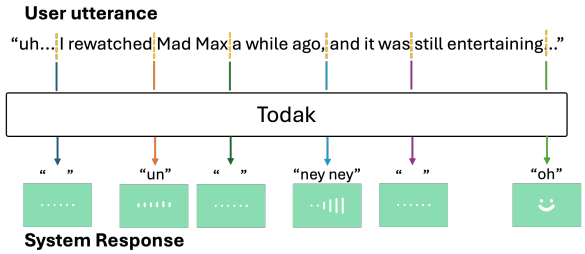


Figure 3: An example of the system response for a user utterance input

further processing. The server-side comprises a web server and a conversation model. The web server transmits the user speech text to the conversation model, which generates the agent's response. This response is then transformed back into speech through a Text-to-Speech (TTS) service[2] and sent back to the client-side. Real-time interaction is essential for the agent's BC responses during user speech. Consequently, bidirectional communication between the client-side and server-side is managed using Socket.IO. Finally, the web browser on the client-side delivers Todak's audio response to the user. To ensure the user's speech is not interrupted, Todak's BC voice volume is set lower than the front-channel voice volume.

Todak's BC is crafted to reflect the multi-modal nature of BC (Young and Lee, 2004), incorporating both verbal and non-verbal expressions. Verbal BC is delivered through voice, whereas non-verbal BC is communicated to the user via various motion graphics inspired by voice waveforms. Figure 2 displays a screenshot of the Todak agent from the user's perspective. The agent's non-human-like design was chosen to prevent any interference from participants' potential biases regarding the agent's appearance (Van Vugt et al., 2006).

Todak's BC expressions are functionally categorized to capture the diverse range of conversational cues. Following (Cutrone, 2010), Todak's BCs are divided into three categories: "continuer," which signals attentiveness to the speaker; "understanding," which indicates comprehension and agreement; and "empathic response," which demonstrates empathy and emotion. Additionally, the absence of BC expression is defined as "no expression." A detailed description of verbal and non-verbal expressions according to Todak's BC categories can be found in Appendix A.

The BC category and expression are determined by Todak's conversation model based on the user's utterance input. As shown in Figure 3, the user's input is streamed, and Todak responds with the corresponding BC expression in real time. The conversation model concurrently predicts both the BC category and the user's emotion, ensuring that the "empathic response" category aligns with the user's emotion (positive to positive, negative to negative). Consequently, the conversation model consists of three components: a BC category prediction module, an emotion inference module, and a BC multi-modal expression module. The BC category prediction module is implemented based on Jang et al. (2021), utilizing the language model KE-T5[3] encoder. The emotion inference module is based on Lim et al. (2021) and infers five emotions: "neutral," "happiness," "surprise," "sadness," and "anger." The BC multi-modal expression module determines the final expressions. For each inferred BC category, verbal and non-verbal expressions are randomly selected with equal probability and combined in the module. The "no expression" category allows BC to exhibit either single or multi-modal characteristics (Young and Lee, 2004). Through this random combination, it is possible to have single-modal BC with either only verbal or only non-verbal expressions, as "no expression" indicates the absence of an expression in the respective modality. Thus, the BC multi-modal expression module can generate diverse BC responses, enhancing the naturalness of the conversation.

4 Methods

4.1 Overview

To examine the effect of BC on conversational engagement, we conducted a between-subject experiment. Participants were randomly divided into two groups: one interacted with Todak with BC (Todak_BC) and the other without BC (Todak_NoBC).

[2]https://clova.ai/voice/

[3]https://huggingface.co/KETI-AIR/ke-t5-base

Subject sensitivity	Subject	Session	Question	Question type
-	-	Start	(Greetings)	-
low	meal menu	1	“Can you please tell me what you ate yesterday?"	open-ended
		2	“Which meal have you enjoyed the most?"	closed-ended
	media content	3	“Can you tell me about a movie or TV show you’ve enjoyed recently?"	open-ended
		4	“Do you like genres similar to those you just mentioned?"	closed-ended
medium	personality	5	“Do your favorite genres reflect your personality?"	closed-ended
		6	“How do you think your personality differs from how others perceive it?"	open-ended
high	stress & worries	7	“What do you do when you feel stressed or worried?	open-ended
		8	“Tell me about someone you feel comfortable talking to about your worries."	open-ended
		9	“Can you tell me about something you’ve been worrying about lately?"	open-ended
-	-	End	(Closing remarks)	-

Table 1: Designed conversation sessions

The interactions were structured into nine sessions to maintain consistency in conversation subjects. Conversational engagement was evaluated using four metrics: conversation duration, number of utterances, self-disclosure, and topic diversity.

4.2 Participants

Participants were randomly assigned to either the Todak_BC or Todak_NoBC groups, and we confirmed that the groups were well-balanced in terms of gender, age, and prior experience with voice assistants. Initially, 64 participants were recruited through snowball sampling and evenly assigned to the two groups. However, due to differences in participant consent for conversation log collection, the final data was collected from 55 participants. The Todak_BC group consisted of 30 participants (15 males, 15 females), with a mean age of 24.69 (SD=6.81) and an average voice assistant experience score of 3.77 (SD=1.19). The Todak_NoBC group included 25 participants (12 males, 13 females), with a mean age of 25.00 (SD=7.07) and an average voice assistant experience score of 3.62 (SD=1.17). Voice assistant experience was measured on a 1-5 Likert scale.

4.3 Conversation session design

To ensure subject consistency across groups, we provided structured guidance with nine session subjects. A "session" is defined as a distinct segment of the overall conversation with a specific subject and guiding question. The subjects and questions for each session are detailed in Table 1. The conversation comprises nine sessions, excluding the start and end sections. Sessions consist of primary questions and follow-up questions to maintain conversational flow. The nine session questions naturally fall into two categories: closed questions, which can be answered with "yes," "no," or a short answer, and open questions, which require more detailed responses (Foddy and Foddy, 1993). The conversation subjects were designed to progress from light topics, answerable through immediate recall, to deeper and potentially more sensitive issues, similar to real conversations. The conversation begins with low-sensitivity subjects like meal menus and favorite media content and gradually transitions to higher-sensitivity subjects, including personality traits and, eventually, stress and worries, allowing for a deeper exploration of the participant’s thoughts and feelings.

4.4 Procedure

The user experiment followed this sequence: obtaining consent, briefing participants about the experiment, conducting a pre-survey, and then interacting with the agent. Participants freely responded to questions in each session and moved to the next session by pressing the session button shown in Figure 2. The average duration of the experiment was approximately 30 minutes, and participants were compensated $25 for their participation.

4.5 Measures

4.5.1 Conversation persistence: conversation duration and number of utterances

Qualitatively assess participants’ conversation engagement by measuring conversation persistence, which includes conversation duration and the number of utterances (Ghazarian et al., 2020). Conversation duration for each session is defined as the time span from start to end, measured in seconds. The experiment system records a timestamp whenever the session button is clicked. Using these timestamps, the duration of each session is calculated by subtracting the current session’s timestamp from the next session’s timestamp. To assess overall conversational persistence, the total conversation duration

H1: Conversation Persistence								
Session (Question type & subject sensitivity)	Conversation duration (sec) Mean (SD)		Statistic	p-value	Number of utterances Mean (SD)		Statistic	p-value
	Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC		
1 (open-ended & low)	44.0 (30.5)	43.0 (34.2)	U=367	0.446	7.8 (5.4)	4.6 (2.4)	U=215	0.003**
2 (closed-ended & low)	44.5 (25.4)	28.0 (13.6)	U=182	0.000***	4.2 (2.8)	2.3 (1.9)	U=193	0.000***
3 (open-ended & low)	52.5 (34.4)	51.8 (35.3)	U=375	0.507	8.6 (7.8)	6.2 (6.7)	U=280	0.054
4 (closed-ended & low)	37.2 (20.4)	27.9 (20.1)	U=244	0.014*	4.4 (2.7)	3.0 (1.9)	U=262	0.027*
5 (closed-ended & medium)	51.1 (20.8	41.4 (25.0)	U=266	0.033*	6.6 (3.5)	4.8 (3.6)	U=254	0.020*
6 (open-ended & medium)	48.4 (24.2)	41.7 (24.3)	U=325	0.201	7.3 (5.3)	5.8 (3.8)	U=316	0.158
7 (open-ended & high)	48.8 (33.3)	33.8 (25.2)	U=257	0.023*	7.2 (4.9)	4.8 (3.6)	U=264	0.030*
8 (open-ended & high)	51.2 (32.5)	45.5 (38.6)	U=306	0.123	8.0 (6.6)	6.4 (6.9)	U=288	0.071
9 (open-ended & high)	79.2 (82.3)	39.5 (55.3)	U=187	0.000***	10.8 (8.1)	5.3 (3.6)	U=190	0.000***
Total	**498.8 (194.7)**	**369.7 (158.7)**	**U=234**	**0.009****	**64.9 (36.4)**	**43.1 (27.96)**	**U=230**	**0.007****

Total N = 55, *< 0.05, **< 0.01, ***< 0.001.

Table 2: Statistical test results for the conversation persistence

H2 Conversation Context Richness								
Session (Question type & subject sensitivity)	H2-a: Self-disclosure Mean (SD)		Statistic	p-value	H2-b:Topic diversity Mean (SD)		Statistic	p-value
	Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC		
1 (open-ended & low)	4.87 (0.82)	4.32 (0.48)	U=229	0.003**	0.239 (0.14)	0.207 (0.11)	t(53)=0.943	0.177
2 (closed-ended & low)	6.50 (1.04)	5.56 (1.26)	U=211	0.002**	0.165 (0.09)	0.146 (0.07)	t(53)=0.887	0.190
3 (open-ended & low)	5.73 (1.26)	5.28 (1.40)	U=322	0.178	0.322 (0.20)	0.318 (0.20)	t(53)=0.077	0.469
4 (closed-ended & low)	5.33 (0.99)	5.00 (1.12)	U=300	0.090	0.172 (0.10)	0.182 (0.16)	t(53)=-0.297	0.616
5 (closed-ended & medium)	5.56 (1.43)	4.80 (1.12)	U=252	0.015*	0.281 (0.18)	0.221 (0.17)	t(53)=1.264	0.106
6 (open-ended & medium)	5.97 (1.38)	5.76 (1.42)	U=364	0.424	0.269 (0.17)	0.259 (0.18)	t(53)=0.221	0.413
7 (open-ended & high)	5.80 (1.27)	4.96 (1.46)	U=268	0.031*	0.334 (0.20)	0.306 (0.22)	t(53)=0.499	0.310
8 (open-ended & high)	6.23 (1.38)	5.64 (1.15)	U=308	0.111	0.346 (0.19)	0.311 (0.19)	t(53)=0.685	0.248
9 (open-ended & high)	8.03 (1.40)	7.00 (1.94)	U=222	0.003**	0.331 (0.22)	0.225 (0.16)	t(53)=2.014	0.025*
Total	**54.07 (6.86)**	**48.32 (7.00)**	**U=230**	**0.007****	**0.516 (0.10)**	**0.460 (0.14)**	**t(53)=1.696**	**0.048***

Total N = 55, *< 0.05, **< 0.01, ***< 0.001.

Table 3: Statistical test results for the conversation context richness

for each participant is obtained by summing the durations of all sessions.

The number of utterances refers to the count of speech instances generated by a user during interactions with Todak. An utterance is defined based on the recognition results of the Web Speech API, which uses Speech-To-Text (STT) technology[4]. The STT module signals the end of an utterance with an information flag: 'false' indicates the utterance is ongoing, while 'true' indicates it has ended. An utterance is considered complete when this flag is 'true'. The number of utterances for each session is the count of utterances within that session. To observe overall conversation persistence, the total number of utterances for each participant is calculated by summing the utterances across all sessions.

4.5.2 Context richness: self-disclosure and topic diversity

Qualitatively assess participants' conversational engagement through context richness measures, including self-disclosure and topic diversity in conversation logs. Self-disclosure was measured using the guidelines for the three-dimension (information, thoughts, and feelings) by Barak and Gluck-Ofri (2007). Scores range from 1 to 3, with higher scores indicating greater self-disclosure. We used OpenAI GPT-4[5], which provides human-level, high-quality annotations (Gilardi et al., 2023), for scoring, employing the "Self-disclosure Analyzer" tool to automatically measure self-disclosure dimensions. The prompt of the "Self-disclosure Analyzer" and an example can be seen in Appendix B. To validate the annotations, we calculated Cohen's kappa scores for 10 participants' logs with two linguistic experts, showing high inter-rater reliability (0.79). Scores for each session were measured as the sum of the scores for information, thoughts, and feelings. The total score for each participant was then calculated by summing the scores across all sessions.

Topic diversity measures how many different topics are covered within a single conversational session. To measure topic diversity in participants' utterances, we used a two-step process. First, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) was applied to infer N topics: N=5 for intra-session analysis and N=20 for total session analysis. Al-

[4] https://developer.mozilla.org/en-US/docs/Web/API/SpeechRecognition

[5] https://openai.com/gpt-4

Session	Self-disclosure											
	Information				Thoughts				Feelings			
	Mean (SD)		Statistic	p-value	Mean (SD)		Statistic	p-value	Mean (SD)		Statistic	p-value
	Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC		
1	2.67 (0.48)	2.32 (0.48)	U=245	0.006**	1.10 (0.31)	1.00 (0.00)	U=338	0.056	1.10 (0.31)	1.00 (0.00)	U=338	0.056
2	2.07 (0.45)	1.80 (0.65)	U=288	0.036**	2.03 (0.41)	1.76 (0.44)	U=282	0.012**	2.40 (0.50)	2.00 (0.50)	U=246	0.003**
3	1.73 (0.45)	1.64 (0.49)	U=340	0.233	1.97 (0.49)	1.84 (0.55)	U=333	0.181	2.03 (0.72)	1.80 (0.58)	U=309	0.106
4	1.50 (0.57)	1.32 (0.56)	U=309	0.095	1.97 (0.41)	1.80 (0.50)	U=317	0.088	1.87 (0.43)	1.88 (0.33)	U=369	0.573
5	1.87 (0.57)	1.68 (0.56)	U=315	0.118	2.30 (0.65)	1.92 (0.57)	U=258	0.013*	1.43 (0.68)	1.20 (0.41)	U=318	0.108
6	2.27 (0.58)	2.20 (0.65)	U=357	0.368	2.37 (0.56)	2.40 (0.65)	U=357	0.641	1.33 (0.61)	1.16 (0.47)	U=334	0.184
7	2.17 (0.65)	1.80 (0.58)	U=265	0.017*	2.33 (0.55)	1.96 (0.61)	U=261	0.013*	1.30 (0.60)	1.20 (0.76)	U=344	0.259
8	2.50 (0.51)	2.36 (0.57)	U=330	0.194	2.43 (0.50)	2.24 (0.60)	U=316	0.124	1.30 (0.60)	1.04 (0.35)	U=304	0.038*
9	2.73 (0.52)	2.56 (0.65)	U=325	0.146	2.80 (0.48)	2.64 (0.64)	U=331	0.153	2.50 (0.57)	1.80 (0.76)	U=188	0.000***
Total	**19.50 (3.12)**	**17.68 (3.35)**	**U=266**	**0.032***	**19.30 (2.42)**	**17.56 (2.95)**	**U=270**	**0.037***	**15.27 (2.88)**	**13.08 (1.98)**	**U=204**	**0.002****

Total N = 55, *< 0.05, **< 0.01, ***< 0.001.

Table 4: Statistical test results based on the self-disclosure three dimensions: information, thoughts, and feelings

though LDA can struggle with data sparsity in short texts (Wu et al., 2020), we chose LDA for this study because the average length of utterances in our dataset (35.8 tokens) is longer than what is typically considered as short text (4.1 to 10.3 tokens) (Wu et al., 2020), making it suitable for our analysis. Next, we computed topic diversity scores, ranging from 0 to 1, using three metrics: Proportion of Unique Words (PUW), average pairwise Jaccard Distance (JD), and Inverted Rank-Biased Overlap (IRBO) (Dieng et al., 2020; Tran et al., 2013; Bianchi et al., 2021). The detailed procedure can be found in Appendix C. The average of these metrics was defined as the topic diversity score for each session and for the total sessions.

4.6 Analysis

To verify our hypotheses, we conducted a mean comparison test on the conversational measurement data of the Todak_BC and Todak_NoBC groups, analyzing approximately 6 hours and 30 minutes of conversation logs collected from the user experiment. Depending on the results of the data normality test, we used either the parametric Student's t-test or the non-parametric Mann-Whitney U test for the mean comparison.

5 Results

The mean of the Todak_BC group was significantly higher than that of the Todak_NoBC group for all conversational engagement measures, including conversation duration (p=0.009), number of utterances (p=0.007), self-disclosure (p=0.007), and topic diversity (p=0.048), as shown in Table 2 and Table 3. Additionally, the effect size for each measure was examined to assess the practical significance of these differences. Specifically, conversation duration showed a rank-biserial correlation of $r_{rb} = 0.38$, number of utterances had $r_{rb} = 0.39$, self-disclosure also had $r_{rb} = 0.39$, and topic diversity exhibited a Cohen's d of 0.46, all of which indicate moderate effects. However, the differences varied across individual sessions. The results for each metric and session are detailed in the following subsections.

5.1 How Todak's back-channeling influences conversation persistence

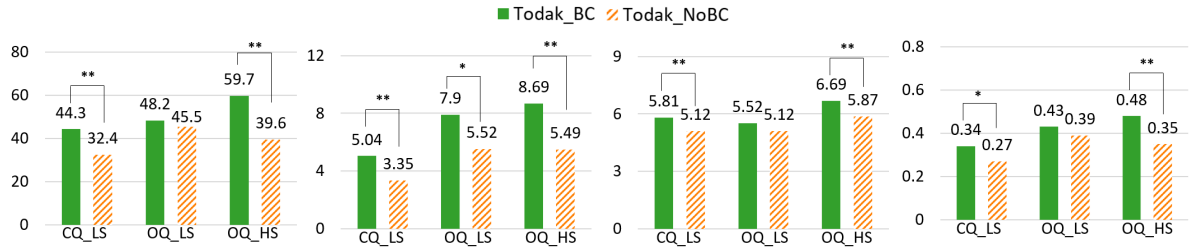
Conversation duration

The mean conversation duration of the Todak_BC group was significantly higher than that of the Todak_NoBC group across all sessions (p=0.009). Specifically, the Todak_BC group averaged 498.8 seconds compared to 369.7 seconds for the Todak_NoBC group, indicating a 130-second longer conversation with the agent. Statistical significance varied across sessions. Sessions 2 (p=0.000), 4 (p=0.014), 5 (p=0.033), 7 (p=0.023), and 9 (p=0.000) had significantly longer durations for the Todak_BC group, with the most notable differences in session 2 (16.5 seconds longer) and session 9 (40 seconds longer). No significant differences were found in sessions 1, 3, 6, and 8.

Number of utterances

The mean number of utterances in the Todak_BC group was significantly higher than in the Todak_NoBC group (p=0.007), with the Todak_BC group averaging 64.9 utterances compared to 43.1, indicating about 21 more utterances on average. The differences in utterances between the two groups across sessions showed similar significance patterns to conversation duration. Sessions 1 (p=0.003), 2 (p=0.000), 4 (p=0.027), 5 (p=0.020), 7 (p=0.030), and 9 (p=0.000) had significantly more utterances in the Todak_BC group. No significant differences were found in sessions 3, 6, and 8. The most notable differences were in sessions 2 (1.9 more utterances) and 9 (16.1 more utterances) for the Todak_BC group.

The experimental results for the two metrics of conversation persistence showed that BC expressed



(a) Conversation duration (s) (b) Number of utterances (c) Self-disclosure (d) Topic diversity

Figure 4: Mean comparison results for the four conversational engagement measures between Todak_BC and Todak_NoBC in CQ_LS, OQ_LS, and OQ_HS ($* < 0.05$, $** < 0.01$). Detailed statistical results are in Appendix D.

by the AI can contribute to conversation persistence similarly to human-human interactions. These results also revealed distinct characteristics between sessions with and without statistically significant differences. Sessions with significant differences, except for session 1, often involved closed-ended questions or high subject sensitivity. Sessions 2, 4, and 5, all closed-ended, and sessions 7 and 9, characterized by high subject sensitivity, showed significant differences for both metrics. Detailed analysis of these trends is presented in section 6.

5.2 How Todak's back-channeling influences conversation context richness

Self-disclosure

The mean self-disclosure score in the Todak_BC group was significantly higher than in the Todak_NoBC group across total sessions (p=0.007). The Todak_BC group averaged 54.07 compared to 48.32 for the Todak_NoBC group, indicating higher self-disclosure when BC was present. Statistical significance varied across sessions. In sessions 1 (p=0.003), 2 (p=0.002), 5 (p=0.015), 7 (p=0.031), and 9 (p=0.003), the Todak_BC group had significantly higher scores, while sessions 3, 4, 6, and 8 showed no significant differences. The most notable differences were in session 2 (6.50 vs. 5.56) and session 9 (8.03 vs. 7.00).

Results indicate sessions with significant differences often involved closed-ended questions or higher subject sensitivity. Closed-ended sessions (2, 4, 5) showed notable differences, especially sessions 2 (p=0.002) and 5 (p=0.015), while session 4 showed marginal significance (p=0.090). High subject sensitivity sessions (7 and 9) also showed significant differences, with the Todak_BC group disclosing more personal information.

The analysis of self-disclosure across the three dimensions shows the Todak_BC group consistently exhibited higher levels. For information, significant differences were found in sessions 1 (p=0.006), 2 (p=0.036), and 7 (p=0.017). For thoughts, differences were observed in sessions 2 (p=0.012), 5 (p=0.013), and 7 (p=0.013). For feelings, differences were noted in sessions 2 (p=0.003), 8 (p=0.038), and 9 (p=0.000). Overall mean scores were consistently higher for the Todak_BC group across all dimensions (see Table 4). Overall, BC expressed by the AI agent significantly enhances self-disclosure across information, thoughts, and feelings. Notably, in sensitive subjective sessions like session 9, the difference between the two groups is the largest (p=0.000), suggesting that BC creates a more engaging and comfortable environment for sharing deeper personal experiences.

Topic diversity

The mean topic diversity score in the Todak_BC group was significantly higher than in the Todak_NoBC group across total sessions (p=0.048). This indicates a broader range of topics discussed when BC is present, suggesting that back-channeling encourages participants to explore a wider variety of subjects. Examining topic diversity across individual sessions, statistical significance varied. Notably, only in session 9 (p=0.025), where participants discussed current concerns with high subject sensitivity, the Todak_BC group exhibited a significantly higher topic diversity score compared to the Todak_NoBC group (0.331 vs. 0.225). This implies that AI agents' BC can enhance topic diversity, especially in sensitive discussions. BC seems to foster an environment where participants feel more comfortable delving into a broader array of topics. Overall, considering both self-disclosure and topic diversity, the analysis highlights that, similar to human-human interactions, the AI agent's BC significantly enhances the richness and variety of topics discussed, especially in sensitive sessions. This supports the notion that BC not only helps maintain conversational engagement but also enriches conversational context. By providing timely and

Session	Group	Example
2	Todak_BC	*"I loved dinner time the most. The grilled steak and grape dessert were amazing."* (P-22)
	Todak_NoBC	*"The ramen was delicious."* (P-30)
9	Todak_BC	*"I've been struggling with time management lately and have been trying different approaches. I have a scheduling plan, but it's quite challenging because I have a lot to do. I'm not sure how to control it effectively, and I think I need to update my approach. I worry that if I miss even one small task, it could negatively affect the overall plan since everything is interconnected. I want to learn and improve my English, and I also have a strong desire to keep up with my exercise. However, managing all these activities together has made time management difficult. Although I am managing better than before, there are still areas that need improvement, and I think I could be more efficient. I'm not the type to plan things like walks well in advance; I tend to act more spontaneously. This isn't always the best approach, especially when deadlines are suddenly moved up. This has caused some difficulties."* (P-5)
	Todak_NoBC	*"I am currently studying in graduate school and have been struggling a lot with writing my thesis. However, I recently found some materials, and I am hopeful that further reviewing these materials will help with my thesis writing."* (P-11)

Table 5: Conversation logs examples (The logs were translated from Korean to English, and personal or sensitive information was edited.)

relevant acknowledgments, BC encourages speakers to continue their discourse, contributing to a more engaging and comprehensive conversational experience. This finding underscores the importance of implementing effective BC strategies in AI conversational agents to facilitate deeper and more meaningful interactions.

6 Conversational engagement analysis based on session characteristics

Based on hypothesis testing insights, we conducted further analysis to explore factors influencing BC effectiveness. We categorized the nine sessions into three groups: closed-ended questions with low or medium subject sensitivity (CQ_LS: Sessions 2, 4, 5), open-ended questions with low or medium subject sensitivity (OQ_LS: Sessions 1, 3, 6), and open-ended questions with high subject sensitivity (OQ_HS: Sessions 7, 8, 9). Comparing the four conversational engagement metrics between Todak_BC and Todak_NoBC within these categories revealed factors affecting BC effectiveness. Results are shown in Figure 4.

6.1 Question type

The type of question appears to influence BC effectiveness, with closed-ended questions showing greater differences in various metrics compared to open-ended questions. In CQ_LS sessions, Todak_BC showed significant differences across all metrics. In OQ_LS sessions, BC's influence was less pronounced, with significant differences observed only in the number of utterances, as shown in Figure 4b. Although the number of utterances measure showed statistically significant differences between the two groups in OQ_LS, the significance was lower compared to CQ_LS. An example from session 2, a CQ_LS session, is provided in Table 5. The session involved a closed-ended question where participants were asked to briefly answer about one of the menus mentioned in earlier sessions. The Todak_BC participants tended to provide detailed explanations following their initial answer, whereas the Todak_NoBC participants gave brief answers and moved on to next.

6.2 Subject sensitivity

Subject sensitivity also seems to influence BC effectiveness, with high-sensitivity topics showing greater differences across all metrics compared to low-sensitivity topics. In OQ_HS sessions, Todak_BC showed significant differences across all metrics, including topic diversity. While, OQ_LS sessions, BC's influence was less pronounced, with significant differences only in the number of utterances, as shown in Figure 4b. Todak_BC participants tended to share their concerns in detail, expressing multiple issues and emotions to ensure the listener's understanding. In contrast, Todak_NoBC participants quickly summarized a specific problem and ended the session. An example from session 9, an OQ_HS session, is provided in Table 5

7 Discussion

The findings from this study highlight that the presence of BC in AI-driven conversations significantly enhances conversational engagement, especially in contexts involving closed-ended questions or high subject sensitivity. The difference in responses between the two groups may result from

BC promoting intimacy and trust (Rost and Wilson, 2013; Gardner, 2001), making speakers more inclined to share their thoughts or emotions (McCabe et al., 2002). Participants in the Todak_BC group might feel more comfortable and engaged when discussing sensitive topics, with BC providing affirmation and encouraging deeper conversation.

These results are particularly relevant in contexts like counseling, where short, low-information responses are discouraged (Koch et al., 2004; Nor, 2020). This implies that incorporating BC in AI agents for counseling could significantly improve outcomes. Various proposals to enhance conversational engagement with AI agents include adding humanization features (Xu et al., 2022; Kang and Kang, 2024) or using self-disclosure strategies (Lee et al., 2020). However, this study uniquely demonstrates the effectiveness of BC as a strategy, making these results significant.

In conclusion, designing conversational strategies for AI agents should consider the dynamic use of BC, taking into account the nature of the conversation. Future research should explore the nuanced effects of BC in various settings and identify optimal BC strategies for different interactions.

8 Conclusion

We investigated the effectiveness of BC in improving conversational engagement between humans and AI agents. Our user experiment results demonstrated that BC can indeed influence human conversational engagement, contributing to both quantitative and qualitative improvements in conversations. This highlights the importance of considering BC when designing conversational strategies for future AI agent services.

Limitations

Despite robust findings, this study has some limitations. The overlap between session categories may introduce some ambiguity in the interpretation of results. Additionally, the sample size, while adequate, could be expanded in future studies to increase the generalizability of the findings. Further research should also explore the long-term effects of BC on user engagement and the potential for BC to improve outcomes in therapeutic settings. We also plan to conduct research on the application of BC based on the conversation patterns identified in the additional analysis section to facilitate its dynamic use.

Ethical Statement

This study was approved by the Institutional Review Board (IRB) of the CHA medical center (CHAMC 2021-08-033-003). To preserve the privacy of participants in the counseling data, we employed two methods: anonymization and data deletion. The personally identifiable information of the participants was fully anonymized, and we removed data that could potentially lead to the inference of personal information. The data were securely managed, with only the researchers involved in this study having access for both annotation and experimental purposes.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220608/2022-0-00608 and No. 2021-0-01343).

References

Antonio Andriella, Rubén Huertas-García, Santiago Forgas-Coll, Carme Torras, and Guillem Alenyà. 2020. Discovering sociable: using a conceptual model to evaluate the legibility and effectiveness of backchannel cues in an entertainment scenario. In *The 29th IEEE International Conference on Robot and Human Interactive Communication*, pages 752–759. IEEE.

Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, 10(3):407–417.

Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 759–766.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tianhua Chen and Mike Lucock. 2022. The mental health of university students during the covid-19 pandemic: An online survey in the uk. *PloS one*, 17(1):e0262562.

Eugene Cho, Nasim Motalebi, S Shyam Sundar, and Saeed Abdullah. 2022. Alexa as an active listener: how backchanneling can elicit self-disclosure and

promote user experience. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23.

Pino Cutrone. 2010. The backchannel norms of native english speakers: A target for japanese l2 english learners. *Language Studies Working Papers*, 2:28–37.

Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2010. Influence of personality traits on backchannel selection. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010*, pages 187–193. Springer.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

William Foddy and William H Foddy. 1993. *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge University Press.

Rod Gardner. 2001. *When listeners talk: Response tokens and listener stance*, volume 92. John Benjamins Publishing.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452.

Eunbin Kang and Youn Ah Kang. 2024. Counseling chatbot design: The effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. *International Journal of Human–Computer Interaction*, 40(11):2781–2795.

Sangmin Kim, Sukyung Seok, Jongsuk Choi, Yoonseob Lim, and Sonya S Kwak. 2021. Effects of conversational contexts and forms of non-lexical backchannel on user perception of robots. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3042–3047. IEEE.

Lynn C Koch, Connie McReynolds, and Phillip D Rumrill. 2004. Basic counseling skills. *Counseling theories and techniques for rehabilitation health professionals*, pages 227–243.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " i hear you, i feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

Han Z Li, Yanping Cui, and Zhizhang Wang. 2010. Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International journal of psychological studies*, 2(1):25.

Yeongbeom Lim, San Kim, Jin Yea Jang, Saim Shin, and Minyoung Jung. 2021. Ke-t5-based text emotion classification in korean conversations. In *Annual Conference on Human and Language Technology*, pages 496–497. Human and Language Technology.

Rosemarie McCabe, John Skelton, Christian Heath, Tom Burns, and Stefan Priebe. 2002. Engagement of patients with psychosis in the consultation: conversation analytic study commentary: Understanding conversation. *Bmj*, 325(7373):1148–1151.

Isabel Donya Meywirth and Jana Götze. 2022. Can you tell that i'm confused? an overhearer study for german backchannels by an embodied agent. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, pages 89–93.

Kelly A Morrow and Cecilia T Deidan. 1992. Bias in the counseling process: How to recognize and avoid it. *Journal of Counseling & Development*, 70(5):571–577.

Mohd Zarawi Mat Nor. 2020. Counselling: What and how. In *Counseling and Therapy*. IntechOpen.

Shaghayegh Nosrati, Maryam Sabzali, Ako Heidari, Tahere Sarfi, and Sina Sabbar. 2020. Chatbots, counselling, and discontents of the digital life. *Journal of Cyberspace Studies*, 4(2):153–172.

Gain Park, Jiyun Chung, and Seyoung Lee. 2023. Effect of ai chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model. *Current Psychology*, 42(32):28663–28673.

Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *Journal of medical Internet research*, 23(3):e24850.

Michael Rost and JJ Wilson. 2013. *Active listening*. Routledge.

H Sacks. 1978. A simplest systematics for the organization of turn taking for conversation.

Benjamin J Sadock and Virginia A Sadock. 2011. *Kaplan and Sadock's synopsis of psychiatry: Behavioral sciences/clinical psychiatry*. lippincott williams & wilkins.

Emmanuel A Schegloff. 1982. *Discourse as an interactional achievement: Some uses of" uh huh" and other things that come between sentences*. Analyzing discourse: Text and talk/Georgetown University Press.

Dwayne Simpson, Grace A Rowan-Szal, George W Joe, David Best, Ed Day, and Angela Campbell. 2009. Relating counselor attributes to client engagement in england. *Journal of Substance Abuse Treatment*, 36(3):313–320.

Jackson Tolins and Jean E Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.

Nam Khanh Tran, Sergej Zerr, Kerstin Bischoff, Claudia Niederée, and Ralf Krestel. 2013. Topic cropping: Leveraging latent topics for the analysis of small corpora. In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries*, pages 297–308. Springer.

Georgiana Shick Tryon. 1990. Session depth and smoothness in relation to the concept of engagement in counseling. *Journal of Counseling Psychology*, 37(3):248.

Henriette C Van Vugt, Johan F Hoorn, Elly A Konijn, and Athina de Bie Dimitriadou. 2006. Affective affordances: Improving interface character engagement through interaction. *International Journal of Human-Computer Studies*, 64(9):874–888.

Patricia McCarthy Veach, Dianne M Bartels, and Bonnie S LeRoy. 2007. Coming full circle: a reciprocal-engagement model of genetic counseling practice. *Journal of genetic counseling*, 16:713–728.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Harry Weger Jr, Gina Castle Bell, Elizabeth M Minei, and Melissa C Robinson. 2014. The relative effectiveness of active listening in initial interactions. *International Journal of Listening*, 28(1):13–31.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online. Association for Computational Linguistics.

Ying Xu, Jianyu Zhang, and Guangkuan Deng. 2022. Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Frontiers in Psychology*, 13:902782.

Richard F Young and Jina Lee. 2004. Identifying units in interaction: Reactive tokens in korean and english conversations. *Journal of Sociolinguistics*, 8(3):380–407.

A BC Expressions of Todak

Examples of verbal and non-verbal expressions for each BC category are illustrated in Figure 5. For "continuer," which shows that Todak is actively listening, the verbal expressions typically consist of short syllables or their repetitions, while the non-verbal expressions involve short and fast movements. This is represented by six dots that quickly shrink and grow sequentially or briefly expand and contract vertically, mirroring the short and fast verbal cues. The "understanding" category

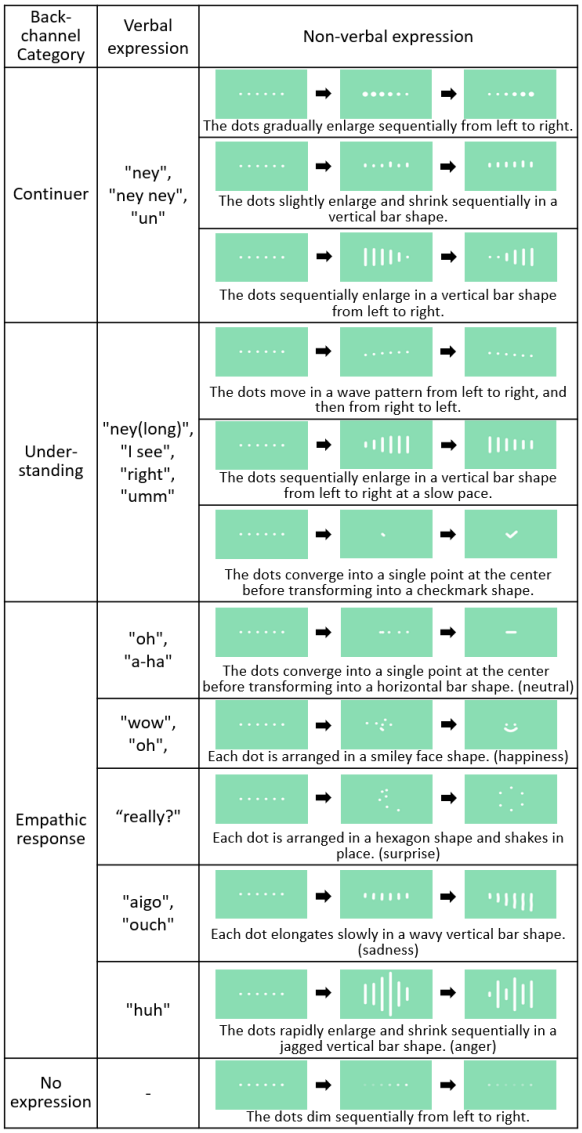


Figure 5: Verbal and non-verbal back-channel expressions by categories

is designed to convey comprehension of the spoken content, with both verbal and non-verbal expressions being relatively longer and slower compared to "continuer." The "empathic response" category is designed to reflect five emotional reactions: neu-

Prompt:
Measure the speaker's level of self-disclosure within a given utterance.
Measurements should be made for all of three categories: Information, Thoughts, and Feelings.
provides a single score for each of the three categories

Information level 1: utterances that provide only general or routine information without personal references, e.g. "I'm traveling to India for at least a month", "Yesterday Brazil beat Ecuador in a soccer match".
Information level 2: utterances that provide general information about the speaker, such as age, occupation, description of family members, interests, or hobbies.
Information level 3: The utterance reveals personal information about the speaker or the speaker's acquaintances, such as personal traits (characteristics), personal experiences, and problematic behaviors.

Thoughts level 1: No expression of any thoughts or opinions about a specific subject that could reveal the speaker's character/personality, only general thoughts, e.g., "It's dangerous to feed dogs human food," "Using cannabis for medical purposes can lead to dependence," etc.
Thoughts level 2: utterances that express the speaker's personal thoughts about past events or future plans, e.g., "I want to study biology when I go to college," "I remember the day my mom died," etc.
Thoughts level 3: utterances that express thoughts about the speaker's personal characteristics, appearance, health, sensitive topics, or hopes. Examples include, "I hate myself for hurting people and then apologizing for it. It's pathetic."

Feelings level 1: corresponds to an utterance with no expression of emotion. Monotone presentation of facts or personal ideas without any expression of emotion or sentiment.
Feelings level 2: Mildly expresses emotions about everyday feelings, confusion, discomfort, routine worries, complaints, or deficiencies. Examples include "I got a B in math and was frustrated. I envied my girlfriends who got higher grades." or "I'm tired of my boss. He makes me nervous."
Feelings level 3: Deep emotions such as joy, elation, humiliation, anguish, anxiety, depression, fear, or pain are expressed. For example, "I feel hopeless. I don't want to live anymore." "There's no one home, I'm so scared, oh my God. I'm shaking, I'm shaking with fear." etc.

Input utterance:
Yesterday I had a simple breakfast of coffee and a bagel, and for lunch I went to the cafeteria and chose pork cutlet from two options. I didn't eat dinner because I'm watching my weight.

Output response:
Information: Level 1 - The utterance includes routine information about daily activities and meal choices without any personal references beyond routine behavior.
Thoughts: Level 2 - The speaker mentions a personal decision regarding not eating dinner due to watching their weight, which reflects a thought about their future actions.
Feelings: Level 1 - There is no expression of emotion; the statement is a straightforward presentation of facts.

Table 6: Self-disclosure analyzer: ChatGPT prompts, input utterance examples, and corresponding responses.

tral, happiness, surprise, sadness, and anger. Verbal expressions are selected to match these emotions, while non-verbal expressions are portrayed through various animations, such as smiling faces for happiness, trembling for surprise, tear shapes for sadness, and intense vertical movements for anger. Additionally, the absence of BC expression is defined as "no expression." In this state, there are no verbal expressions, and the non-verbal expression involves the six dots on the screen gradually dimming from left to right, repeating this cycle as the idle state.

B Self-Disclosure Analyzer

Table 6 shows the self-disclosure analyzer prompts and an example of an input utterance and output response. The prompt was composed based on the guidelines of Barak and Gluck-Ofri (2007).

C Measuring Topic Diversity

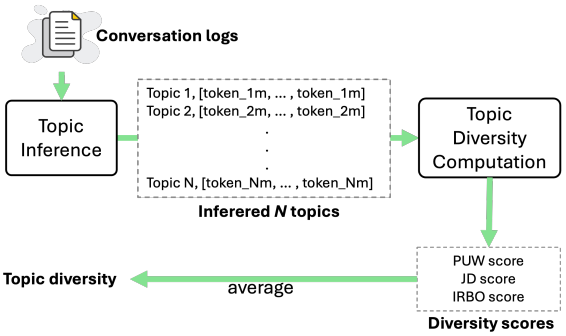


Figure 6: Procedure for measuring the topic diversity

The two-step procedure for measuring topic diversity can be found in Figure 6. The topic infer-

Session category	Conversation duration				Number of utterances			
	Mean (SD)		Statistic	p-value	Mean (SD)		Statistic	p-value
	Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC		
CQ_LS	44.27 (16.35)	32.44 (14.95)	U=218	0.004**	5.04 (2.40)	3.35 (2.04)	U=216	0.004**
OQ_LS	48.29 (22.67)	45.49 (19.00)	U=358	0.390	7.09 (5.19)	5.52 (3.62)	U=271	0.040*
OQ_HS	59.72 (33.99)	39.60 (29.05)	U=219	0.004**	8.69 (5.49)	5.49 (4.18)	U=227	0.006**

Total N = 55, *< 0.05, **< 0.01, ***< 0.001.

Table 7: Statistical test results for the conversation persistence in CQ_LS, OQ_LS, and OQ_HS

Session category	Self-disclosure				Topic diversity			
	Mean (SD)		Statistic	p-value	Mean (SD)		Statistic	p-value
	Todak_BC	Todak_NoBC			Todak_BC	Todak_NoBC		
CQ_LS	5.81 (0.90)	5.12 (0.96)	U=225	0.005**	0.344 (0.13)	0.274 (0.11)	t(53)=2.14	0.019*
OQ_LS	5.52 (0.82)	5.12 (0.88)	U=296	0.089	0.433 (0.11)	0.391 (0.15)	t(53)=1.19	0.120
OQ_HS	6.69 (1.01)	5.87 (1.10)	U=216	0.004**	0.476 (0.15)	0.350 (0.17)	t(53)=2.95	0.002**

Total N = 55, *< 0.05, **< 0.01, ***< 0.001.

Table 8: Statistical test results for the conversation context richness in CQ_LS, OQ_LS, and OQ_HS

ence module outputs N topics and a list of M tokens (set to 10) for each topic from conversation logs. The topic diversity computation module then calculates three diversity scores (PUW, JD, and IRBO) using the lists of tokens. The final topic diversity in our study is the average of the three diversity scores.

We referred to the LDA implementation for the topic inference module from this site: (https://wikidocs.net/40710). For the topic diversity computation (PUW, JD, and IRBO), we referred to the code from GitHub (https://github.com/silviatti/topic-model-diversity).

D Conversational Engagement Analysis by Question Type and Subject Sensitivity

Conversational engagement statistical analysis results by session characteristics, including question type and subject sensitivity, are presented in Table 7 and Table 8. When calculating topic diversity by session category, the number of topics (N) was set to 10.