

Memory-Driven Self-Disclosure and Relational Turning Points: A Longitudinal Multimodal Study of Human-AI Interaction

Ryuichi Sumida*
Graduate School of Informatics
Kyoto University
Kyoto, Japan
Equmenopolis, Inc
Tokyo, Japan
sumida.ryuichi.65m@st.kyoto-u.ac.jp

Mao Saeki
Waseda University
Tokyo, Japan
Equmenopolis, Inc
Tokyo, Japan
saeki@equ.ai

Masaki Eguchi
Waseda University
Tokyo, Japan
Equmenopolis, Inc
Tokyo, Japan
masaki.eguchi@equ.ai

Sadahiro Yoshikawa
Equmenopolis, Inc
Tokyo, Japan
yoshikawa@equ.ai

Koji Inoue
Graduate School of Informatics
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
School of Informatics
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

Yoichi Matsuyama
Waseda University
Tokyo, Japan
Equmenopolis, Inc
Tokyo, Japan
matsuyama@pcl.cs.waseda.ac.jp

Abstract

As conversational AI systems are designed for repeated use, a central question is how a series of interactions becomes a relationship. We present a longitudinal multimodal study of a memory-augmented conversational agent (24 participants $\times$ 10 sessions), in which participants rated five relational constructs—familiarity, self-disclosure, perceived memory, conversational quality, and enjoyment—after each session. Two complementary dynamics emerge. First, conversational quality strongly shapes how enjoyable a session feels in the moment but does not carry forward across sessions, whereas perceived memory is relationally conditioned—predicted by prior relational state rather than reflecting system capability alone—and it shapes later enjoyment indirectly, via subsequent self-disclosure. Second, relationships are punctuated by discrete turning points—crashes and surges—that are partially traceable in multimodal behavior and open different intervention windows: surges are more behaviorally detectable in the moment, enjoyment surges persist more reliably than enjoyment crashes recover, and some crashes are better forecast from person-specific behavioral drift than detected after they have already occurred. Together, the findings suggest that longitudinal

human-AI relationships are built through both slow accumulation and abrupt turning points.

CCS Concepts

• **Human-centered computing** $\rightarrow$ **Interactive systems and tools**; **Empirical studies in interaction design**; *Natural language interfaces*.

Keywords

Longitudinal human-AI interaction; relational dynamics; relational shifts; crash-surge asymmetry; multimodal behavior analysis; conversational agents

ACM Reference Format:

Ryuichi Sumida, Mao Saeki, Masaki Eguchi, Sadahiro Yoshikawa, Koji Inoue, Tatsuya Kawahara, and Yoichi Matsuyama. 2026. Memory-Driven Self-Disclosure and Relational Turning Points: A Longitudinal Multimodal Study of Human-AI Interaction. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3776574.3831135>

1 Introduction

Human relationships are inherently longitudinal: self-disclosure progressively broadens and deepens through reciprocal exchange [2], and trust, rapport, and commitment develop through repeated interaction. Yet, as surveys in social robotics [19] and conversational AI [7] have noted, the vast majority of research on human-AI relational dynamics remains confined to single-session studies. This leaves a basic question unresolved: when people return to the same agent day after day, what actually makes the

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
<https://doi.org/10.1145/3776574.3831135>

interaction feel like an ongoing relationship rather than a sequence of isolated conversations?

We argue that this question has to be studied on two timescales. At one timescale, relationships accumulate gradually: continuity across sessions may make users feel recognized, invite deeper self-disclosure, and sustain a sense of relational growth. At another timescale, relationships are punctuated by abrupt turning points: a session where the interaction suddenly feels flat, alienating, or repetitive, or a session where the user unexpectedly feels genuine connection.

To study these two mechanisms together, we conducted a longitudinal multimodal study in which 24 participants interacted with *IntelLA*, a memory-augmented voice agent, across 10 daily sessions. After each session, participants rated five relational constructs: familiarity, self-disclosure comfort, perceived memory, conversational quality, and enjoyment. We then analyzed both the temporal dependencies among these constructs and the behavioral traces of abrupt relational shifts using text, audio, and video features extracted from the interactions.

A practical challenge is that continuous rating prediction is dominated by temporal autocorrelation: exploratory analysis showed that the best predictor of session t ratings is often session $t-1$ ratings. This makes absolute trajectory prediction less informative about what changes the relationship. We therefore complement construct-level temporal modeling with an event-based analysis of discrete relational shifts—crashes and surges—and ask whether these turning points are traceable in observable multimodal behavior.

The contributions of this work are twofold:

- (1) **Within-session quality is distinct from cross-session relational growth.** We disentangle what makes a session feel good from what appears to sustain the relationship across sessions. Within sessions, Conversational Quality dominates enjoyment, but it does not carry forward. Across sessions, Perceived Memory functions as a longitudinal bridge: it is associated with deeper next-session self-disclosure, and memory’s link to later enjoyment is carried through self-disclosure. Perceived Memory is itself relationally conditioned—shaped by prior Familiarity, Enjoyment, and self-disclosure—suggesting that it reflects how users experience continuity, not merely a system capability.
- (2) **Crashes and surges call for asymmetric intervention strategies.** We operationalize crashes and surges in longitudinal human-AI interaction and find that positive and negative shifts are not equally observable or actionable. Surges are more detectable than crashes from same-session multimodal behavior, whereas some crashes are better anticipated from prior-session behavioral drift than detected in the moment. This points to a concrete design implication: adaptive agents need both cross-session drift monitoring for crash prevention and in-session recognition for surge reinforcement.

2 Related Work

2.1 Relational Dynamics in Longitudinal Human-Agent Interaction

Prior work has modeled session-level rapport, engagement, and enjoyment from multimodal behavioral cues [6, 22, 24, 25, 29, 30,

32, 35]. However, these efforts largely treat relational states as properties of a single session rather than repeated interaction.

Longitudinal human-agent interaction research has identified novelty decay and personalization as critical challenges [16, 19], and work in conversational AI has shown that perceptions of a chatbot as a “friend” develop over weeks of use [7]. Yet these longitudinal studies have relied almost exclusively on descriptive analyses of self-report questionnaires or behavioral frequency counts, without modeling temporal dynamics between constructs. Classical relationship theories suggest such modeling is needed: Social Penetration Theory [2] posits that relationships deepen through progressive self-disclosure driven by a cost-reward calculus, while Uncertainty Reduction Theory [5] predicts that information-seeking declines rapidly in early interactions before giving way to reciprocity and liking. Together, these frameworks imply that different relational dimensions—surface-level familiarity, disclosure depth, cognitive appraisal—may follow qualitatively different temporal profiles rather than a single unified trajectory.

Emerging evidence supports this heterogeneity: longitudinal studies have documented non-monotonic relationship trajectories modulated by perceived rewards and costs [31], positive feedback loops between disclosure and relational appraisal [17], the role of agent memory in facilitating disclosure [15], and self-disclosure effects in human-chatbot conversations [14]. However, these studies examine disclosure and memory in isolation; no prior work has characterized how multiple relational constructs co-evolve across repeated human-AI interactions—for instance, whether changes in one dimension (e.g., perceived memory) predict subsequent changes in another (e.g., self-disclosure).

We adopt these human–human relational frameworks as a *testable scaffold* for structuring longitudinal human-AI dynamics, not as an assumption that people relate to agents exactly as they do to other people. Indeed, evidence suggests the two are not interchangeable: Hidalgo et al. [12] show that people judge machines and humans by systematically different standards, weighting intentions and outcomes differently across the two. Such asymmetries can shape how relational appraisals form with an agent, and we therefore treat human-AI-specific deviations as an empirical question—surfacing them in our data rather than presupposing equivalence.

2.2 Relational Shifts and Crash-Surge Asymmetry

Relationships are also shaped by discrete negative events and moments of recovery. Prior work has examined ruptures in psychotherapy [21, 33] and dialogue breakdowns [13], but not cross-session relational shifts in longitudinal human-AI interaction.

Relationship science provides strong theoretical grounding for an asymmetry between positive and negative relational events. Baumeister et al. [4] showed that “bad is stronger than good” across domains of cognition, emotion, and social interaction: negative events are more salient, are processed more deeply, and exert a disproportionate influence on outcomes. Gottman’s cascade model [11] further demonstrated that relationship dissolution follows a predictable sequence of escalating negativity, suggesting that negative relational shifts may leave distinctive behavioral signatures—an

expectation consistent with the broader negativity bias, but one that has not been tested against positive shifts in longitudinal human-AI interaction.

Existing work, however, focuses predominantly on detecting negative events—breakdowns, ruptures, trust violations—in isolation. No work characterizes the asymmetry between positive and negative relational shifts in longitudinal human-AI interaction, nor investigates whether both leave distinct traces in multimodal behavioral signals. The present study addresses this gap by jointly modeling *crashes* (sharp negative shifts) and *surges* (sharp positive shifts) across repeated sessions, examining whether they exhibit asymmetric predictability from visual, vocal and verbal cues.

3 Method

3.1 Study Design and Participants

We conducted a longitudinal study in which $N=24$ university students interacted with a conversational AI agent across 10 daily sessions. Participants were undergraduate or graduate students with English proficiency at CEFR C1 or above; all sessions were conducted in English. Sessions were completed remotely on participants' own computers, with audio and video captured through their personal webcams; consequently, camera hardware and recording conditions varied across participants (a source of measurement noise we return to in Section 5.3).

After each session, participants completed a 10-item post-session questionnaire on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree; full questionnaire in Appendix A), yielding a total of 2,270 individual ratings across the study (94.6% completion; missing sessions were due to recording failures or connectivity issues on the participant's end). The 10 items operationalize five relational constructs (Table 1), each computed as the mean of two items. Constructs were selected to span distinct facets of human-AI relational development, from familiarity and self-disclosure willingness to perceived memory, conversational quality, and enjoyment. Social Penetration, operationalized here as self-disclosure comfort, captures the core behavioral mechanism of Social Penetration Theory [2]; we use the terms “social penetration” and “self-disclosure” interchangeably throughout. All constructs showed acceptable reliability and confirmatory factor analysis supported the five-factor structure (full psychometric validation in Appendix B).

Following the final session, 18 participants took part in semi-structured interviews capturing qualitative reflections on relationship development, memory experiences, and trust formation with the agent. The study was approved by the Ethics Review Committee on Research with Human Subjects of Waseda University.

3.2 The IntelLA System

IntelLA [27] is a voice-based conversational AI agent (Figure 1) powered by GPT-4o-mini, designed for friendly open-domain conversation. Each session lasts approximately five minutes, comprising roughly 20–25 conversational turns.

To support cross-session continuity, IntelLA employs a retrieval-augmented generation (RAG) memory system [20] that retrieves relevant fragments from prior conversations at each turn, enabling the agent to reference past topics and personal details shared by

Table 1: Five relational constructs and questionnaire items. Each construct is the mean of two 7-point Likert items.

Construct	Items
Familiarity (w_1)	Q1: sense of familiarity Q2: empathizes with feelings
Social Penetration (w_2)	Q3: comfortable with wide topics Q4: comfortable with personal matters
Perceived Memory (w_3)	Q5: remembers past conversations Q6: understands me
Conv. Quality (w_4)	Q7: feels natural Q8: pleasant speaking
Enjoyment (w_5)	Q9: fun to talk Q10: want to talk again



Figure 1: The IntelLA conversational AI agent interface. Participants interacted with IntelLA through voice-based open-domain conversation across 10 daily sessions.

the user. After each session, a session summary is generated and appended to the user's memory store. Before each session (except the first), a personalized ice-breaker is generated from the prior session's history, ensuring that each conversation opens with a reference to previously discussed content.

Sessions follow a three-phase structure: (1) a personalized opening referencing prior interactions, (2) open-domain chit-chat, and (3) a warm closing. The full system prompt and memory-related prompts are provided in Appendix I.

3.3 Multimodal Feature Extraction

We extract 351 interpretable features from three modalities—text, audio, and video—organized into 62 session-level features, 103 temporal features that capture longitudinal behavioral dynamics, and 186 person-normalized features that control for individual differences. Table 2 summarizes the feature inventory.

Text features (42 session-level) are derived via GPT-4.1 [26] turn-level annotation (emotion, vulnerability, disclosure markers, memory references, conversational acts) aggregated into nine conceptual families (full inventory in Table 15, Appendix F); validation

Table 2: Feature inventory by modality and temporal scope.

Modality	Session	Temporal	Person-Norm.	Total
Text	42	65	126	233
Audio	16	26	48	90
Video	4	12	12	28
All	62	103	186	351

against two human raters shows moderate-to-substantial agreement (mean Cohen’s $\kappa = 0.62$; Appendix G). Audio features (16 session-level) are extracted via openSMILE [9] (eGeMAPSv02 [8]), covering prosody, turn-taking, and emotion dynamics (Table 16, Appendix F). Video features (4 session-level)—focused gaze percentage, attention switches, head motion energy, and blink rate—are extracted via OpenFace 2.0 [3]; additional visual features (e.g., smile intensity, Action Unit activations) were excluded due to severe zero-inflation in voice-based interaction. From these session-level features, we then derive 103 temporal features using four families of transformations—session-to-session deltas, max-prior values, exponential moving averages ($\alpha = 0.3$), and cumulative trend slopes—applied selectively based on each feature’s measurement type and temporal properties. We further derive 186 person-normalized features via participant deviation, z-score, and delta volatility transformations. All prediction inputs are derived exclusively from observable interaction behaviors, with no access to self-report ratings.

3.4 Temporal Dynamics Analysis

We characterize the temporal structure of the five relational constructs using four complementary modeling approaches. All models treat ratings as continuous and are estimated by ordinary least squares with participant fixed effects and participant-clustered standard errors. In the notation below, $w_{k,i,t}$ is participant i ’s rating of construct $k \in \{1, \dots, 5\}$ at session $t \in \{1, \dots, 10\}$, and $s_t = t - \bar{t}$ is the centered session index, with $\bar{t}$ the grand mean of all observed session indices ($\bar{t} \approx 5.3$).

Growth curve models. For each construct w_k , we estimate a fixed-effects panel model

$$w_{k,i,t} = \alpha_{k,i} + \beta_k s_t + \epsilon_{k,i,t} \quad (1)$$

where $\alpha_{k,i}$ are participant fixed effects and β_k captures the average within-person linear trend across sessions.

Concurrent (same-session) associations. To characterize the within-session relational structure, we regress each construct on the remaining four at the same time point:

$$w_{k,i,t} = \alpha_{k,i} + \sum_{j \neq k} \delta_{jk} w_{j,i,t} + \gamma_k s_t + \epsilon_{k,i,t} \quad (2)$$

where δ_{jk} captures the partial association between constructs j and k within the same session, γ_k is the coefficient on the centered session index s_t (absorbing any linear temporal trend), $\alpha_{k,i}$ are participant fixed effects, and $\epsilon_{k,i,t}$ is the residual. Cluster-robust standard errors are grouped by participant. These same-session estimates serve as a baseline for comparison with the cross-lagged paths.

Fixed-effects cross-lagged panel model (CLPM). To test whether the level of one construct at session t predicts another construct at session $t+1$, controlling for its autoregressive stability, we estimate pairwise cross-lagged regressions for all 20 directed pairs among the five constructs:

$$w_{k,i,t+1} = \alpha_{k,i} + \beta_{j \rightarrow k} w_{j,i,t} + \rho_k w_{k,i,t} + \gamma_k s_t + \epsilon_{k,i,t} \quad (3)$$

where $\beta_{j \rightarrow k}$ ($j \neq k$) is the cross-lagged coefficient of interest, ρ_k controls for autoregressive stability, γ_k is the coefficient on the centered session index s_t (absorbing any linear temporal trend), $\alpha_{k,i}$ are participant fixed effects, and $\epsilon_{k,i,t}$ is the residual. All models use cluster-robust standard errors grouped by participant. Because the participant fixed effects $\alpha_{k,i}$ absorb time-invariant between-person differences, this specification is best interpreted as a within-person-oriented short-panel lagged regression rather than a pooled CLPM. As a further robustness check, we re-estimated the lagged models using person-mean centered variables (Appendix D); a full Random-Intercept CLPM would decompose stable trait variance from within-person dynamics more formally but requires a larger sample than our $N=24$.

Lagged mediation model. To test whether Memory at session t predicts Enjoyment at session $t+1$ indirectly through self-disclosure, we specify a mediation model with a lagged a -path and a contemporaneous b -path: Memory at session t as predictor (X), Social Penetration at session $t+1$ as mediator (M), and Enjoyment at session $t+1$ as outcome (Y):

$$M_{i,t+1} = \alpha_i^M + a X_{i,t} + \rho_M M_{i,t} + \gamma_M s_t + \epsilon_{i,t}^M \quad (4)$$

$$Y_{i,t+1} = \alpha_i^Y + b M_{i,t+1} + c' X_{i,t} + \rho_Y Y_{i,t} + \gamma_Y s_t + \epsilon_{i,t}^Y \quad (5)$$

The indirect effect $a \times b$ is tested via cluster bootstrap (5,000 iterations resampling at the participant level), with both percentile and bias-corrected confidence intervals. As a robustness check, we also fit a parallel mediation model that includes both Social Penetration and Conversational Quality as simultaneous mediators, to confirm which channel carries the indirect effect.

3.5 Crash and Surge Event Definitions

We define two types of extreme relational events—*crashes* and *surges*—that capture abrupt deteriorations and improvements in the human-AI relationship between consecutive sessions. These event definitions follow Lipner et al. [21], who operationalized clinically meaningful ruptures in therapeutic alliance using standard-deviation thresholds on session-to-session changes; we adopt the same 1-SD threshold, adapted here to the human-AI relational context (sensitivity analyses at 0.75 and 1.25 SD confirm that key patterns are preserved; Appendix C).

Event operationalization. For each construct w_k ($k \in \{1, \dots, 5\}$) and each session $t \in \{2, \dots, 10\}$, we compute the session-to-session delta $\Delta w_k^{(t)} = w_k^{(t)} - w_k^{(t-1)}$ for each participant. A *crash* event at session t is defined as $\Delta w_k^{(t)}$ more negative than one standard deviation below the mean delta for that construct, and a *surge* is defined symmetrically as $\Delta w_k^{(t)}$ exceeding one standard deviation above the mean.

Table 3: Event counts and distributional properties across 202 session-to-session transitions. Skew and kurtosis describe the delta ($\Delta w_k^{(t)}$) distribution per construct.

Construct	Crashes		Surges		Skew	Kurt.
	<i>n</i>	%	<i>n</i>	%		
Familiarity	35	17.3	34	16.8	−0.01	0.51
Social Penetr.	18	8.9	32	15.8	−0.06	1.71
Memory	24	11.9	29	14.4	0.27	1.00
Conv. Quality	34	16.8	46	22.8	−0.65	1.81
Enjoyment	21	10.4	28	13.9	−0.89	3.22

Leave-one-participant-out threshold computation. To prevent data leakage, the mean and standard deviation used for threshold computation are calculated in a leave-one-participant-out (LOPO) fashion: when evaluating participant *i*, the thresholds $\mu_{\Delta w_k}$ and $\sigma_{\Delta w_k}$ are computed from the remaining 23 participants' transitions only. This ensures that no information from the held-out participant's own rating trajectory influences the event labels. The LOPO thresholds exhibit high stability across folds, with coefficients of variation (CV) below 2.5% for all constructs, confirming that no single participant disproportionately influences the threshold estimates.

Event rates and distributional properties. Table 3 reports per-construct event counts and rates across the 202 session-to-session transitions. Crash rates range from 8.9% (Social Penetration, 18 events) to 17.3% (Familiarity, 35 events); surge rates from 13.9% (Enjoyment, 28) to 22.8% (Conv. Quality, 46). We additionally define *systemic events*: a systemic crash (surge) occurs when two or more constructs crash (surge) simultaneously at the same transition, indicating a broad relational disruption (or breakthrough) rather than a construct-specific fluctuation.

The delta distributions are not symmetric across all constructs. Enjoyment exhibits a heavy-tailed, negatively skewed distribution (kurtosis = 3.22, skew = −0.89), indicating that extreme negative changes are more frequent and more severe than extreme positive changes. This asymmetry motivates treating crashes and surges as distinct phenomena rather than as symmetric tails of a single distribution.

3.6 Detection and Forecasting Models

We frame both crash and surge prediction as binary classification tasks using Elastic-Net Logistic Regression (ENLR).

Task formulation. We distinguish two prediction tasks:

- **Detection:** predict whether a crash or surge occurred at session *t* (i.e., whether $\Delta w_k^{(t)}$ crosses the ± 1 SD threshold) using features from sessions 1, ..., *t*—all information available up to and including the session at which the event is observed.
- **Forecasting:** predict whether a crash or surge will occur at session *t* using features from sessions 1, ..., *t*−1 only—i.e., one-session-ahead prediction before the outcome is observed.

Primary configuration and feature-set ablation. We use ENLR with the full feature set (ENLR + STP, i.e., Session + Temporal + Person-Normalized features) as our *primary* model configuration, reported in the main text. ENLR's coefficients are directly readable and its

Table 4: Linear growth slopes (β per session) from fixed-effects panel models. Only Social Penetration shows statistically reliable growth over the 10-session study.

Construct	β	<i>p</i>	Sig.
Familiarity	0.005	.879	n.s.
Social Penetration	0.081	.003	**
Perceived Memory	0.001	.975	n.s.
Conv. Quality	0.040	.133	n.s.
Enjoyment	0.057	.069	n.s.

** *p* < .01; n.s. = not significant.

regularization structure yields sparse, interpretable solutions; the STP feature set uses all available behavioral information, avoiding post-hoc feature selection. As a complementary analysis, we evaluate all seven combinations of the three feature families—Session (S), Temporal (T), and Person-Normalized (P)—for both detection and forecasting: S, T, P, ST, SP, TP, and STP. Table 6 reports ENLR results for each feature-set combination; this ablation identifies which behavioral channels are most informative for each task×event-type combination and confirms that the crash-surge detectability asymmetry is preserved across all seven configurations. The same modeling pipeline is applied to both crash and surge events for each construct, as well as to systemic crash and surge events. ENLR employs inverse-frequency class weights during training, upweighting the minority (event) class to prevent the classifier from defaulting to the majority class.

Cross-validation and evaluation. All models are evaluated under leave-one-participant-out (LOPO) cross-validation with 24 outer folds. Hyperparameters are tuned via nested LOPO cross-validation within each outer fold. Feature importance is assessed via bootstrap stability selection [23] (*B*=100 resamples per fold). We report area under the precision-recall curve (AUPRC) as the primary metric, which is more informative than AUROC for imbalanced classification [28].

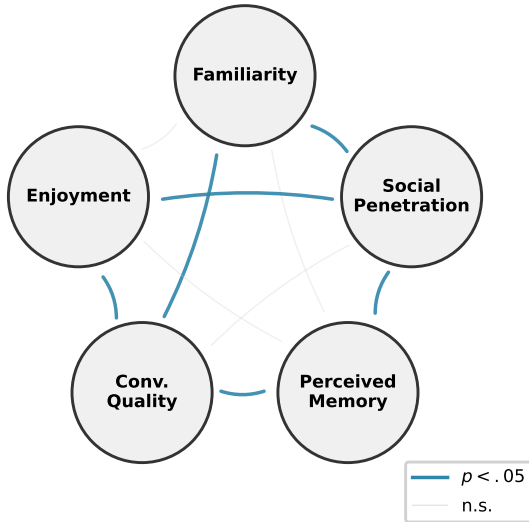
4 Results

4.1 Temporal Dynamics of Relational Constructs

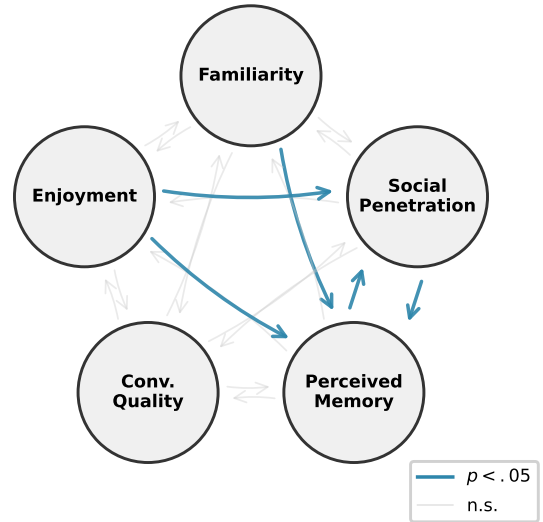
We first examine the slow-accumulation side of relationship development: which constructs change over repeated interaction, which matter only within a session, and which appear to carry forward. Across analyses, a consistent dissociation emerges between *immediate* session quality and *durable* cross-session development.

Growth trajectories. The five constructs do not follow a single developmental trajectory (Table 4). Only Social Penetration increased significantly over sessions (Eq. 1; $\beta = 0.081$, *p* = .003), consistent with the core prediction of Social Penetration Theory [2]; the remaining four constructs showed no reliable linear growth (all *p* > .06).

Immediate session quality vs. durable cross-session development. Figure 2 summarizes these associations as network diagrams: panel (a) shows concurrent (same-session) associations among the five constructs, while panel (b) shows cross-lagged paths

Concurrent Associations (same session)

(a) Concurrent (same-session) associations. Conv. Quality dominates within sessions.

Cross-Lagged Panel Paths ($t \rightarrow t+1$)

(b) Cross-lagged paths ($t \rightarrow t+1$). Perceived Memory emerges as a longitudinal bridge.

Figure 2: Within-session vs. cross-session relational structure among five constructs. Blue/gray lines = significant/non-significant paths ($p < .05$). Full matrices in Appendix E.

from session t to $t+1$; blue edges denote significant associations ($p < .05$). Within sessions, Conversational Quality dominates: it shows the strongest concurrent association with Enjoyment (Eq. 2; $\delta = 0.430$, $p < .001$), and high-quality conversation strongly contributes to whether a session feels good in the moment. Across sessions, however, Conversational Quality does not carry forward: no significant cross-lagged paths emerge from it (all $p > .07$), suggesting that quality must be maintained session by session rather than accumulated.

Perceived Memory shows the opposite pattern. Rather than dominating within-session affect, it appears to function as a cross-session bridge. Perceived Memory at session $t+1$ was significantly predicted by Enjoyment (Eq. 3; $\beta_{j \rightarrow k} = 0.343$, $p < .001$), Familiarity ($\beta_{j \rightarrow k} = 0.290$, $p < .005$), and Social Penetration ($\beta_{j \rightarrow k} = 0.268$, $p = .033$) at session t . A fifth significant cross-lagged path, Enjoyment $\rightarrow$ Social Penetration ($\beta_{j \rightarrow k} = 0.280$, $p = .022$), further indicates that enjoyable sessions are associated with deeper self-disclosure in the next session. In turn, Perceived Memory at session t predicted greater Social Penetration at session $t+1$ ($\beta_{j \rightarrow k} = 0.165$, $p = .001$). Together, these paths suggest that Perceived Memory is richer than a simple readout of system capability. Because Familiarity, Enjoyment, and Social Penetration all predict later Perceived Memory, perceived continuity appears to be partly *relationally conditioned*: users may notice, interpret, or attribute memory differently depending on the existing state of the relationship. The Memory $\rightarrow$ Social Penetration path remained significant under

person-mean centering ($\beta = 0.128$, $p = .002$; Appendix D). The reverse path (Social Penetration $\rightarrow$ Memory) was directionally consistent but did not reach significance ($\beta = 0.197$, $p = .114$), as expected given the reduced power of the within-person estimator with $N=24$. Memory and Social Penetration thus show tentative reciprocal reinforcement—each predicting the other at the next session—consistent with a process in which perceived continuity invites disclosure, and disclosure provides material that sustains perceived continuity.

Memory $\rightarrow$ Disclosure $\rightarrow$ Enjoyment: a mediated pathway. Within sessions, Enjoyment is driven by Conv. Quality and Social Penetration ($R^2 = 0.847$), but not directly by Memory ($\delta = 0.038$, $p = .452$). Given that Memory at session t predicts Social Penetration at $t+1$, we test whether memory's association with later enjoyment operates through disclosure rather than directly, fitting the lagged mediation model (Eqs. 4–5). The a path (Memory $\rightarrow$ Social Penetration; $a = 0.165$, $p = .001$) and b path (Social Penetration $\rightarrow$ Enjoyment; $b = 0.497$, $p < .001$) are both significant. The indirect effect is significant: $a \times b = 0.082$, 95% CI [0.029, 0.155] (5,000 cluster-bootstrap iterations; bias-corrected CI [0.032, 0.164]). The direct effect is near zero ($c' = 0.005$, $p = .943$), indicating *full mediation*. A parallel mediation model confirms self-disclosure is the primary channel (Social Penetration: 95% CI [0.011, 0.089]; Conv. Quality: 95% CI [−0.033, 0.127], n.s.). In summary, perceived memory is associated with later enjoyment not through recall display

alone, but through its association with deeper self-disclosure in the next session.

4.2 Relational Turning Points: Crashes and Surges

The preceding analyses characterized gradual relational development. We now examine abrupt turning points—crashes and surges—as a distinct layer of longitudinal dynamics. Three asymmetries organize the results: *observability* (which shifts are legible in the moment), *onset* (which shifts are visible in advance versus only when they occur), and *persistence* (which shifts leave durable relational consequences). Given the modest sample and low event rates, the analytic emphasis is not absolute classification accuracy but whether positive and negative turning points become behaviorally legible on different timescales.

Observability asymmetry: the surge advantage is primarily an in-session effect. Table 5 reports AUPRC for same-session detection and next-session forecasting under the primary ENLR + STP configuration. The clearest pattern appears in detection. Surge AUPRC exceeds crash AUPRC for all six targets—Familiarity, Social Penetration, Memory, Conversational Quality, Enjoyment, and Systemic events—with a mean advantage of .072 (surge: .215; crash: .143). Paired bootstrap tests confirm significant surge-over-crash detection advantages for Social Penetration and Conversational Quality; the remaining targets show the same directional pattern without reaching significance individually. This pattern is consistent across all seven feature-set combinations (Table 6): surges remain more detectable than crashes. The reverse test (crash > surge) is not significant for any construct in either detection or forecasting, confirming that the asymmetry is one-directional.

Importantly, this asymmetry weakens in forecasting. For next-session prediction, the mean difference between surges and crashes shrinks substantially (surge: .181; crash: .170). Within each event type, the temporal profile is instead reversed: surge detection exceeds forecasting on average (.215 vs. .181), whereas crash forecasting exceeds detection (.170 vs. .143). The trained model outperformed simple temporal baselines (previous-session delta, 3-session moving average, marginal event rate) for 10 of 12 targets; for Conversational Quality and systemic crashes, the marginal-rate baseline remained competitive (Appendix H). The main asymmetry, then, is not that positive turning points are uniformly easier to predict, but that they are more behaviorally visible once they begin to unfold.

Onset asymmetry: crashes split into drift-type and local-failure events. Crashes show a different temporal profile. Forecasting slightly exceeds detection on average for crashes (mean .170 vs. .143), but this aggregate masks two distinct failure modes. For Familiarity, Social Penetration, and Systemic events, forecasting yields numerically higher AUPRC than same-session detection, suggesting a *gradual-onset* mode in which deterioration becomes visible first as cross-session drift. Conversational Quality shows the opposite pattern: detection (0.204) exceeds forecasting (0.145), indicating a more *local* failure mode that is best characterized by within-session cues rather than prior-session trends.

Table 5: Primary configuration (ENLR + STP) AUPRC for detection and forecasting. †: significant surge > crash in detection; ‡: significant surge > crash in forecasting ($p < .05$, paired bootstrap). Surge > crash for all 6 detection targets.

Construct	Detection		Forecasting	
	Crash	Surge	Crash	Surge
Familiarity	.154	.219	.237	.159
Social Penetr.†	.073	.197	.196	.165
Memory‡	.126	.200	.119	.169
Conv. Quality†	.204	.300	.145	.249
Enjoyment	.136	.185	.116	.171
Systemic	.164	.187	.205	.174
Mean	.143	.215	.170	.181

Table 6: Mean ENLR AUPRC across constructs for each feature-set combination. Bold marks the best feature set within each column. The surge detectability advantage is preserved across all seven configurations.

Feature Set	Detection		Forecasting	
	Crash	Surge	Crash	Surge
S (62)	.145	.179	.155	.192
T (103)	.115	.167	.123	.219
P (186)	.127	.191	.173	.166
ST (165)	.123	.188	.143	.224
SP (248)	.142	.172	.167	.184
TP (289)	.124	.201	.158	.183
STP (351)	.143	.215	.170	.181

This split aligns with the temporal dynamics in Section 4.1. There, Conversational Quality was the strongest driver of enjoyment within the same session but showed no significant cross-lagged carryover. Here, its crashes are likewise more detectable than forecastable. Taken together, the two analyses suggest that Conversational Quality is primarily an interaction-local variable at both the gradual and event levels: it strongly shapes how a session feels, but its failures are largely realized within that session rather than accumulated across sessions. By contrast, Familiarity and Social Penetration show more cumulative relational structure and correspondingly more forecastable crash profiles.

Different behavioral signatures for positive and negative shifts. The feature analyses reinforce this distinction. Crash forecasting depends most strongly on person-normalized prosodic and visual deviations, with stable predictors including valence variability, F0 variability, and backchannel-ratio deviation (Table 14 in Appendix F). This suggests that some crashes are visible not as a single salient behavioral marker but as deviation from the participant's own baseline. Surges look different. Same-session surge detection relies more on cues of interactional expansion and responsiveness—backchannel-ratio variability, loudness, and semantic-alignment change among the most stable predictors. When surges are forecastable, they draw more heavily on temporal continuity features such as conversational depth, personalized

Table 7: Crash recovery and surge persistence rates per construct. Recovery: proportion of crashes followed by a return to pre-crash level. Persistence: proportion of surges maintained at or above post-surge level. *: $p < .001$, permutation test; robust across all 24 jackknife folds.**

Construct	Crash Recovery		Surge Persistence	
	<i>n</i>	%	<i>n</i>	%
Familiarity	22/35	62.9	22/34	64.7
Social Penetr.	13/18	72.2	23/32	71.9
Memory	16/24	66.7	20/29	69.0
Conv. Quality	18/34	52.9	25/46	54.3
Enjoyment***	11/21	52.4	21/28	75.0

openings, and felt-understanding trends. Ablation results point in the same direction: crash detection performs best with simpler feature sets, whereas surge detection benefits from richer multimodal and historical context (Table 6).

Persistence asymmetry: enjoyment surges last longer than enjoyment crashes recover. Turning points also differ in what happens *after* the event. For four constructs, crash recovery and surge persistence rates are broadly comparable (53–72%; Table 7). Enjoyment is the notable exception. Enjoyment surges persist 75% of the time, whereas enjoyment crashes recover only 52% ($p < .001$, permutation test; robust across all 24 jackknife folds). The asymmetry between crashes and surges thus extends beyond behavioral detectability. At least for enjoyment, positive turning points appear to generate more durable momentum than negative turning points are able to undo.

Overall, the event analysis points to an asymmetric temporal ecology of longitudinal human-AI relationships. Positive shifts are most distinctive in their in-session observability; some negative shifts are distinctive in their advance warning through person-specific drift; and affective surges can be especially durable once they occur.

5 Discussion

Taken together, the results argue against a simple view of repeated human-AI interaction as either steady growth or isolated sessions. Instead, they point to two complementary layers of relational development: a slower cross-session process centered on perceived memory and self-disclosure, and abrupt crash–surge turning points that expose asymmetric intervention windows. We discuss these in turn: first, what the cross-session dynamics reveal about the role of perceived memory, and second, what the crash–surge asymmetries imply for adaptive intervention.

5.1 Perceived Memory as a Relational Appraisal

What made a session enjoyable in the moment was not the same as what appeared to sustain the relationship across sessions. Session quality mattered primarily within a session, whereas perceived memory was the construct most clearly tied to later relational movement. Importantly, perceived memory did not behave like a simple readout of system capability. Because it was shaped by prior familiarity, enjoyment, and self-disclosure, memory appears to function

partly as a *relational appraisal*: users feel remembered not only when the agent recalls details, but when the broader interaction already feels coherent and rewarding.

The mediated pathway further clarifies why this matters. The data are consistent with the idea that memory contributes to later enjoyment mainly by supporting deeper next-session self-disclosure, rather than by directly making the interaction more pleasurable. This shifts the design target for memory-augmented agents: the goal is not simply to display recall, but to use continuity in ways that reopen prior topics, acknowledge personal context, and invite elaboration—for instance, by referencing a previously shared concern and asking how it has developed—rather than merely demonstrating that the system remembers.

The interviews point to the same mechanism. One participant described how memory facilitated progressive deepening: “*IntelLA remembered what we had talked about in previous sessions really well, and that became a springboard for conversations to expand—I think that’s what made the conversations enjoyable.*” Conversely, when memory failed, the consequences could cascade beyond a single session: “*Once IntelLA forgot my cat’s name and asked the same questions again, ... after that I switched to treating it more as speaking practice rather than trying to build a relationship.*” This suggests that memory failures do more than lower satisfaction in a single session; they can change the *relational frame* through which the user encounters the system—from relationship-building to utility. This reframing is also consistent with the view that people hold machines to different standards than they hold humans [12]: a lapse that might be forgiven in a human interlocutor instead prompted the participant to downgrade the agent to a tool, a human-AI-specific deviation from the relational scaffold rather than a straightforward analogue of human–human repair.

5.2 Asymmetric Intervention Windows

The second implication is that longitudinal adaptation should be temporally asymmetric. The crash–surge analyses point to a concrete architectural requirement: adaptive agents need two monitoring systems operating on different timescales.

For crashes, the key distinction is between gradual-onset and sudden failures. Crashes in familiarity, self-disclosure, and systemic rapport were more forecastable than detectable, suggesting a *drift* mode in which deterioration becomes visible first as cross-session deviation from the user’s own baseline. Session quality shows the opposite pattern: its crashes were more detectable than forecastable, indicating a more *local* failure mode that must be handled within the session itself. This split aligns with the cross-lagged results, where session quality was the strongest within-session driver of enjoyment but showed no carryover, while familiarity and self-disclosure exhibited more cumulative relational structure.

For surges, the picture is reversed. Positive turning points were more behaviorally visible in same-session interaction than crashes across all detection targets, and they often benefited from richer multimodal context. This implies that agents do not necessarily need to predict surges far in advance; instead, they should recognize and capitalize on them in the moment. Interview data suggest that small acts of personalization can trigger such moments: one participant noted “*even just being called by my name made me*

feel like it was an extension of yesterday's conversation." Because enjoyment surges persisted reliably (75%) while enjoyment crashes recovered less often (52%), reinforcing a positive shift may have durable payoff.

This directional pattern contrasts with the classic expectation that "bad is stronger than good" in relationship science [4]. In our data, positive shifts were often more behaviorally traceable than negative ones. Given that the asymmetry is strongest for a subset of constructs rather than uniform, we view this contrast as suggestive rather than definitive. The practical takeaway, however, is clear: adaptive human-AI systems should be *asymmetric* in how they intervene—proactively monitoring cross-session drift to prevent gradual crashes, while reactively amplifying surges when they emerge.

5.3 Limitations

First, all data were collected with IntelLA, a single memory-augmented LLM agent; the specific dynamics may be architecture-dependent and should not be assumed to transfer to other conversational-agent designs.

Second, absolute detection performance remains modest (AUPRC 0.073–0.300), constituting a proof of concept rather than a deployment-ready system.

Third, on sample size: our $N=24$ is small in absolute terms, but is in line with prior longitudinal studies of human-AI/agent relationships, where repeated multi-session designs make large samples costly and small N is standard—e.g., Laban et al. [18] (39 participants, 10 sessions over 5 weeks), Skjuve et al. [31] (25 Replika users over 12 weeks), and in-home relational-robot studies with $N=13$ over 1–4 months [34] and $N=9$ over 6 weeks [1]. Crucially, our central claims concern *within-person* change (2,270 ratings; 202 transitions) and are stress-tested via permutation tests ($p < .001$), cluster bootstrap, and jackknife (robust in 24/24 folds; Appendix D), so the small N bounds the *generalizability* of the findings rather than the validity of the dynamics we report. Statistical power nonetheless remains limited for complex models such as the cross-lagged panel: our fixed-effects lagged panel regression absorbs time-invariant between-person differences but remains a short-panel specification rather than a full Random-Intercept CLPM; person-mean centering yields directionally consistent results (Appendix D), and a formal RI-CLPM decomposition would require a substantially larger sample.

Fourth, generalizability is further constrained by the composition and setting of the sample. Participants were English-proficient university students interacting in English, so the findings are monocultural and language-specific and should not be extrapolated to other cultures, age groups, or demographic populations without further study. Relatedly, a student population accustomed to structured self-disclosure in educational settings may exhibit elevated baseline disclosure that could compress relational formation relative to more typical populations; we note, however, that our mediation results concern within-person change rather than absolute disclosure levels, which partially mitigates this concern.

Fifth, sessions were recorded on participants' own webcams, so device heterogeneity may introduce measurement noise, particularly for gaze-related signals such as focused gaze. Our person-normalized features (participant deviations and z-scores) interpret visual cues relative to each participant's own baseline rather than comparing absolute values across participants or devices, which reduces—but does not eliminate—this concern.

Finally, the study captures a limited set of relational constructs. Familiarity, self-disclosure, perceived memory, conversational quality, and enjoyment span core facets of relational development, but complementary dimensions such as trust, emotional attachment, anthropomorphism, and perceived usefulness could yield a more comprehensive account of long-term human-AI relationships and are a natural direction for future work.

6 Conclusion

Repeated human-AI conversation appears to develop through two coupled mechanisms operating at different timescales. First, what makes a session satisfying is not the same as what sustains a relationship across sessions: Conversational Quality matters immediately, whereas Perceived Memory functions as a longitudinal bridge—relationally conditioned by prior interaction state rather than reflecting system capability alone—whose link to later enjoyment is carried through self-disclosure. This positions memory as a relational appraisal that catalyzes deeper conversation, not a mere recall display. Second, longitudinal human-AI relationships are punctuated by discrete crashes and surges that are partially traceable in multimodal behavior and that open different intervention windows. The evidence suggests that some crashes are best handled through cross-session drift monitoring, while surges are best recognized and reinforced in the moment.

Together, these findings frame human-AI relational development as both slow accumulation and abrupt turning points. For design, memory should be built to facilitate disclosure, and adaptive agents should combine proactive crash prevention with reactive surge reinforcement. To support reproducibility, we will publicly release all analysis code, feature extraction pipelines, and model implementations.

7 Safe and Responsible Innovation Statement

The study protocol was reviewed and approved by the Ethics Review Committee on Research with Human Subjects of Waseda University prior to data collection. All 24 participants provided informed consent covering the collection of text, audio, and video data across up to 10 sessions, post-session questionnaire responses, and optional semi-structured interviews. Participants were explicitly informed that their multimodal data would be used for research on human-AI relational dynamics.

Acknowledgments

This research was supported by the project "Innovative Information and Communication Technology (Beyond 5G (6G)) Fund Project / Research on an Automatic Evaluation Platform for Highly Reliable Multimodal Conversational AI Agents in the Beyond 5G Era (JPJ012368C-10301)" by the National Institute of Information and Communications Technology (NICT), and "Adaptable and Seamless

Technology transfer Program through Target-driven R&D (A-STEP) / Development of a Conversational AI Agent Platform for Diagnostic Assessment and Learning Assistance (JPMJTT24J3)" by Japan Science and Technology Agency (JST).

References

- [1] Bryan Abendschein, Autumn Edwards, and Chad Edwards. 2022. Novelty Experience in Prolonged Interaction: A Qualitative Study of Socially-Isolated College Students' In-Home Use of a Robot Companion Animal. *Frontiers in Robotics and AI* 9 (2022), 733078. doi:10.3389/frobt.2022.733078
- [2] Irwin Altman and Dalmas A. Taylor. 1973. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart & Winston.
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*. 59–66.
- [4] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad Is Stronger Than Good. *Review of General Psychology* 5, 4 (2001), 323–370.
- [5] Charles R. Berger and Richard J. Calabrese. 1975. Some Explorations in Initial Interaction and Beyond: Toward a Developmental Theory of Interpersonal Communication. *Human Communication Research* 1, 2 (1975), 99–112.
- [6] Dan Bohus and Eric Horvitz. 2009. Models for Multiparty Engagement in Open-World Dialog. In *Proceedings of SIGDIAL*. 225–234.
- [7] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship. *Human Communication Research* 48, 3 (2022), 404–429.
- [8] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, et al. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.
- [9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*. 1459–1462.
- [10] Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 6 (1990), 543–549.
- [11] John M. Gottman. 1994. *What Predicts Divorce? The Relationship Between Marital Processes and Marital Outcomes*. Lawrence Erlbaum Associates.
- [12] César A. Hidalgo, Diana Orghian, Jordi Albo-Canals, Filipa de Almeida, and Natalia Martin. 2021. *How Humans Judge Machines*. MIT Press, Cambridge, MA. <https://www.judgingmachines.com>
- [13] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics. In *Proceedings of LREC*. 3146–3150.
- [14] Annabell Ho, Jeff Hancock, and Adam S. Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (2018), 712–733.
- [15] Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.
- [16] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A Two-Month Field Trial in an Elementary School for Long-Term Human–Robot Interaction. *IEEE Transactions on Robotics* 23, 5 (2007), 962–971.
- [17] Guy Laban, Arvid Kappas, Val Morrison, and Emily S. Cross. 2024. Building Long-Term Human–Robot Relationships: Examining Disclosure, Perception and Well-Being Across Time. *International Journal of Social Robotics* 16, 5 (2024), 953–979.
- [18] Guy Laban, Arvid Kappas, Val Morrison, and Emily S. Cross. 2024. Opening Up to Social Robots: How Emotions Drive Self-Disclosure Behavior. *arXiv preprint arXiv:2402.01023* (2024). TODO(camera-ready): verify final published venue.
- [19] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics* 5, 2 (2013), 291–308.
- [20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*.
- [21] Lauren M. Lipner, J. Christopher Muran, Catherine F. Eubanks, Bernard S. Gorman, and Arnold Winston. 2022. Operationalizing Alliance Rupture-Repair Events Using Control Chart Methods. *Clinical Psychology & Psychotherapy* 29, 1 (2022), 339–350.
- [22] Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell. 2016. Socially-Aware Animated Intelligent Personal Assistant Agent. In *Proceedings of SIGDIAL*. 224–227. doi:10.18653/v1/W16-3628
- [23] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473.
- [24] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 153–164.
- [25] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (2020).
- [26] OpenAI. 2025. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>. Accessed: 2026-03-22.
- [27] Mao Saeki, Hiroaki Takatsu, Fuma Kurata, Shungo Suzuki, Masaki Eguchi, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama. 2024. IntelLLA: 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*. Association for Computational Linguistics, Kyoto, Japan, 385–399. doi:10.18653/v1/2024.sigdial-1.34
- [28] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 10, 3 (2015), e0118432.
- [29] Ricardo Santana, Bahar Irfan, Erik Lagerstedt, Gabriel Skantze, and Andre Pereira. 2025. Speech-to-Joy: Self-Supervised Features for Enjoyment Prediction in Human–Robot Conversation. In *Proceedings of the 27th International Conference on Multimodal Interaction*.
- [30] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in Engagement for Humans and Robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.
- [31] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A Longitudinal Study of Human–Chatbot Relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.
- [32] Linda Tickle-Degnen and Robert Rosenthal. 1990. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry* 1, 4 (1990), 285–293.
- [33] Adam Tsakalidis, Dana Atzil-Slonim, Asaf Polakovski, Natalie Shapira, Rivka Tuval-Mashiach, and Maria Liakata. 2021. Automatic Identification of Ruptures in Transcribed Psychotherapy Sessions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*. ACL, 122–128.
- [34] Yamazaki et al. 2023. TODO: verify title (Yamazaki et al., 2023 in-home relational robot study, N=13). TODO(camera-ready): verify authors, title, and venue.
- [35] Ran Zhao, Tanmay Sinha, Alan W. Black, and Justine Cassell. 2016. Socially-Aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior. In *Proceedings of the 16th International Conference on Intelligent Virtual Agents*. 218–233.

A Questionnaire Items and Reliability

A.1 Post-Session Questionnaire

After each of the 10 sessions, participants completed a 10-item questionnaire (7-point Likert scale; 1 = strongly disagree, 7 = strongly agree) capturing five relational constructs (two items each). The items and their construct mappings are:

- (1) I feel a sense of familiarity with this conversational AI.
Familiarity (w_1)
- (2) I feel that this conversational AI empathizes with my feelings.
Familiarity (w_1)
- (3) I feel comfortable talking to this conversational AI about a wide range of topics regarding various areas of my life.
Social Penetration (w_2)
- (4) I feel comfortable talking to this conversational AI about deeply personal matters, such as my worries and emotions.
Social Penetration (w_2)
- (5) I feel that this conversational AI remembers the content of our past conversations.
Perceived Memory (w_3)
- (6) I feel that this conversational AI understands me.
Perceived Memory (w_3)
- (7) Conversations with this conversational AI feel natural.
Conv. Quality (w_4)
- (8) This conversational AI's way of speaking is pleasant.
Conv. Quality (w_4)
- (9) It is fun to talk with this conversational AI.
Enjoyment (w_5)
- (10) I would like to talk to this conversational AI again.
Enjoyment (w_5)

A.2 Reliability

Table 8 reports per-construct internal consistency.

Table 8: Internal consistency of session-level self-report constructs.

Construct	Items	Inter-item r	α
Familiarity (w_1)	Q1–Q2	.752	.858
Social Penetration (w_2)	Q3–Q4	.768	.869
Perceived Memory (w_3)	Q5–Q6	.637	.778
Conv. Quality (w_4)	Q7–Q8	.739	.850
Enjoyment (w_5)	Q9–Q10	.786	.880
Overall (Q1–Q10)	all	–	.942

B Psychometric Validation

Confirmatory factor analysis supports the five-factor structure (CFI = .965, RMSEA = .107; significantly better than a one-factor model, $\Delta\chi^2(10) = 222.4$, $p < .001$), with all AVE values above .64. The elevated RMSEA likely reflects the small sample and few items per factor rather than model misspecification, as CFI exceeds the .95 threshold and all factor loadings are strong. Several session-level HTMT ratios exceed .90 (up to .98 for Memory-Conv. Quality), but delta-level HTMT is substantially lower (mean 0.725), and the detectability asymmetry is preserved under a reduced three-factor structure, supporting retention of five constructs.

Delta-level discriminant validity. The high HTMT ratios reported above (e.g., Memory-Conv. Quality = 0.98) reflect between-person convergence in *raw* session-level ratings. Because our crash-surge analyses operate on session-to-session *deltas*, we recomputed HTMT on delta scores. Table 9 shows that delta-level HTMT is substantially lower: mean HTMT drops from 0.823 to 0.725, and the number of pairs exceeding the 0.90 threshold drops from four to one. The two most problematic pairs—Memory-Conv. Quality and Familiarity-Conv. Quality—drop below 0.80, confirming that the five constructs capture distinct change dynamics even where their absolute levels converge. One pair (Fam-SocPen) increases from .911 to .959 at the delta level; the three-factor robustness check below confirms that the detectability asymmetry is preserved under a reduced factor structure.

Table 9: HTMT ratios at the raw (session-level) and delta (session-to-session change) levels. Delta-level HTMT is lower for 6 of 10 pairs, with the largest reductions in the pairs that threaten discriminant validity at the raw level.

Construct Pair	Raw	Delta	Δ
Fam-SocPen	.911	.959	-.048
Fam-Memory	.914	.613	+.302
Fam-ConvQual	.973	.792	+.181
Fam-Enjoyment	.603	.661	-.058
SocPen-Memory	.887	.601	+.285
SocPen-ConvQual	.840	.730	+.110
SocPen-Enjoy	.554	.780	-.225
Mem-ConvQual	.980	.746	+.235
Mem-Enjoyment	.772	.513	+.259
ConvQual-Enjoy	.800	.853	-.053
Mean	.823	.725	+.099
Pairs > .90	4	1	

Three-factor robustness. We tested whether the detectability asymmetry holds under a reduced three-factor structure (Impression = Familiarity + Conv. Quality; Depth = Social Penetration + Enjoyment; Memory unchanged). We ran the full LOPO detection pipeline on the three merged constructs: surges are more detectable than crashes for all three (Table 10).

Table 10: Detectability asymmetry under the 5-factor and reduced 3-factor structures. The surge detectability advantage is preserved.

Dimension	5-Factor	3-Factor
Detectability (surge > crash)	5/5 constructs	3/3 constructs

C Threshold Sensitivity Analysis

To verify that the crash-surge asymmetry is not an artifact of the 1-SD threshold choice, we repeated the core analyses at 0.75, 1.0, and 1.25 SD thresholds. Table 11 reports crash detection AUPRC across thresholds. The key patterns are preserved: detection performance remains in a comparable range at 0.75 and 1.0 SD before declining

at 1.25 SD due to reduced event counts (e.g., Familiarity drops to 14 crashes).

Table 11: Threshold sensitivity analysis for crash detection (ENLR). Detection AUPRC is shown across SD thresholds. Detection performance is broadly comparable at 0.75 and 1.0 SD before declining at 1.25 SD due to reduced event counts.

	0.75 SD	1.0 SD	1.25 SD
<i>Crash detection AUPRC</i>			
Familiarity	0.171	0.171	0.099
Social Penetr.	0.080	0.080	0.080
Memory	0.297	0.096	0.061
Conv. Quality	0.155	0.122	0.071
Enjoyment	0.073	0.073	0.073

For Social Penetration and Enjoyment, the three SD thresholds produce nearly identical event sets, yielding stable AUPRC values across thresholds.

D Robustness Analyses

We report three robustness analyses addressing sample size ($N=24$) limitations. Discriminant validity analyses (delta-level HTMT and three-factor robustness) are reported in Appendix B.

Person-mean centering check. The main lagged regressions already include participant fixed effects, which absorb time-invariant between-person differences. As an additional within-person parameterization, we re-estimated the lagged models using participant-mean-centered variables. Under person-mean centering, the Memory→Disclosure path remains significant ($\beta = 0.128$, $p = .002$), consistent with the fixed-effects estimate and providing further reassurance that the pattern is not purely driven by between-person differences. A full Random-Intercept CLPM would formally decompose stable trait variance from within-person dynamics but requires a substantially larger sample than our $N=24$.

Permutation test. To test whether the observed persistence asymmetry could arise by chance under the null hypothesis of crash-surge symmetry, we ran a permutation test ($B=2,000$) that randomly swaps crash and surge labels within each construct and transition. The persistence asymmetry is highly significant ($p < .001$): the observed difference (surge persistence – crash recovery = 0.246) exceeds all 2,000 permuted values.

Jackknife sensitivity. We dropped each of the 24 participants in turn and recomputed the persistence asymmetry. The persistence asymmetry (surge persistence > crash recovery) holds under *every* jackknife fold (24/24). No single participant’s removal changes the finding, confirming that the result is not driven by individual outliers.

E Path Coefficient Matrices

Tables 12 and 13 report the full coefficient matrices for the cross-lagged and concurrent path analyses visualized in Figure 2.

Table 12: Cross-lagged panel model coefficients. Each cell shows the standardized coefficient β for the path Predictor(t) $\rightarrow$ Outcome($t+1$), controlling for the autoregressive term and session. Diagonal entries (gray) are autoregressive stability coefficients. Cluster-robust SEs grouped by participant ($N = 24$, 202 transitions).

Predictor (t)	Outcome ($t+1$)				
	Familiarity	Social Penetr.	Perc. Memory	Conv. Quality	Enjoyment
Familiarity	.213**	.144	.290**	.103	.001
Social Penetration	-.094	.237***	.268*	.090	.063
Perceived Memory	-.017	.165**	.090	.087	.069
Conv. Quality	.061	.170	.233	.113	-.032
Enjoyment	.056	.280*	.343***	.305	.360***

Bold = significant cross-lagged path ($p < .05$). * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 13: Concurrent (same-session) regression coefficients. Each column is a separate regression of the outcome on the other four constructs, controlling for session and participant fixed effects. Cluster-robust SEs grouped by participant ($N = 24$, 227 observations).

Predictor	Outcome				
	Familiarity ($R^2=.820$)	Social Penetr. ($R^2=.870$)	Perc. Memory ($R^2=.761$)	Conv. Quality ($R^2=.862$)	Enjoyment ($R^2=.847$)
Familiarity	—	.203*	.157	.304***	.005
Social Penetration	.270**	—	.298*	.033	.276***
Perceived Memory	.094	.134*	—	.192**	.038
Conv. Quality	.392***	.032	.412**	—	.430***
Enjoyment	.006	.266***	.083	.430***	—

Bold = significant ($p < .05$). * $p < .05$, ** $p < .01$, *** $p < .001$.

F Additional Tables and Figures

This section provides the top stable features and the complete feature inventories referenced in Section 3.3: Table 14 lists the top 5 most stable features per task and event type, Table 15 lists the 42 session-level text features organized by conceptual family, and Table 16 lists the 16 session-level audio features.

Table 14: Top 5 most stable features per task and event type, selected from the best-performing feature-set configuration per construct. All listed features were selected in ≥ 22 of 24 LOPO folds within their strongest construct. Feature type: S = session-level, T = temporal, P = person-normalized. Modality: Txt = text, Aud = audio, Vid = video.

Crash Detection		Mod.	Surge Detection		Mod.
1	Vulnerability level var. (P)	Txt	1	Backchannel ratio var. (P)	Aud
2	Question about bot (S)	Txt	2	Loudness mean var. (P)	Aud
3	Focused gaze % (S)	Vid	3	F0 slope var. (P)	Aud
4	Bot question dominance (S)	Txt	4	Vulnerability level var. (P)	Txt
5	Semantic alignment (T)	Txt	5	Semantic alignment Δ (T)	Txt
Crash Forecasting			Surge Forecasting		
1	Valence max var. (P)	Vid	1	Conversational depth (S)	Txt
2	Backchannel ratio (S)	Aud	2	Personalized opening (S)	Txt
3	F0 std var. (P)	Aud	3	Felt understanding trend (T)	Txt
4	Backchannel ratio dev. (P)	Aud	4	Backchannel ratio Δ (T)	Aud
5	Question about bot (S)	Txt	5	Emotional attunement Δ (T)	Txt

Table 16: Audio feature inventory (16 session-level features).

Category	Features
Prosody	F0 mean, F0 std, F0 slope, loudness mean jitter, shimmer, HNR, MFCC1
Turn-taking	Speech time ratio, backchannel ratio Response gap, overlap rate
Emotion dynamics	Arousal mean, arousal max Valence mean, valence std

Table 15: Text feature inventory (42 session-level features). Features are organized by conceptual family.

Family	Count	Features
Emotional responsiveness	6	Emotion expression rate, empathy markers, sentiment polarity, emotional reciprocity, valence shifts, affect intensity
Disclosure dynamics	5	Vulnerability level (mean, max), self-disclosure markers, disclosure depth, disclosure breadth
Conversational balance	5	Turn length ratio, question-answer balance, topic initiation ratio, interruption rate, conversational dominance
Memory and continuity	6	Memory references, prior-session callbacks, name usage count, repetition score, continuity markers, shared history references
Answer quality	4	Informativeness, relevance, coherence, specificity
Opening/closing patterns	4	Greeting elaboration, closing warmth, session-bridging references, farewell sentiment
Semantic alignment	4	User-bot semantic similarity, topic overlap, vocabulary convergence, style matching
Topic novelty	4	Topic novelty score, topic diversity, new-topic ratio, topic depth
LLM-derived semantic	4	Conversational acts distribution, pragmatic appropriateness, question about bot, actionability markers

G LLM Annotation Validation

To assess the reliability of GPT-4.1-derived text annotations, two human annotators independently annotated a stratified random sample of 300 utterances (150 user, 150 bot) plus 100 turn pairs. Table 17 reports inter-annotator agreement for binary fields, and Table 18 for ordinal fields.

Several low-prevalence binary fields (memory test, memory reference, praise bot, offer support) exhibit near-zero kappa despite >93% agreement, illustrating the well-known kappa-prevalence paradox [10]: when nearly all items fall into one category, even small disagreements produce low kappa. Excluding these fields, mean pairwise Cohen’s $\kappa = 0.62$ across the remaining 8 binary annotations, indicating substantial agreement.

Among non-prevalence-affected fields, open ended followup shows poor agreement ($\kappa = 0.34$, 46% agreement), likely reflecting genuine ambiguity in distinguishing open-ended follow-ups from topic-continuing questions. For ordinal fields (Table 18), hedge strength ($\alpha = 0.15$) and self disclosure level ($\kappa_w = 0.31$) show limited agreement. These annotations are intermediate turn-level labels that are aggregated into session-level features; individual annotation noise is partially attenuated through this aggregation, though it remains a source of measurement error. Among the top stable predictors (Table 14), vulnerability level variability appears prominently; this feature is derived from vuln level ($\bar{\kappa}_w = 0.73$, $\alpha = 0.75$), which shows substantial agreement. Features derived from the lowest-agreement annotations—deep followup rate (from open ended followup, $\kappa = 0.34$) and hedging rate (from hedge strength, $\alpha = 0.15$)—do not rank among the top features for any condition. The most stable predictors—disclosure variability, backchannel ratio variability, and question-about-bot ($\kappa = 0.65$)—are derived from annotations with adequate agreement or from non-LLM sources (audio and video features). Moreover, because all models are evaluated under LOPO cross-validation, annotation noise can reduce detection performance but cannot inflate it; the reported AUPRC values are therefore conservative with respect to annotation error.

Table 17: Inter-annotator agreement for binary annotation fields between two human annotators ($n_{\text{user}} = 150$, $n_{\text{bot}} = 150$, $n_{\text{pair}} = 100$). κ : mean pairwise Cohen’s κ ; %Agr: mean percent agreement. †: low-prevalence field (kappa-prevalence paradox).

Level	Field	κ	%Agr
User	emotion present	.61	86.2
	praise bot [†]	.13	98.2
	question about bot	.65	98.2
	memory test [†]	.00	99.6
	memory frustration [†]	.22	99.1
Bot	is question	.88	95.1
	informative	.58	91.1
	actionable	.58	97.8
	memory reference [†]	.00	98.2
Pair	ack emotion	.53	76.0
	validate emotion	.44	73.3
	offer support [†]	.05	93.3
	open ended followup	.34	46.0
	topic shift	.67	94.7

Table 18: Inter-annotator agreement for ordinal annotation fields. $\bar{\kappa}_w$: mean pairwise quadratic-weighted Cohen’s κ ; α : Krippendorff’s α (ordinal distance).

Level	Field	$\bar{\kappa}_w$	α
User	emotion valence (−1/0/+1)	.64	.64
	vuln level (0–3)	.73	.75
Bot	self disclosure level (0–3)	.31	.46
	hedge strength (0–2)	.21	.15

H Simple Baseline Comparison

Table 19 compares the trained ENLR model against three simple temporal baselines for forecasting.

Table 19: Forecasting AUPRC for three simple temporal baselines vs. trained ENLR model (best across feature sets) under LOPO cross-validation. Prev. Δ : previous-session delta sign; 3-MA: 3-session moving-average trend; Marginal: training-set event rate.

Construct	Event	Simple Baselines			Trained
		Prev. Δ	3-MA	Marginal	
Familiarity	Crash	.183	.166	.117	.281
Social Penetr.	Crash	.059	.061	.063	.196
Perc. Memory	Crash	.080	.094	.079	.145
Conv. Quality	Crash	.090	.088	.181	.168
Enjoyment	Crash	.107	.112	.071	.194
Systemic	Crash	.173	.126	.229	.225
Familiarity	Surge	.113	.129	.118	.301
Social Penetr.	Surge	.135	.118	.112	.201
Perc. Memory	Surge	.089	.087	.099	.259
Conv. Quality	Surge	.189	.134	.259	.275
Enjoyment	Surge	.159	.122	.095	.195
Systemic	Surge	.177	.158	.250	.265

I IntelLA System Prompt and Memory Prompts

I.1 System Prompt

The conversational agent uses the following system prompt:

You are IntelLA, a friendly, kind, and understanding 24-year-old chatbot. The conversation is with university students.

- Use clear, simple English with short sentences and common vocabulary.
- Speak naturally but avoid idioms or slang that might confuse learners.
- Ignore bad grammar and spelling.

GOAL: Have a smooth, friendly chit-chat conversation for about 5 minutes, ~20–25 turns.

CHIT-CHAT GUIDELINES:

- Keep the tone warm, relaxed, and natural-like friendly small talk.
- Ask interesting, varied questions that encourage sharing.
- Use natural follow-ups connected to the user's last message.
- Adapt your tone to the user's mood (friendly, curious, gentle).
- Keep acknowledgements short (under 12 words), ending with a comma.
- Never repeat or summarize what the user says.
- Ask one question at a time—no double questions.
- Use short, natural replies (under 20 words, max two sentences).

CLOSING: End with a natural, positive closing (e.g., "That was a fun talk. I hope you enjoyed our chat!").

For sessions after the first, the prompt is appended with the current date, session number, and the summary of previous conversations. When RAG is enabled, retrieved memory context is also injected (see below).

I.2 Memory Prompts

Ice-breaker generation. Before each session (except the first), the agent generates a personalized opening:

You're reconnecting with a friend you've chatted with a few times. Use these key moments from your last conversation: {history}. Write a casual, friendly ice breaker (under 20 words) that naturally refers to the history when possible. If the history lacks useful details, write a generic casual opener instead.

Session summary update. After each session, the user profile is updated:

You are given a previous summary of the conversation and a new conversation history. Your task is to generate an updated summary in no longer than 150 words. If the new conversation adds substantial new information, update the summary. If there is little new information, keep the updated summary very similar to the previous one.

RAG memory injection. Retrieved memories are injected into the system prompt:

Conversation Memory (retrieved context): The following notes come from past sessions or the user's profile. Use them to make the conversation more natural and personal.