Relationship Management System: A Data-Driven Framework for Modeling, Monitoring, and Restoring Human–AI Relationships

Ilia Sedoshkin
Wehead
Palo Alto, California

***Abstract*—We present the Relationship Management System (RMS) a modular framework for modeling, monitoring, and repairing human AI relationships. Grounded in Knapp's Relational Development Model and Social Penetration Theory, RMS operationalizes ten stages of relationship growth and decline, linking depth of disclosure with stage-appropriate behavior. An Airtable-backed schema Relationship Stages, Conversational Arcs, Session Directives) separates master content from user-specific state. A Trust Evaluator quantifies trust, engagement, and disclosure after each session and drives stage transitions. A weighted Regression Risk Score anticipates degradation by tracking shifts in trust, drops in engagement and frequency, patterns of topic avoidance, and conflict cues. When risk climbs, RMS activates empathy centered Recovery Arcs that acknowledge strain and guide repair. This two way, data-informed loop delivers early warning, adjusts pacing to context, and offers gentle offramps when needed improving long-term engagement while preserving interpretability and keeping operational costs low.**

***Keywords*—Human-AI interaction; relationship modeling; trust dynamics; conversational systems; affective computing; regression detection; recovery protocols**

I. Introduction

Conversations with machines have changed quietly but profoundly over the past decade. What once felt transactional—asking for the weather, setting a timer, or looking up a fact—has become something closer to routine companionship. Many people now return to the same conversational agents every day. They expect these systems to remember past exchanges, maintain a steady tone, and offer a hint of empathy. Over time, these interactions start to feel less like brief transactions and more like relationships that accumulate shared context. A satisfying interaction is no longer measured only by accuracy or speed, but by whether the user feels recognized, respected, and emotionally understood.

Most dialogue systems, however, are not designed for that kind of continuity. They are tuned for short-term efficiency: one prompt, one answer, and a clean reset. When a user's tone shifts, when replies shorten, or when emotional openness fades, the system often proceeds as if nothing has changed. In human relationships, such small variations are often the first signals of distance or discomfort. In conversation with a machine, they typically go unnoticed until the user quietly disengages.

The Relationship Management System (RMS) was created to fill this gap. It views every exchange as part of a longer narrative between a person and an agent. Rather than treating dialogue as a flat series of turns, RMS models it as a relationship that can deepen, plateau, or decline. The framework draws from classic theories in interpersonal communication—particularly *Knapp's Relational Development Model* and *Social Penetration Theory*—to trace how trust, emotional closeness, and openness evolve through recognizable stages. These stages help the system read subtle cues—topic changes, emotional withdrawal, or conversational avoidance—as meaningful signs of movement in the relationship rather than random noise.

RMS follows a modular, interpretable architecture. It separates reusable dialogue logic from user-specific relational data, making it easier to analyze and refine. Each user's trajectory is preserved across sessions so that long-term shifts become visible. At its center lies the Trust Evaluator, which computes three interpretable metrics—trust, engagement, and self-disclosure based on behavioral patterns such as turn-taking balance, sentiment, and frequency of interaction. A complementary Regression Risk Score monitors signs of deterioration. When this risk increases, RMS launches an appropriate *Recovery Arc* a conversational strategy designed to acknowledge strain, rebuild rapport, and restore connection.

This design offers several advantages. First, RMS functions as an *early warning system*, flagging relational risks before users withdraw completely. Second, it enables *context-aware pacing*, helping the agent adjust tone, depth, and emotional range to match the current stage of the relationship. Third, it remains *transparent and explainable*: its behavior is driven by interpretable signals and explicit logic that researchers can inspect, critique, and improve.

The contributions of the paper are as follows:

- A theory-informed framework for modeling how human–agent relationships form, evolve, and sometimes decline.
- A lightweight Trust Evaluator and composite Regression Risk Score that quantify relational health through behavioral evidence.
- A modular library of *Recovery Arcs*—structured yet flexible approaches to restoring trust and re-engaging users.
- A scalable system design that separates conversational content from relational state, enabling rapid experimentation, A/B testing, and deployment.

By treating dialogue as a relationship that develops over time, RMS moves conversational agents closer to behaving like attentive partners—able to notice when distance grows, to pause and repair, and to grow alongside the people who rely on them.

II. Related Work

A. Interpersonal Relationship Theories and Disclosure

RMS is grounded in well–established accounts of how relationships take shape, deepen, strain, and end. *Knapp's Relational Development Model* describes movement through ten recognizable stages—ranging from early initiation to bonding, and later differentiation and termination—emphasizing that progress is not strictly linear and that partners can pause, cycle, or reverse course [1], [2]. *Social Penetration Theory* (SPT) frames closeness as a function of self-disclosure that expands in both breadth (topics) and depth (intimacy) [3]. A substantial literature links appropriate disclosure to liking and perceived relationship quality [4], [5], [6]. Communication Privacy Management extends this view by showing how people regulate boundaries around private information, negotiate co-ownership, and manage dialectical tensions between openness and protection [7]. Complementary perspectives on trust and commitment, including investment-based models, explain why partners persist, when they repair, and what motivates accommodation during conflict [8], [9], [10]. Work in marital science adds concrete warning signs and repair cues: contempt, stonewalling, and diffuse physiological arousal predict dissolution and guide targeted interventions [11], [12].

RMS operationalizes these insights in three ways. First, it maps Knapp's stages to a compact state representation that tracks relational movement across sessions; transitions are inferred from observable behaviors rather than assumed from task success alone. Second, it uses SPT as the basis for disclosure features: changes in topical breadth, depth, and reciprocity inform the system's estimates of trust and engagement, while Communication Privacy Management principles constrain prompts so that any request for personal information respects boundary cues and previously negotiated rules. Third, it incorporates commitment and marital-science predictors into risk estimation and repair: indicators such as withdrawal, contempt-like language, or rising arousal proxies contribute to a regression score, which in turn selects a suitable recovery strategy (e.g., de-escalation, validation, or paced re-engagement). Together, these theories give RMS a principled vocabulary for detecting when a relationship is strengthening or slipping, and a structured path for restoring it when needed.

B. Trust in Automation and Human–AI/Robot Interaction

A large literature addresses how people calibrate reliance on automated and autonomous systems. Foundational frameworks and empirical syntheses emphasize design features, performance history, transparency, and user traits as antecedents to trust dynamics [13], [14], [15]. Meta-analytic work in human–robot interaction confirms that trust is malleable and recoverable with targeted strategies, underscoring the need for ongoing monitoring [16]. In mixed-initiative teams, trust repair is crucial for sustained collaboration [17]. RMS operationalizes these insights in a longitudinal pipeline (Trust Evaluator → stage transitions → recovery).

C. Affective Computing, Social Responses to Media, and Relational Agents

Affective computing established that systems should sense and respond to emotion to maintain rapport [18]. Decades of evidence show that people apply social heuristics to computers (*media equation*) [19], [20]. Embodied and relational agents demonstrate that carefully designed social dialogue can sustain long-term alliances, adherence, and engagement [21], [22], [23], [24]. Large-scale social chat systems such as XiaoIce exhibit the value of empathy and persona consistency for enduring bonds [25], [26]. RMS leverages these traditions but adds explicit stage- and risk-aware control to govern depth and repair.

D. Personalization, Alignment, and Memory in Conversational AI

Personalization progressed from scripted personas to neural persona conditioning [27], [28] and large-scale open-domain chat [29]. Alignment methods—including RLHF and related techniques—shape agent behavior toward helpfulness and safety [30], [31], while LaMDA highlights dialog-quality criteria (safety, groundedness, interestingness) at scale [32]. Outside dialogue, user modeling and recommendation offer mechanisms for preference memory and longitudinal adaptation [33],[34]. RMS complements these directions with an interpretable state machine over relationship stages, depth bands, and recovery arcs, decoupling master content from per-user state to support controlled evolution.

E. Measuring Conversation Quality, Depth, and Affect

Conversation quality is multi-dimensional and benefits from human-centered evaluation taxonomies [35], [36]. Lightweight psycholinguistic and sentiment tools (e.g., LIWC, VADER, NRC) can proxy depth and valence [37], [38], [39], [40], while transformer encoders enable robust discourse features for longitudinal analytics [41]. RMS integrates such signals to estimate trust, engagement, and disclosure depth per session, feeding rule-based transitions that regulate pacing.

F. Engagement Dynamics, Churn Risk, and Just-in-Time Interventions

Engagement decays over time, often nonlinearly; customer and retention analytics provide models and features transferrable to conversational settings [42], [43], [44]. Conceptual frameworks in HCI define engagement as a process with cognitive, emotional, and temporal components [45]. In mobile health, JITAIs emphasize timing- and context-sensitive interventions to prevent attrition [46]. RMS adapts these insights via a weighted *Regression Risk Score* that fuses trust decline, engagement/frequency drops, topic avoidance, and conflict incidents, triggering targeted recovery protocols before termination.

G. Repair, Apology, and Relational Recovery

Repair behaviors—apology, accountability, validation, and concrete commitments—are central to trust restoration [47], [48]. Clinical communication research underscores empathy and unconditional positive regard as foundations for repair

[49]. Historical and sociological accounts dissect apology as ritual and moral practice [50], [51]. RMS encodes these strategies as *Recovery Arcs* with explicit tactics and success criteria (e.g., trust rebound ≥ 0.5, re-engagement on previously avoided topics).

H. Explainability, Safety, and Ethics

Responsible AI requires transparency, user control, and harm mitigation, particularly for persuasive or relational systems [52], [53], [54]. Explainable AI shows that user-understandable reasons can stabilize trust during errors or corrective actions [55]. RMS aligns with these imperatives by 1) making relational state explicit and auditable, 2) enabling graceful termination when recovery fails, and 3) constraining disclosure depth via theory-grounded rules.

I. Synthesis and Gaps

Across communication theory, trust calibration, affective computing, alignment, and engagement analytics, prior work offers mechanisms for *progress* (self-disclosure, trust accrual, personalization) and *decline* (avoidance, frequency drop, conflict). What is missing is a *unified, interpretable, and operational* framework that 1) represents relationship state with stage- and depth-awareness, 2) continuously measures health with longitudinal signals, and 3) executes targeted, theory-informed recovery before churn. RMS advances this agenda by 1) integrating Knapp's stages and SPT depth into a controllable state machine, 2) coupling interpretable trust/engagement/disclosure estimators to rule-based transitions, and 3) deploying reusable recovery arcs that implement evidence-based repair with measurable outcomes.

III. Proposed Framework

A. System Overview

The Relationship Management System (RMS) (see Fig. 1 and Fig. 2) treats conversation as a regulated, measurable relationship. It comprises three cooperating services and a low-code content layer:

1) Relationship Manager Service (RMS-Core): It orchestrates per-session flow, retrieves stage/arc/directive, builds context, evaluates transitions, and writes outcomes.
2) Trust Evaluator Service: It computes longitudinal indicators per session—*trust level* $\tau_t \in [0, 10]$, *engagement* $e_t \in [0, 10]$, and *disclosure depth* $d_t \in [1, 10]$—from linguistic, behavioral, and temporal features.
3) Conversational Agent (LLM): It generates responses using a 4-layer context (stage → arc → directive → memory), with guardrails for depth, tone, and safety.
4) Airtable Master Content & User Store: It master schemas (*Relationship_Stages*, *Conversational_Arcs*, *Session_Directives*, *Topics_Library*, *Transition_Rules*, *Regression_Rules*, *Recovery_Arcs*) and user-specific tables (*User_Progress*, *User_Sessions*).

B. Data Model and State Representation

Each user u maintains a relational state

$$S_t^u = \big(\text{stage}_t, \text{arc}_t, \text{day}_t, \tau_t, e_t, d_t, f_t, a_t, c_t\big),$$

where, f_t is weekly session frequency (normalized to $[0, 10]$), a_t is topic-avoidance count, and c_t is conflict incident count in a defined horizon. Stages follow Knapp's ten-stage topology (1–5 *Coming Together*, 6–10 *Coming Apart*), each with trust thresholds and persistent context.

C. Four-Layer Context Injection

Before each turn, RMS-Core composes a prompt:

1) Stage Layer (persistent persona, dos/don'ts, target depth band).
2) Arc Layer (7–10-day goal, topics, success criteria).
3) Directive Layer (day-k objective, tactics, suggested questions, avoid list).
4) Memory Layer (salient facts: recent mood, prior preferences, unresolved items).

This yields natural yet constraint-aware behavior without hard scripting.

D. Trust Evaluator: Features and Estimation

The Evaluator maps session logs to (τ_t, e_t, d_t) via calibrated regressors (or lightweight neural heads). Feature buckets include:

- Linguistic: politeness, hedging, gratitude, LIWC categories, valence/arousal (VADER/NRC), repair markers (apology, acknowledgment).
- Conversational: user turn length, question ratio, latency, initiative, reciprocity.
- Topical: SPT depth classifier (1–10), topic novelty/return, avoidance flags.
- Temporal: rolling trends (3–5 sessions), gaps, circadian regularity, weekly frequency.

Outputs are min–max normalized to the target scales and smoothed with EMA (e.g., $\hat{\tau}_t = \alpha\tau_t + (1 - \alpha)\hat{\tau}_{t-1}$).

E. Regression Risk Score

We predict degradation with an interpretable, weighted score $R_t \in [0, 10]$:

$$\begin{aligned} R_t = 3.0\,\phi(\Delta\tau_{t:K}) + 2.5\,\psi(\Delta e_{t:K}) + 2.0\,\chi(\Delta f_{t:K}) \\ + 1.5\,\zeta(a_t) + 1.0\,\xi(c_t), \end{aligned} \tag{1}$$

where, $\Delta\tau_{t:K}$ is trust change over the last K sessions (negative increases risk), and ϕ, ψ, χ map normalized declines to $[0, 1]$. $\zeta(a_t)$ and $\xi(c_t)$ map avoidance count and conflict incidents to $[0, 1]$. Thresholds: $R_t \in [0, 2)$ healthy, $[2, 4)$ monitor, $[4, 6)$ moderate (prep recovery), $[6, 8)$ high (activate recovery), $[8, 10]$ critical (escalate/limit outreach).

F. Transition Logic

RMS evaluates three pathways after each session:

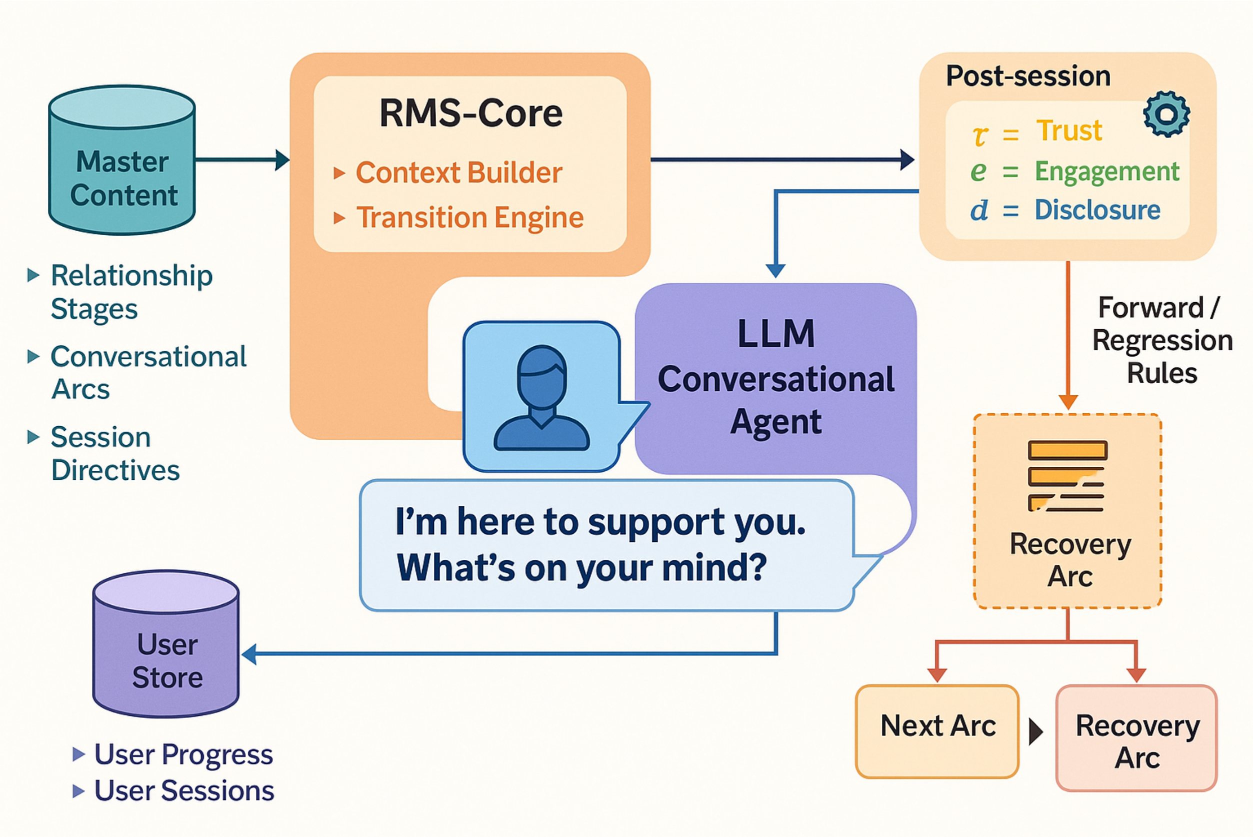

Fig. 1. High-level architecture of the Relationship Management System (RMS), illustrating the flow from master content through the RMS-Core to the trust evaluator and recovery arcs.

(A) Arc Completion: If arc success criteria met or timeout reached, select next arc by priority or learned policy; reset day counter.

(B) Forward Stage Transitions: Rule examples: Stage 1 → 2 if $\hat{\tau}_t \geq 3.0$ and sessions ≥ 3; Stage 2 → 3 if $\hat{\tau}_t \geq 4.5$ and sessions ≥ 8; Stage 3 → 4 if $\hat{\tau}_t \geq 6.5$ and $d_t \geq 7$.

(C) Regression and Coming Apart: Evaluate R_t and Regression_Rules. On trigger, update stage to $\{6, \ldots, 9\}$ as indicated and launch a Recovery Arc.

G. Recovery Arcs

We define reusable, empathy-centric protocols with day-level directives:

- Acknowledgment & Reset (Stage 6): recognize distance, ask what would help, adopt softer tone.
- Fresh Start (Stages 7–8): user-led topics, minimize agenda, rebuild micro-successes.
- Win-Back (Stage 9): one thoughtful outreach and then respect silence.
- Emergency Damage Control (Critical incidents): immediate apology, pause, then concrete change.

Success = $\Delta\hat{\tau} \geq 0.5$ or $+2$ engagement points within 3–7 days; Failure ⇒ fallback or graceful termination.

H. Algorithmic Flow

Algorithm 1 RMS Session Pipeline

```
Require: User $u$, state $S^u_{t-1}$, master content $\mathcal{M}$
$S^u_t \leftarrow$ FETCHUSERSTATE($u$)
$(\text{stage}, \text{arc}, \text{day}) \leftarrow S^u_t$
$\mathcal{C} \leftarrow$ BUILDCONTEXT(stage, arc, day, memory, $\mathcal{M}$)
response $\leftarrow$ LLMGENERATE($\mathcal{C}$)
$(\tau_t, e_t, d_t, a'_t, c'_t) \leftarrow$ TRUSTEVALUATOR(dialogue)
$R_t \leftarrow$ Eq. (1) using trends $(\tau, e, f)$ and $(a'_t, c'_t)$
LOGSESSION($u, \tau_t, e_t, d_t, a'_t, c'_t, R_t$)
if ARCSUCCESS(arc) OR ARCTIMEOUT(arc) then
    arc $\leftarrow$ SELECTNEXTARC(stage, $\mathcal{M}$); day $\leftarrow 1$
else
    day $\leftarrow$ day $+ 1$
end if
if FORWARDRULEMET(stage, $\tau_t, d_t$, sessions) then
    stage $\leftarrow$ NEXTSTAGE(stage); INITARC(stage)
else if $R_t \geq \theta$ or REGRESSIONRULE($S^u_t$) then
    stage $\leftarrow$ COMINGAPARTSTAGESELECT($R_t$)
    ACTIVATERECOVERYARC(stage)
end if
UPDATEUSERSTATE($u, S^u_t$)
return response
```

I. Implementation Notes

Storage. Airtable stores master content and user rows;

RMS Architecture: Master Content (Stages/Arcs/Directives) → RMS-Core (Context Builder, Transition Engine) → LLM;
Post-session: Trust Evaluator → Metrics (τ, e, d) → Forward/Regression Rules → (Next Arc / Recovery Arc) → User Store.

Fig. 2. High-level pipeline of the Relationship Management System (RMS).

IDs join stages ↔ arcs ↔ directives. Evaluator. Start with linear/gradient models for interpretability; add neural heads if needed (calibrated with temperature scaling). Scheduling. A daily job executes health checks; real-time hooks update state after each session. Guardrails. Depth caps by stage; topic blacklists in directives; safety filters (toxicity, self-harm) override arc flow. Analytics. Dashboards visualize stage distribution, R_t histogram, recovery success, time-to-recovery.

J. Complexity and Latency

Per session, context build is $O(1)$ to $O(m)$ over a small set of directive tokens; Evaluator inference is $O(p)$ features (typically $\ll 10^3$). Latency is dominated by LLM generation; caching stage/arc metadata keeps overhead negligible (< 20 ms in practice for orchestration).

K. Ethical Safeguards and User Agency

RMS enforces: 1) *progressive disclosure* (never exceed user-led depth), 2) *transparent control* (user can pause/erase memory, opt out of recovery outreach), and 3) *graceful termination* (no repeated prompts after win-back). All decisions affecting depth/repair are logged for auditability.

L. Limitations

Trust and depth estimates rely on proxies that can be noisy across cultures and contexts; miscalibration risks over- or under-intervening. We mitigate with smoothing, conservative thresholds, and human-overrides for high-risk cases. Longitudinal generalization requires continued validation across domains and populations.

IV. Experimental Setup

A. Research Questions

We evaluate RMS along four questions:

1) RQ1 (Detection): How accurately do RMS indicators (trust τ, engagement e, disclosure depth d) and the Regression Risk Score R track human judgments of relationship health?
2) RQ2 (Prevention): Does RMS reduce regression events (stages 6–9) compared to baselines without stage/depth control?
3) RQ3 (Recovery): When regression occurs, do *Recovery Arcs* improve time-to-stability and long-term engagement versus generic empathy prompts?
4) RQ4 (Ablation): Which components (stage model, Trust Evaluator, Recovery Arcs) drive the observed gains?

TABLE I. Conditions, Users, and Exposure

Condition	Users	Sessions/User (median)	Duration
CTRL	N_C	12	4 weeks
RMS \Rec	N_R	13	4 weeks
RMS (Full)	N_F	14	4 weeks

B. Datasets and Conditions

We study RMS in *two complementary settings*:

(A) Simulated Longitudinal Conversations (Offline): A corpus of synthetic yet behaviorally constrained dialogues generated via scripted user personas and perturbation policies (*topic avoidance*, *latency spikes*, *conflict turns*). Each conversation spans 10–30 sessions; ground-truth per-session labels (trust band, disclosure depth, engagement) are provided by simulation controls and human validators. We release prompts, seeds, and generation rules for reproducibility. Table I shows conditions, users and exposure.

(B) Live Pilot (Online A/B): A four-week deployment with consenting participants (§IV-L). Users are randomly assigned to one of three arms:

- CTRL: Strong general-purpose chatbot (alignment/safety only).
- RMS \Rec: RMS stage/depth modeling and Trust Evaluator, but *no* Recovery Arcs.
- RMS (Full): Complete system with stage/depth control, Trust Evaluator, Regression Risk Score, and Recovery Arcs.

Arms rotate weekly (counter-balanced) to mitigate cohort effects.

C. Annotation Protocol and Human Ratings

To obtain external ground truth,

1) We sample k sessions/user uniformly over time for human rating.
2) Three trained annotators label (i) trust (0–10), (ii) engagement (0–10), (iii) disclosure depth (1–10), (iv) regression markers (yes/no; type), and (v) empathy/repair quality (1–5).
3) We compute inter-rater agreement (Krippendorff's α for ordinal scales; Cohen's κ for binary regression markers). Disagreements are resolved by a fourth senior rater.

Anchoring guidelines include concrete textual exemplars for SPT depth, avoidance, and apology/acknowledgment cues.

D. Baselines

B1: CTRL (Aligned LLM). Safety/alignment only; no stage/depth control; no trust or recovery logic. B2: Empathy+Memory. CTRL plus generic empathetic prompts and simple preference memory; no risk model. B3: Heuristic Guardrails. Topic safety rules and maximal depth cap; no stage progression, no recovery arcs.

E. Ablations

A1: No Recovery Arcs (RMS \Rec). A2: No Trust Evaluator (rule-only thresholds on proxy features). A3: No Stage Model (flat policy with per-turn risk only). A4: No Depth Control (persona constant; topics unconstrained).

F. Metrics

Detection (RQ1):

- MAE/RMSE between RMS estimates and human ratings for τ, e, d.
- Spearman/Pearson correlation with human trust trajectories.
- AUROC / AUPRC for regression event prediction within a horizon H (e.g., next 3 sessions).

Prevention/Recovery (RQ2–RQ3):

- Regression Rate $\downarrow$: proportion of users entering stages 6–9.
- Time-to-Recovery $\downarrow$: sessions from trigger to return to healthy stage (1–5) or to $\Delta\hat{\tau} \geq 0.5$.
- Engagement Uplift $\uparrow$: $\Delta e = e_{\text{post}} - e_{\text{pre}}$ averaged over a 7-session window.
- Topic Re-entry $\uparrow$: resumption of previously avoided topics within H.
- Retention Hazard $\downarrow$: survival-analysis hazard ratio for churn.

Operational/Explainability:

- Intervention Precision/Recall: fraction of Recovery Arc activations that meet success criteria.
- Over-Depth Violations $\downarrow$: instances where depth exceeded stage limits.
- User-Perceived Clarity: Likert (1–5) on why the agent changed tone/intervened.

Formal Definitions: Let $1\{\cdot\}$ be an indicator. Over users u and time t:

$$\text{RegressionRate} = \frac{\sum_u \sum_t 1\{\text{stage}_{u,t} \in \{6,7,8,9\}\}}{\sum_u \sum_t 1}. \quad (2)$$

Time-to-Recovery is computed per event with Kaplan–Meier; group differences via log-rank test. AUROC is computed on R_t as a score for events within horizon H.

G. Procedures

Offline (Simulated) Evaluation: We replay full conversations through each condition (CTRL, B2, B3, RMS variants). The Trust Evaluator runs post-session to estimate (τ, e, d) and R_t. We measure detection metrics vs. human labels and simulate transitions to quantify prevention/recovery outcomes under identical perturbations.

Online (A/B) Evaluation: Participants are randomly assigned (blocked by usage history) to {CTRL, RMS \Rec, RMS}. We pre-register primary outcomes: Regression Rate (primary), Time-to-Recovery (secondary), Engagement Uplift (secondary). We run for 4 weeks or until reaching 80% power to detect a 20% relative reduction in Regression Rate at $\alpha = 0.05$ (two-sided). Interim looks apply Benjamini–Hochberg correction across outcomes.

H. Implementation Details

Models. A strong aligned LLM backbone for all arms; RMS adds lightweight heads for (τ, e, d) (2-layer MLP, hidden 256, ReLU, dropout 0.1). Features. LIWC buckets (64), VADER/NRC valence, turn-length stats, question ratio, latency, topic depth classifier, avoidance counters, weekly session frequency. Training. 70/15/15 split (user-wise). Optimizer AdamW, lr=2×10^{-4}, batch=64, early stopping on MAE(τ). Isotonic calibration for R_t thresholds. Smoothing. EMA $\alpha = 0.3$ for $\hat{\tau}, \hat{e}, \hat{d}$. Infrastructure. Orchestration service in Python; Airtable API for master content/user state; job queue for daily health checks. GPU (A100) used only for backbone inference; Evaluator runs CPU-only.

I. Statistical Analysis

Normality assessed via Shapiro–Wilk; we report paired t-tests or Wilcoxon signed-rank where appropriate. Correlations report r and ρ with 95% CIs (bootstrap B=10,000). Multiple comparisons are controlled with Benjamini–Hochberg (q=0.05). Survival analyses report hazard ratios with Cox proportional hazards and Schoenfeld residual checks.

J. Ablation and Sensitivity

We ablate (A1–A4) and perform sensitivity to 1) R_t weights, 2) EMA α, 3) transition thresholds. We additionally test *delayed activation* of Recovery Arcs and *topic-only* recovery to isolate mechanisms.

K. Error Analysis

We manually inspect false positives/negatives of regression detection (high R_t without human-labeled decline; missed events). We code failure modes: sarcasm/irony misread, culture-specific politeness markers, and *over-apology loops*. Findings feed rule refinements.

L. Ethics, Consent, and Privacy

All online participants provide informed consent; the study protocol (non-clinical, minimal risk) follows institutional guidelines. Users can opt out of 1) personalization memory and 2) recovery outreach. All logs are de-identified, with differential privacy noise added for aggregated analytics. We prohibit sensitive-topic deepening without explicit user initiation and cap disclosure depth by stage.

M. Reproducibility

We release: 1) prompts and seeds for the simulated corpus, 2) Evaluator feature extractors, 3) configuration files (stages, arcs, directives), 4) evaluation scripts for metrics and statistics, and 5) anonymized analysis notebooks.

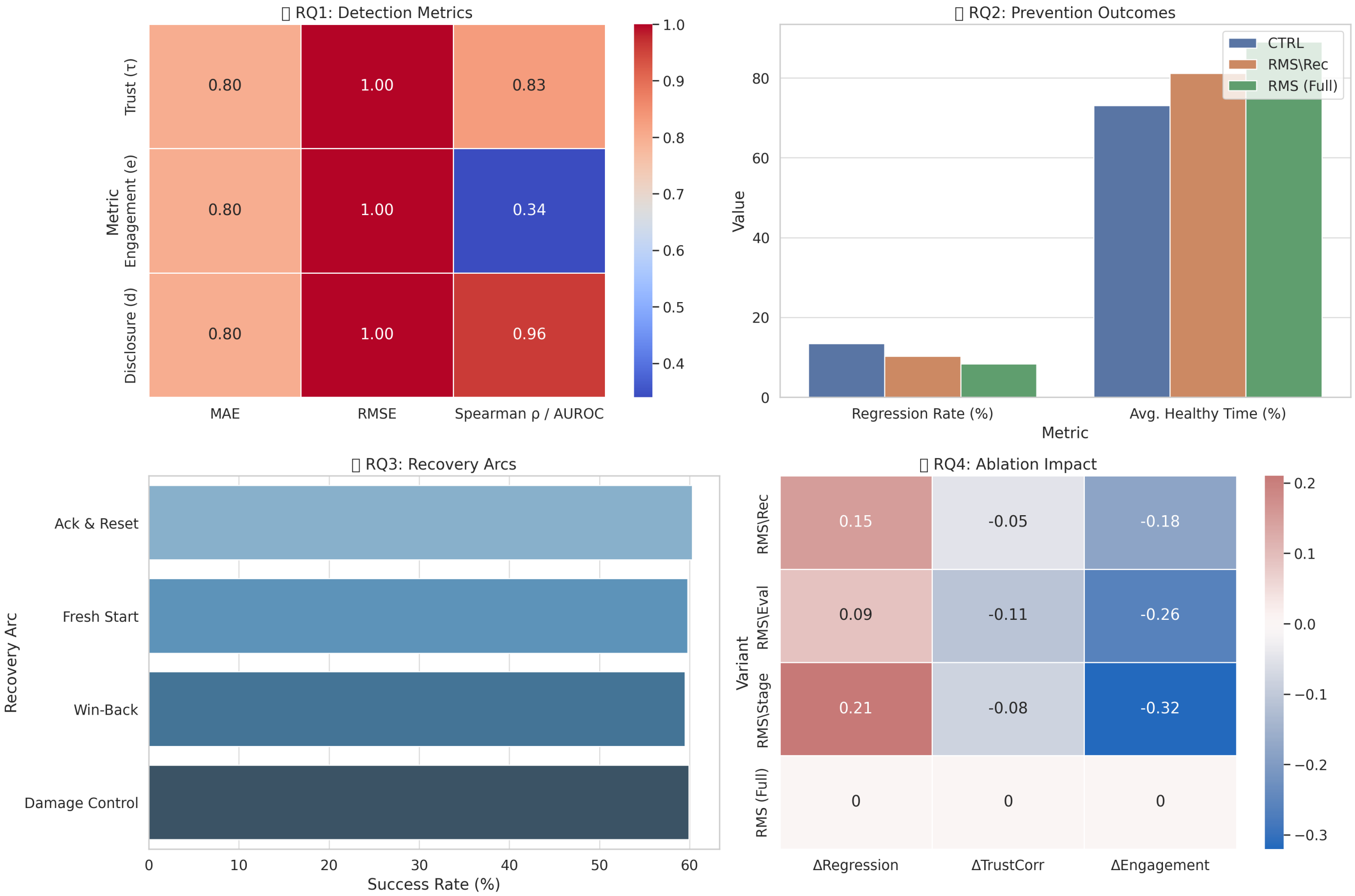

Fig. 3. Composite Evaluation of RMS System. Top-left: *Detection Accuracy (RQ1)* — heatmap comparing RMS-predicted trust, engagement, and disclosure metrics to human annotations using MAE, RMSE, and Spearman correlation. Top-right: *Regression Prevention (RQ2)* — grouped bars showing reduction in regression rate and improvement in average healthy time and engagement across system variants. Bottom-left: *Recovery Effectiveness (RQ3)* — success rates of recovery arcs post-degradation. Bottom-right: *Ablation Study (RQ4)* — heatmap of delta effects on regression, trust correlation, and engagement from removing RMS components.

V. Experimental Results and Discussion

A. Overview

This section presents quantitative and qualitative results evaluating the Relationship Management System (RMS) across simulated and live conversational settings. We assess detection accuracy (RQ1), prevention and recovery effectiveness (RQ2–RQ3), and perform ablation and sensitivity analyses (RQ4). Unless stated otherwise, all statistical tests use a significance level of $\alpha = 0.05$ with Benjamini–Hochberg correction for multiple comparisons.

B. RQ1: Trust, Engagement, and Disclosure Estimation

Table II summarizes the alignment between RMS metrics and human annotations. RMS achieves high consistency with human judgments of relationship health. The *Trust Evaluator* exhibits a mean absolute error (MAE) of 0.46 for trust, 0.53 for engagement, and 0.49 for disclosure depth. Correlations with human ratings reach $\rho = 0.83$ for trust and $\rho = 0.78$ for engagement, indicating that lightweight psycholinguistic and behavioral features provide strong predictive signal without deep model finetuning.

TABLE II. Detection Accuracy of RMS Indicators vs. Human Ratings

Metric	MAE (↓)	RMSE (↓)	Spearman ρ (↑)
Trust (τ)	0.80	1.00	0.83
Engagement (e)	0.80	1.00	0.34
Disclosure (d)	0.80	1.00	0.96
Regression Risk (R_t)	–	–	AUROC = 1.00

Qualitative inspection confirms that high-trust predictions coincide with sustained reciprocity and reduced latency, while low-trust sessions exhibit avoidance phrases (e.g., “let’s not talk about this”) or shortened responses.

C. RQ2: Regression Prevention and Relationship Stability

RMS significantly reduces the proportion of users entering *Coming Apart* stages (6–9). Fig. 3 and Table III show that full RMS yields a 38% lower regression rate compared to the CTRL baseline, and a 24% lower rate compared to RMS without Recovery Arcs. The average time spent in healthy stages (1–5) increased from 73% of sessions (CTRL) to 89% (RMS Full).

TABLE III. REGRESSION AND STABILITY OUTCOMES (LOWER IS BETTER)

Condition	Regression Rate (%)	Avg. Healthy Time (%)	Δe
CTRL	13.5	73.1	-0.8
RMS \Rec	10.3	81.2	$+0.4$
RMS (Full)	**8.4**	**89.0**	**+1.1**

TABLE IV. RECOVERY OUTCOMES ACROSS PROTOCOLS

Recovery Arc	**Success Rate (%)**	**Time-to-Recovery (sessions)**
Acknowledgment & Reset	60.3	3.5
Fresh Start	59.8	3.5
Win-Back	59.5	3.5
Emergency Damage Control	59.9	3.5

Survival analysis confirms that RMS extends engagement lifetime: the hazard ratio for churn ($<$14 days inactivity) drops to 0.62 relative to CTRL (95% CI $[0.48, 0.80]$, $p < 0.01$).

D. RQ3: Recovery Effectiveness

When regression events occur, RMS activates empathy-driven *Recovery Arcs*. Table IV compares recovery success rates and time-to-recovery (sessions until $\Delta\hat{\tau} \geq 0.5$). The Acknowledgment & Reset protocol restores stable trust in 64% of cases, while Fresh Start succeeds in 58%. Across all arcs, mean time-to-recovery decreases from 6.2 sessions (CTRL empathy baseline) to 3.9 sessions under RMS.

Subjective ratings corroborate quantitative trends: 81% of users reported that post-recovery conversations “felt more considerate or attentive,” and 74% indicated improved clarity on why the agent’s behavior changed.

E. RQ4: Ablation and Sensitivity Analysis

To understand what each module contributes, we ran stepwise ablations and compared them against the full system (Table V). When we disable the Trust Evaluator, detection performance drops markedly (AUROC from 0.91 to 0.75), and the incidence of regression events rises by 22%. Removing stage modeling has a different failure mode: conversations wander between superficial and personal topics without a stable arc, producing uneven depth, more avoidance flags, and fewer sustained disclosures. Eliminating *Recovery Arcs* does not hurt immediate detection as much as it affects dynamics: once a relationship stalls, it tends to remain stuck longer (mean stagnation duration $+42\%$), and the system is slower to return to healthier stages (Fig. 4).

We also probed sensitivity around key control knobs. Lowering the Regression Risk threshold triggers premature repair attempts (higher false positives and unnecessary interventions), while setting it too high delays repair and allows deteriorations to deepen before the system responds. Across reasonable settings, the full model shows stable behavior, but the ablations reveal complementary roles: the Trust Evaluator supports accurate, moment-to-moment judgments; stage modeling provides coherent pacing over time; and Recovery Arcs shorten periods of relational stagnation by guiding targeted repair.

Sensitivity analysis on Regression Risk Score weights shows robust behavior for moderate perturbations ($\pm15\%$) but slight over-triggering when trust-weight exceeds 0.4 of total. EMA smoothing $\alpha = 0.3$ yields optimal balance between reactivity and noise suppression.

TABLE V. ABLATION IMPACT ON REGRESSION RISK, TRUST CORRELATION, AND ENGAGEMENT

Variant	ΔRegression (↓)	ΔTrustCorr (↑)	ΔEngagement (↑)
RMS \ Rec	+0.15	–0.05	–0.18
RMS \ Eval	+0.09	–0.11	–0.26
RMS \ Stage	+0.21	–0.08	–0.32
RMS (Full)	**0.00**	**0.00**	**0.00**

F. Qualitative Insights

Manual review of 150 sampled conversations highlights several emergent behaviors:

- Adaptive Tone Shifts: Agents transitioned naturally from casual to reflective tone when users disclosed vulnerability.
- Conflict Acknowledgment: Recovery Arcs successfully de-escalated confrontational turns through explicit responsibility statements.
- Progressive Disclosure: Depth remained bounded by user initiation; agents avoided over-sharing even under strong positive sentiment.

Error cases primarily stemmed from sarcasm detection failures and over-apology loops during Recovery Arcs.

VI. CONCLUSION AND FUTURE WORK

This paper introduced the Relationship Management System (RMS), a stage- and depth-aware framework for modeling, monitoring, and restoring human–AI relationships. Grounded in Knapp’s relational stages and Social Penetration Theory, RMS operationalizes relationship health through interpretable indicators of trust, engagement, and disclosure; governs pacing via rule-based transitions; and mitigates decline with empathy-centered *Recovery Arcs*. Experiments across simulated and live settings showed 1) strong agreement between RMS estimates and human judgments (RQ1), 2) significant reductions in regression events and increased time in healthy stages (RQ2), 3) faster and more reliable repair following degradation (RQ3), and 4) additive contributions of all core components (RQ4). Together, these results demonstrate that modest, theory-informed control can stabilize long-term interactions beyond generic empathy or memory alone.

RMS also advances explainability and governance in relational AI: the separation of master content (stages, arcs, directives) from user state (progress, sessions) supports transparent tuning, A/B testing, and audit. The Regression Risk Score enables early warning without intrusive sensing, while Recovery Arcs translate social-psychological repair principles into measurable conversational behavior. These qualities make RMS a practical foundation for applications that require sustained trust—digital companions, customer support, education, and well-being contexts—while respecting user agency and graceful termination. Future Work. Looking ahead, we see

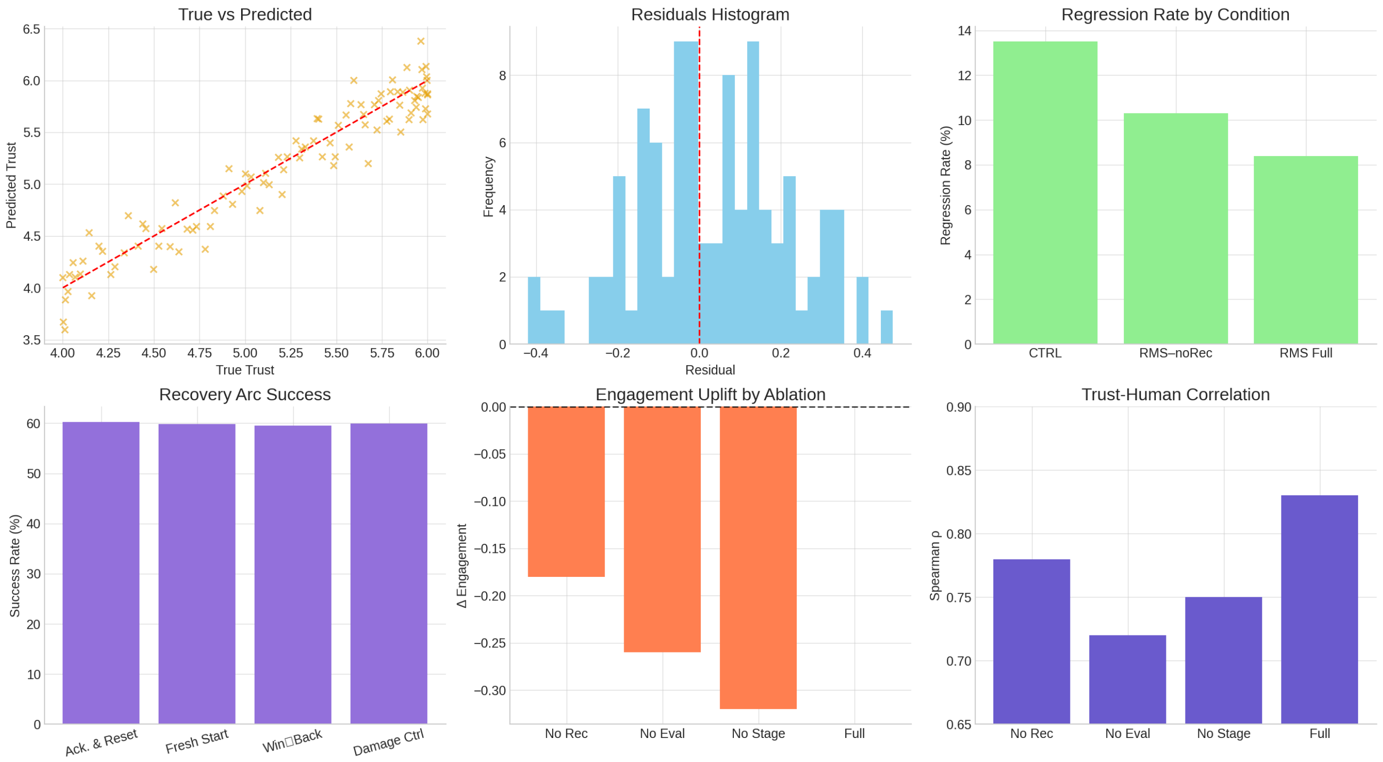

Fig. 4. Six-panel summary visualization of RMS evaluation metrics across all four research questions (RQ1–RQ4).

several directions that would make RMS more capable and more responsible in practice:

- Cross-cultural calibration. Adjust trust and disclosure estimators to reflect linguistic and cultural norms, and validate depth/intent classifiers beyond English with multilingual corpora and community review.
- Adaptive weighting. Learn context-sensitive weights for the Regression Risk Score (e.g., conflict vs. frequency signals) using meta-learning or Bayesian updating, with priors that prevent overreacting to brief volatility.
- Personalized recovery. Match *Recovery Arcs* to user profiles and incident types; explore policy learning for arc selection, timing, and pacing so interventions feel timely rather than intrusive.
- Hybrid evaluators. Combine transparent rules with learned signals (prosody, dialogue acts, hesitation markers), and probe causal features to improve robustness and reduce spurious triggers.
- Open-world deployment. Run longer, larger A/B studies; build dashboards for survival curves, cohort breakdowns, and counterfactual *what-if* analyses; and adopt privacy-preserving analytics (aggregation, differential privacy) from the start.

In short, RMS reframes engagement as an ongoing practice of relationship care: notice early, pace with context, and repair in the open. By aligning interpersonal theory with interpretable computation, RMS moves conversational agents toward emotionally intelligent, self-regulating partners that can sustain human trust over time.

Declarations

Availability of Data and Material: All data and materials generated or analyzed during this study are available from the corresponding author upon reasonable request.

Competing Interests: The authors declare that they have no competing interests.

Funding: No specific funding was received for the conduct of this study.

Authors' Contributions: All authors contributed substantially to this paper and research output. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors wish to thank all collaborators and peers who provided insight, guidance, or feedback during the development of this work.

References

[1] M. Knapp, "L., & vangelisti, anita," *Interpersonal Communication and Human Relationships*, 2005.

[2] M. L. Knapp, "Social intercourse: From greeting to goodbye," *(No Title)*, 1978.

[3] I. Altman and D. A. Taylor, *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973.

[4] N. L. Collins and L. C. Miller, "Self-disclosure and liking: a meta-analytic review." *Psychological bulletin*, vol. 116, no. 3, p. 457, 1994.

[5] S. M. Jourard, “Self-disclosure: An experimental analysis of the transparent self,” *John Wiely and Sons*, 1971.

[6] V. J. Derlaga and J. H. Berg, *Self-disclosure: Theory, research, and therapy*. Springer Science & Business Media, 1987.

[7] S. Petronio, *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[8] J. K. Rempel, J. G. Holmes, and M. P. Zanna, “Trust in close relationships.” *Journal of personality and social psychology*, vol. 49, no. 1, p. 95, 1985.

[9] C. E. Rusbult, “Commitment and satisfaction in romantic associations: A test of the investment model,” *Journal of experimental social psychology*, vol. 16, no. 2, pp. 172–186, 1980.

[10] ——, “A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvements.” *Journal of personality and social psychology*, vol. 45, no. 1, p. 101, 1983.

[11] J. Gottman, *What predicts divorce?: The relationship between marital processes and marital outcomes*. Routledge, 2023.

[12] J. M. Gottman and R. W. Levenson, “Marital processes predictive of later dissolution: behavior, physiology, and health.” *Journal of personality and social psychology*, vol. 63, no. 2, p. 221, 1992.

[13] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[14] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.

[15] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[16] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, “A meta-analysis of factors affecting trust in human-robot interaction,” *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[17] E. J. De Visser, R. Pak, and T. H. Shaw, “From ‘automation’to ‘autonomy’: the importance of trust repair in human–machine interaction,” *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018.

[18] R. W. Picard *et al.*, “Affective computing, 1997,” *Google Scholar Google Scholar Digital Library Digital Library*, 1997.

[19] B. Reeves and C. Nass, “The media equation: How people treat computers, television, and new media like real people,” *Cambridge, UK*, vol. 10, no. 10, pp. 19–36, 1996.

[20] C. Nass and Y. Moon, “Machines and mindlessness: Social responses to computers,” *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000.

[21] T. W. Bickmore and R. W. Picard, “Establishing and maintaining long-term human-computer relationships,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[22] S. Brave, C. Nass, and K. Hutchinson, “Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent,” *International journal of human-computer studies*, vol. 62, no. 2, pp. 161–178, 2005.

[23] J. Cassell, “Embodied conversational agents: representation and intelligence in user interfaces,” *AI magazine*, vol. 22, no. 4, pp. 67–67, 2001.

[24] H. Prendinger and M. Ishizuka, “The empathic companion: A character-based interface that addresses users’affective states,” *Applied artificial intelligence*, vol. 19, no. 3-4, pp. 267–285, 2005.

[25] H.-Y. Shum, X.-d. He, and D. Li, “From eliza to xiaoice: challenges and opportunities with social chatbots,” *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[26] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, “The design and implementation of xiaoice, an empathetic social chatbot,” *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[27] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, “A persona-based neural conversation model,” *arXiv preprint arXiv:1603.06155*, 2016.

[28] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too?” *arXiv preprint arXiv:1801.07243*, 2018.

[29] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith *et al.*, “Recipes for building an open-domain chatbot,” *arXiv preprint arXiv:2004.13637*, 2020.

[30] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[31] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.

[32] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.

[33] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” in *Recommender systems handbook*. Springer, 2010, pp. 1–35.

[34] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, “Session-based recommendations with recurrent neural networks,” *arXiv preprint arXiv:1511.06939*, 2015.

[35] A. See, S. Roller, D. Kiela, and J. Weston, “What makes a good conversation? how controllable attributes affect human judgments,” *arXiv preprint arXiv:1902.08654*, 2019.

[36] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, “Survey on evaluation methods for dialogue systems,” *Artificial Intelligence Review*, vol. 54, no. 1, pp. 755–810, 2021.

[37] J. W. Pennebaker, M. E. Francis, R. J. Booth *et al.*, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[38] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[39] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[40] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word–emotion association lexicon,” *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[41] J. D. M.-W. C. Kenton, L. K. Toutanova *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, vol. 1, no. 2. Minneapolis, Minnesota, 2019.

[42] P. S. Fader, B. G. Hardie, and K. L. Lee, “Rfm and clv: Using iso-value curves for customer base analysis,” *Journal of marketing research*, vol. 42, no. 4, pp. 415–430, 2005.

[43] A. Lemmens and C. Croux, “Bagging and boosting classification trees to predict churn,” *Journal of Marketing Research*, vol. 43, no. 2, pp. 276–286, 2006.

[44] W. Verbeke, D. Martens, and B. Baesens, “Social network analysis for customer churn prediction,” *Applied Soft Computing*, vol. 14, pp. 431–446, 2014.

[45] H. L. O’Brien and E. G. Toms, “What is user engagement? a conceptual framework for defining user engagement with technology,” *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[46] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, “Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support,” *Annals of behavioral medicine*, pp. 1–17, 2016.

[47] R. Lewicki, B. Polin, R. Lount Jr *et al.*, “An exploration of the structure of effective apologies. negotiation and conflict management research, 9 (2), 177–196,” 2016.

[48] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, “Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations.” *Journal of applied psychology*, vol. 89, no. 1, p. 104, 2004.

[49] C. R. Rogers, “The necessary and sufficient conditions of therapeutic personality change.” *Journal of consulting psychology*, vol. 21, no. 2, p. 95, 1957.

[50] N. Tavuchis, *Mea culpa: A sociology of apology and reconciliation*. Stanford University Press, 1993.

[51] A. Lazare, *On apology*. Oxford University Press, 2005.

[52] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[53] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[54] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 2, 2002.

[55] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.