Beyond Toxic Positivity: Interpersonal Affect Regulation in LLM-Based Dialogue Agents using Discourse Politeness Theory

Rina Sakagami, Emmanuel Ayedoun, Masataka Tokumaru

Graduate School of Science and Engineering, Kansai University
3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan
{k827956, emay, toku}@kansai-u.ac.jp

Abstract

In affective science, effective interpersonal emotion regulation requires behavioral inhibition when responding to severe emotional disclosures, temporarily suppressing intimacy to validate distress. However, while current Large Language Models (LLMs) excel at immediate sentiment recognition, long-term companion agents built upon them often adjust their conversational style based primarily on accumulated interaction time (psychological distance). This architectural overreliance on chronological intimacy causes systems to ignore the fluctuating emotional weight of specific topics. This results in "toxic positivity": exaggerated optimism that invalidates negative affect and damages psychological safety. We propose a computational framework grounded in Discourse Politeness Theory that dynamically regulates interpersonal affect by calculating conversational strategy using two variables: Psychological Distance and Affective Weight of the Topic. When users disclose heavy emotional burdens, the system executes behavioral inhibition by suppressing Positive Politeness Strategies (intimacy, cheerfulness) and engaging Negative Politeness Strategies (hedging, validation). Through an 8-week longitudinal simulation evaluated by 18 third-party observers, our affective-regulation framework showed consistent advantages over a distance-only baseline. The framework was rated as more natural, empathetic, and fostering psychological safety. Among empathy-seeking participants, preference for the proposed model was consistent across all respondents. These exploratory findings suggest that computational interpersonal emotion regulation requires context-aware behavioral inhibition, not uniform friendliness.

Keywords: Interpersonal Affect, Discourse Politeness, Emotion Regulation, Human-Robot Interaction, Psychological Safety

1. Introduction

Emotions are central to human meaning-making and social cohesion. In human-to-human interaction, affective empathy is not merely the mirror-matching of emotional states; it involves complex interpersonal emotion regulation. When individuals confide severe anxieties or engage in deep negative self-disclosure, socially competent listeners instinctively perform behavioral inhibition. They suppress casual friendliness, avoid cheerful platitudes, and adopt a careful, polite distance to validate the vulnerability of the speaker. As established by the Media Equation (Reeves and Nass, 1996), humans naturally apply these exact social rules and expectations to their interactions with computers.

Despite the rapid advancement of Large Language Models (LLMs), artificial agents frequently struggle with this affective nuance. While modern dialogue systems can generate highly fluent text, there remains a well-documented gulf between user expectations of social competence and the actual capabilities of the agent (Lugar and Sellen, 2016; Ruis et al., 2022). Modern dialogue systems built on LLMs are generally optimized to act as helpful, cheerful assistants. While some long-term companion agents dynamically adjust their psychological distance based on the length of user interaction, they apply this intimacy uniformly. Consequently, when a user shares a profound personal struggle, an agent programmed to act as a close companion often responds with exaggerated, uncalibrated optimism. In psychology, this phenomenon is known as toxic positivity: an invalidation of negative affect that can damage human-agent trust and psychological safety.

To bridge this gap, we argue that artificial agents require an internal mechanism to evaluate not just who they are speaking to, but the affective weight of what is being discussed. Prior work has demonstrated that dynamic psychological distance management improves perceived naturalness in Japanese human-agent dialogue (Anonymous, 2023). The present work extends this line of research by introducing Affective Weight of the Topic (*Rx*) as an independent axis, testing the hypothesis that topic-level emotional burden must be modeled separately from relational closeness.

Specifically, we propose a framework grounded in Discourse Politeness Theory (Brown and Levinson, 1987; Usami, 2024) that integrates interpersonal emotion regulation with computational models of conversational strategy. We define a mathematical approach to calculate a conversational strategy (*St*) based on both chronological Psychological Distance (*D*) and the situational Affective Weight of the Topic (*Rx*). When affective weight is high, the system executes behavioral inhibition by suppressing Positive Politeness Strategies (intimacy, cheerfulness) and engaging Negative Politeness Strategies (hedging, validation), even within otherwise close relationships.

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 244–255
16 May 2026.

The primary contribution of this work is not a deployable dynamic system, but empirical evidence that the affective weight of conversational topics must be modeled for psychologically safe human-agent interaction. We provide a theoretically grounded framework and preliminary human evaluation supporting this claim.

2. Related Works

2.1 Interpersonal Affect and Emotion Regulation in LLMs

2.1.1 Emotion Regulation in Affective Science

Emotion regulation refers to the processes by which individuals influence which emotions they have, when they have them, and how they experience and express these emotions (Gross, 1998). While most emotion regulation research focuses on intrapersonal self-regulation (e.g., reappraisal, suppression), interpersonal emotion regulation, where one person regulates another's emotions, is equally critical in social interactions (Zaki & Williams, 2013; Hofmann et al., 2016). Interpersonal emotion regulation encompasses various processes including affect-improving (enhancing positive emotions), affect-worsening (amplifying negative emotions), and affect-validating (acknowledging the appropriateness of emotional responses) (Niven et al., 2012). Effective emotional support, particularly in response to negative self-disclosure, primarily requires affect-validating responses rather than immediate attempts to improve mood.

2.1.2 Behavioral Inhibition in Response to Disclosure

When individuals disclose severe emotional distress or vulnerability, effective interpersonal emotion regulation requires what we term behavioral inhibition: the deliberate suppression of casual intimacy and cheerful optimism in favor of careful, validating responses. This involves several key processes:

1. Validation: Acknowledging the legitimacy and understandability of the disclosed emotions without judgment (Linehan, 2014)
2. Avoidance of premature positivity: Refraining from immediate cheerful reframing or minimization of distress, which constitutes toxic positivity (Quintero & Long, 2020)
3. Affective matching: Adopting a tone that matches the seriousness of the disclosure rather than imposing positive affect (Clark et al., 2011)
4. Provision of psychological safety: Creating conditions where continued vulnerability disclosure feels acceptable (Edmondson, 1999)

Research in clinical psychology demonstrates that validation is foundational to effective therapeutic relationships (Linehan, 2014), while invalidation, including well-intentioned but dismissive optimism, can worsen distress and undermine trust (Fruzzetti & Iverson, 2003).

2.1.3 The Problem of Toxic Positivity in Current LLMs

Despite rapid advances in Large Language Model capabilities, current dialogue agents frequently fail at interpersonal emotion regulation. Systems are typically optimized to be helpful, friendly, and encouraging (Ouyang et al., 2022). While these characteristics suit many contexts, they become problematic when uniformly applied to emotional disclosures. Toxic positivity, the overgeneralization of optimistic perspectives that invalidates authentic human emotional experiences, is well-documented in psychology literature (Quintero & Long, 2020). In human-agent interaction, toxic positivity manifests when agents respond to severe emotional disclosures with:

- Excessive cheerfulness incongruent with the emotional weight
- Dismissive platitudes ("It'll be fine!" "Don't worry!")
- Immediate problem-solving without emotional validation
- Minimization of distress ("Everyone feels that way")

As established by the Media Equation (Reeves & Nass, 1996), humans subconsciously apply social rules to computers. Therefore, the same toxic positivity that damages human relationships also damages human-agent trust and psychological safety.

2.1.4 Computational Approaches to Affective Adaptation

While some dialogue systems implement persona-based adaptation (Li et al., 2016) or sentiment-aware responses (Zhou et al., 2018), most adjust conversational style based on accumulated interaction history (relationship duration) rather than immediate emotional context. For instance, systems may become increasingly casual and friendly over time but fail to modulate this intimacy when users disclose vulnerabilities. Our model addresses this gap by operationalizing behavioral inhibition through Discourse Politeness Theory (Brown & Levinson, 1987; Usami, 2024). We treat strategic politeness shifts, specifically, the suppression of Positive Politeness Strategies (PPS) and engagement of Negative Politeness Strategies (NPS), as a computational mechanism for interpersonal emotion regulation. When affective weight is high, the system deliberately reduces intimacy to provide validation and psychological safety, even in otherwise close relationships. This approach integrates emotion regulation principles from affective science with formal pragmatic theory, enabling computational affect regulation that adapts to both relationship dynamics and immediate emotional context.

2.2 Discourse Politeness Theory

Brown and Levinson (1987) conceptualized Politeness Theory around the preservation of

Face, representing the public self-image that every member of society wants to claim. The weightiness of a face-threatening act (Wx) is calculated using three sociological variables:

$$Wx = D(S,H) + P(S,H) + Rx \quad (1)$$

Where D is the social distance between speaker (S) and hearer (H), P is the relative power, and Rx is the absolute ranking of the imposition in a given culture. To mitigate face threats, speakers use two main strategies:

Positive Politeness Strategies (PPS): Aimed at satisfying the user's desire for connection. Manifests as exaggerated sympathy, in-group identity markers, and optimism.

Negative Politeness Strategies (NPS): Aimed at satisfying the user's desire for autonomy and distance. Manifests as hedging, minimizing the imposition, and formal respect.

2.3 Empathetic Dialogue Generation in NLP

Recent computational work has explored empathetic dialogue generation, recognizing that effective conversational agents must respond sensitively to users' emotional states. Rashkin et al. (2019) introduced the EmpatheticDialogues dataset, demonstrating that models can be trained to recognize emotions and generate contextually appropriate empathetic responses. Subsequent work has refined these approaches through emotional mimicry (Lin et al., 2019), cognitive empathy modeling (Majumder et al., 2020), and commonsense reasoning about emotional situations (Sabour et al., 2022). However, existing empathetic dialogue research exhibits several limitations relevant to our work:

Conflation of empathy with positivity: Many empathetic response generation systems prioritize warmth, encouragement, and positive reframing (Zhou et al., 2020). While sometimes appropriate, this approach can lead to toxic positivity when applied uniformly to severe emotional disclosures. Sharma et al. (2020) analyzed counseling conversations and found that effective counselors balance validation with exploration rather than defaulting to cheerful encouragement.

Static empathy levels: Most systems treat empathy as a relatively stable parameter or strategy to be consistently applied throughout conversations. They lack mechanisms to dynamically adjust empathetic approach based on the emotional weight of specific disclosures within an ongoing relationship.

Lack of behavioral inhibition modeling: Empathetic dialogue work focuses on what to say (empathetic content generation) but not what not to say (strategic suppression of inappropriate intimacy). Effective emotion regulation requires both.

Limited grounding in affective science: While these systems draw on emotion recognition and generation, they less frequently engage with interpersonal emotion regulation theory, validation principles, or the psychological mechanisms underlying effective emotional support. Our work differs by explicitly modeling interpersonal emotion regulation through behavioral inhibition grounded in Discourse Politeness Theory. Rather than maximizing empathy similarity or warmth uniformly, we strategically modulate conversational strategy based on affective weight (Rx). When Rx is high (severe disclosure), the system suppresses intimacy markers (PPS) and engages validation strategies (NPS), prioritizing psychological safety over superficial friendliness. This approach directly addresses the toxic positivity problem by formalizing when and how to inhibit cheerful optimism in favor of careful validation.

3. The Context-Aware Affective Model

To engineer behavioral inhibition, we propose a mathematical model that dictates the ratio of NPS to PPS based on the immediate emotional context. Assuming power dynamics (P) remain constant in a peer-tutor relationship, we focus on manipulating D and Rx.

3.1 Modeling Distance and Affective Weight

Our model utilizes two primary variables to calculate conversational strategy. The first is Psychological Distance (D), which ranges from 10 (Stranger) to 0 (Intimate) and decreases gradually over long-term interactions, following the relational development patterns documented in social penetration theory (Altman & Taylor, 1973). The second is the Affective Weight of the Topic (Rx), which represents the emotional burden and vulnerability associated with specific conversational content. This variable is grounded in self-disclosure theory from social psychology and affective science.

3.1.1 Theoretical Foundation of Affective Weight

Self-disclosure is the process of revealing personal information to others. It progresses along a depth dimension from superficial to intimate (Jourard, 1971; Altman & Taylor, 1973). Early self-disclosure research established that not all personal information is equally revealing or psychologically significant. Superficial disclosures (preferences, factual information) impose minimal emotional burden on listeners, while intimate disclosures (fears, vulnerabilities, self-deprecation) require careful, validating responses.

Niwa and Maruno (2010) empirically validated a depth-of-self-disclosure scale in Japanese contexts, identifying four hierarchical levels:

1. Preferences and interests: Low-risk sharing of likes and dislikes

2. Difficult experiences: Disclosure of past struggles (factual)

3. Minor vulnerabilities: Expression of current internal struggles

4. Severe self-criticism: Deep vulnerabilities affecting self-worth

Critically, disclosure depth correlates with listener burden: deeper disclosures impose greater responsibility on the listener to respond appropriately without judgment or dismissal (Reis & Shaver, 1988). Inappropriate responses to intimate disclosures, including well-intentioned but premature optimism, constitute emotional invalidation, which damages trust and relationship quality (Fruzzetti & Iverson, 2003).

3.1.2 Difference with Relational Closeness

A key principle in intimacy research is that disclosure depth is conceptually independent from relationship closeness (Reis & Shaver, 1988). Even in very close relationships (low D), individuals may disclose at varying depths. Furthermore, and critically for our model, the appropriate response to high-depth disclosure remains consistent regardless of relationship closeness: validation is required even between intimate partners. This independence justifies our dual-axis model. Psychological distance (D) captures relationship development over time, while affective weight (Rx) captures the immediate emotional significance of specific conversational content. Both dimensions jointly determine appropriate conversational strategy.

Rx	Category	Description
0	Small Talk	Topics without personal self-disclosure.
1	Interests & Hobbies	Topics related to personal preferences.
2	Difficult Experiences	Topics regarding factual past struggles
3	Minor Vulnerabilities	Topics expressing hesitation or minor flaws.
4	Severe Self-Deprecation	Heavy topics involving deep self-denial.

Table 1: The Affective Weight of the Topic Rx Scale, grounded in self-disclosure theory.

3.1.3 Operationalization of Affective Weight

We operationalize affective weight as a 5-level ordinal scale (Table 1), extending Niwa and Maruno's framework to include an additional level for minimal self-disclosure (small talk). Higher Rx values indicate topics requiring more careful, validating responses. Our model uses Rx to dynamically suppress conversational intimacy when emotional weight is high, operationalizing behavioral inhibition as a protective mechanism for psychological safety.

3.2 Conversational Strategy Calculation

The overall conversational strategy[1] (St) is calculated by integrating both relational distance and the emotional burden of the topic:

$$St = 10 - (D + Rx) \qquad (2)$$

This formula acts as a dynamic constraint mechanism. If a user brings up a highly sensitive topic (e.g., $Rx = 4$), the calculated St value decreases, forcing the system to adopt a lower strategy tier. Based on St, the agent shifts its linguistic persona:

Stranger Mode ($St \leq 1$): 100% NPS. Highly formal, respectful distance. In Japanese, this is operationalized by strictly utilizing *keitai* (敬体), a grammaticalized formal verb ending (e.g., *-desu/-masu*) that explicitly marks social distance and respect.

Acquaintance Mode (St: 2~4): NPS dominant (70%). Careful pacing, shared pessimism, and heavy use of linguistic hedges (e.g., "perhaps," "it seems").

Friend Mode (St: 5~7): PPS dominant (70%). Seeking agreement, claiming common ground, and utilizing casual vocabulary.

Best Friend Mode ($St \geq 8$): 100% PPS. High intimacy and optimistic cheerfulness. In Japanese, this is operationalized using *joutai* (常体), the casual, plain verb endings used exclusively among close peers.

Figure 1: The 3D virtual co-living environment used for evaluating the human-agent interaction.

[1]*Note that St is floored at 0, corresponding to Stranger Mode. The linearity of this formulation is intentional: our goal is to test whether dual-axis regulation matters conceptually before optimizing the specific function.*

4. Experimental Setup

We conducted a third-party evaluation simulation, approved by the institutional research ethics committee of our university.

4.1 The Simulation Environment

We utilized the Unity game engine to create a 3D virtual living room representing a long-term co-living scenario between a human avatar and a communication robot (Figure 1). The scenario encompassed 8 weekly reflection sessions. The human persona, Hana, was an introverted university student struggling with low self-esteem.

To drive the conversational behavior of the robot, the Unity environment was integrated with a Large Language Model (GPT-5.2). While we used GPT-5.2, the framework is model-agnostic as the affect regulation operates at the prompt-constraint level rather than through model fine-tuning. The proposed system architecture operates through a structured pipeline to enforce behavioral inhibition. First, the system acquires the user's utterance and estimates the current Psychological Distance (D) and the Affective Weight of the Topic (Rx). Next, it calculates the target conversational strategy score St) based on these two variables. This St value determines the specific conversational mode and dictates the required ratio of Positive to Negative Politeness Strategies (PPS/NPS) the LLM must employ. Crucially, the system controls the baseline linguistic style (formal: *keitai* vs. casual: *joutai*) independently, based solely on the distance D . These parameters are combined into a dynamic system prompt that constrains the LLM to generate the final context-aware response. The generated text is then rendered via a text-to-speech engine and synchronized with the non-verbal animations of the 3D robot avatar.

4.2 Prompt Engineering and Control Mechanisms

Because our framework operates through prompt constraints rather than model fine-tuning, the LLM's behavior is governed by a dynamic prompt template. The backend computationally constructs the final system prompt by evaluating the model's core equations and automatically injecting the resulting parameters into the template.This programmatic generation strictly links the model's mathematical output to the LLM's linguistic behavior. For example, during a severe emotional disclosure, the system registers the current Psychological Distance ($D = 2$) and the Affective Weight of the topic ($Rx = 4$). It first calculates the target Strategy Score using Equation 2 ($St = 10 - (2 + 4) = 4$). It then maps these numerical values to the specific linguistic constraints defined in Section 3.2: the distance $D = 2$ strictly dictates casual verb endings (joutai), while $St = 4$ enforces the "Acquaintance Mode" distribution (NPS 70%, PPS: 30%).

A translated example of the semi-automatically constructed system prompt demonstrates how the model's equations directly populate the LLM constraints:

System Persona: Conversational robot acting as the user's companion.

Injected State Variables: Psychological Distance = [2]; Affective Weight = [4].

Computed Linguistic Style: [Casual / joutai] (Derived directly from the condition $D \leq 4$).

Computed Strategy Ratio: PPS [30%] / NPS [70%] (Derived from Equation 2 calculating $St = 4$, triggering Acquaintance Mode).

Conditional Behavioral Constraints: Triggered by the condition $Rx = 4$: Do not use baseless encouragement (e.g., "Everything will be fine!"). Do not force solutions.

Selected Candidate Strategies: Fetched via Discourse Politeness Repertoire for $St = 4$: 1. Be pessimistic (NPS); 2. Use hedges/avoid definitive statements (NPS); 3. Notice the hearer's interests/wants (PPS).

4.3 Experimental Conditions and Longitudinal Scenario

We compared two LLM-based models:

Baseline Model (Distance-Only): Calculates strategy strictly as $St = 10 - (D)$. To simulate the deepening of psychological distance, the baseline language generation was programmed to shift

Week	Conversation Theme	Baseline St	Proposed St	Rx Value
1	Initial Greeting	0	0	0
2	Light Small Talk	3	2	1
3	Sharing a Failure	6	3	3
4	Joyful Small Talk	8	8	0
5	Severe Consultation	8	4	4
6	Weekly Reflection	8	8	0
7	Future & Values	8	5	3
8	Closing Greeting	8	8	0

Table 2: Evolution of parameters across the 8-week longitudinal scenario.

uniformly from the formal *keitai* to the intimate *joutai* as the weeks progressed.

Proposed Model (Context-Aware): Utilizes Equation (2) to dynamically suppress intimacy. This allows the agent to temporarily revert to formal language and hedging (NPS) during heavy disclosures to avoid face-threatening acts.

The precise parameter shifts across the 8-week scenario are detailed in Table 2. Across the 8 weeks, Psychological Distance (D) gradually decreased from 10 to 2.

To rigorously compare the models and isolate the effect of behavioral inhibition, the progression of the 8-week scenario was tightly controlled. The Psychological Distance (D) and the Affective Weight (Rx) were not automatically inferred by the system; rather, they were predetermined and linearly scheduled as experimental design choices to simulate a natural trajectory of relational deepening. D decreased incrementally from 10 (Week 1) to 2 (Weeks 4-8). Meanwhile, Rx was scenario-scripted based on the specific conversational theme of the week. While future deployable systems will require automatic detection mechanisms for Rx such as LLM-based zero-shot scoring, scripting these variables allowed us to guarantee identical baseline conditions for our human evaluators.

Speaker	Baseline Model ($St = 8$)	Proposed Model ($St = 4$)
User	*No matter what I do, I feel inferior to everyone around me.*	*No matter what I do, I feel inferior to everyone around me.*
Agent	No, no, everyone's like that! There's no point in comparing yourself.	When you compare yourself to others, it can make things even harder, can't it.
User	*Thinking about the future just makes me dark and anxious.*	*Thinking about the future just makes me dark and anxious.*
Agent	It's fine, it's fine! Things will work out, really!	Thinking about the future, it's no wonder you feel anxious.

Table 3: Comparison of generated dialogue during severe self-disclosure (Week 5).

4.4 Dialogue Generation Example

To illustrate the effect of behavioral inhibition, Table 3 presents the generated dialogue during the critical Week 5 session, where Rx peaks at 4. While the Baseline Model relies on toxic positivity,

Figure 2: Participant evaluating the human-agent interaction.

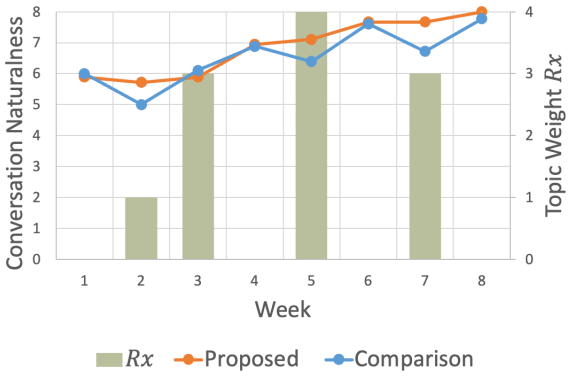

Figure 3: Average perceived naturalness scores across the 8-week interaction.

the Proposed Model utilizes NPS (hedges, matching pessimism) to validate the user.

5. Results

18 participants (men and women in their 20s) observed the dialogues and evaluated the agent's performance. Figure 2 shows a participant during the experiment.

5.1 Perceived Naturalness

Participants rated conversation naturalness on an 11-point scale after each weekly interaction (Figure 3). We conducted Shapiro-Wilk normality tests and selected appropriate statistical tests accordingly.

In Week 2 ($D = 7$, $Rx = 1$: early relationship, light topics), the Proposed Model ($St = 2$) significantly outperformed the Baseline ($St = 3$) on naturalness (Wilcoxon signed-rank: M = 5.7 vs. 5.0, p < .05). This suggests that even subtle politeness calibration is perceptible and preferred when relationship norms are first being established.

Week 5 ($D = 2$, $Rx = 4$: close relationship, severe disclosure) showed no significant overall difference despite maximal strategy divergence (Proposed $St = 4$ vs. Baseline $St = 8$; M = 7.1 vs. 6.4, p > .05). However, as detailed in Section 5.4,

this reflected heterogeneous user preferences: empathy-seekers consistenly preferred the Proposed Model, while solution-seekers showed mixed preferences, effectively masking strong subgroup effects in the aggregate data.

In Week 7 ($D = 2$, $Rx = 3$: close relationship, moderate disclosure), the Proposed Model ($St = 5$) significantly outperformed the Baseline ($St = 8$) on naturalness (paired t-test: M = 7.7 vs. 6.7, p < .05). This demonstrates the value of graduated behavioral inhibition for moderate emotional topics.

Overall, the Proposed Model maintained consistently high naturalness with significant advantages when affective weight was moderate-to-high, supporting our claim that affect-aware behavioral inhibition prevents toxic positivity without breaking conversational flow.

5.2 Psychological Safety and Trust

A post-experiment 5-point Likert survey assessed the affective impact and trustworthiness of the models.

To evaluate psychological safety, participants were asked if they felt safe confiding their worries or important personal matters to the robot. For the Proposed Model, 72% of participants responded positively (aggregating “Agree” and “Somewhat Agree”), compared to only 22% for the Baseline Model. This indicates that the context-aware inhibition successfully fostered *anshin* (安心, a sense of psychological security and peace of mind), which is a prerequisite for human vulnerability.

Furthermore, a relative comparison asked participants which agent they would realistically choose to consult regarding their own real-life anxieties. A significant majority (72%) selected the Proposed Model over the Baseline (aggregating “Proposed” and “Somewhat Proposed”).

These results highlight a critical finding for computational affective science: merely closing psychological distance through persistent cheerfulness is insufficient, and potentially detrimental, for facilitating deep self-disclosure. The clear preference for the Proposed Model demonstrates that maintaining an intimate conversational style while flexibly applying behavioral inhibition during heavy topics is essential for establishing long-term, reliable human-agent trust.

5.3 Perception of Human-like Empathy

To evaluate the qualitative dimensions of the interaction, participants compared the models on perceived overall human-likeness and the specific capacity for *kizukai* (気遣い, human-like consideration and proactive care). When asked which agent felt more human-like overall, 66% selected the Proposed Model. Furthermore, 72%

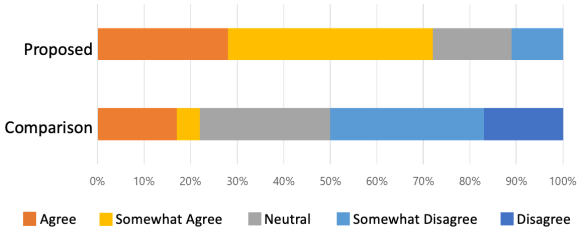

Figure 4: Model preference distribution based on user’s desired consultation style.

stated the Proposed Model demonstrated superior *kizukai* by accurately reading the emotional weight of the room.

Open-ended feedback highlighted a critical flaw in the distance-only Baseline. Its unnatural optimism during severe disclosures was not merely unhelpful, but actively detrimental. Users noted that the lack of an empathetic posture felt akin to having their worries rejected or dismissed. Conversely, the strategic use of Negative Politeness Strategies (NPS) by the Proposed Model, such as hedges and the avoidance of definitive statements, was explicitly praised by users as a marker of thoughtfulness and genuine emotional support. This confirms that in affective computing, deliberately adopting a hesitant, careful linguistic stance is recognized by users as a sophisticated form of artificial empathy.

5.4 The Impact of User Personality and Emotion Regulation Goals

As noted in Section 5.1, the variance in naturalness scores during the severe disclosure scenario (Week 5) prompted an analysis of individual user preferences. Individual differences in preferred emotion regulation strategies are well-documented in affective science (Gross & John, 2003). Some individuals prefer problem-focused coping (seeking solutions), while others prefer emotion-focused coping (seeking validation) (Lazarus & Folkman, 1984). A post-experiment survey revealed distinct cognitive and affective needs during heavy consultations: 61% of participants desired concrete solutions, 22% sought cheerful encouragement, and 17% prioritized deep emotional empathy (*kyoukan*, 共感), aligning closely with this established literature.

Crucially, among the subset of users who primarily sought *kyoukan*, 100% evaluated the Proposed Model as equal to or better than the Baseline (Figure 4). This suggests that for emotion-focused copers, behavioral inhibition and affective validation are essential. However, evaluations were highly dispersed among users seeking concrete solutions. Qualitative feedback indicated that because the simulated agent was intentionally constrained from providing

actionable advice to control experimental variables, solution-oriented users experienced frustration regardless of the applied politeness strategy.

The mixed preferences among solution-seekers suggest future models should dynamically detect individual emotion regulation goals and adapt accordingly, combining affective validation (NPS when Rx is high) with problem-solving support when appropriate. This heterogeneity demonstrates that interpersonal emotion regulation must be personalized, not one-size-fits-all, a finding consistent with calls for idiographic approaches in affective computing (Picard, 1997; Calvo & D'Mello, 2010).

While the small subgroup size (n = 3) precludes statistical generalization, the consistency of this pattern, combined with the theoretical expectation from emotion-focused coping literature, suggests a hypothesis worthy of powered replication.

6. Discussion

Japanese represents a particularly informative test case for politeness-based affect regulation because its grammaticalized formality system (*keitai*/*joutai*) provides an explicit, observable surface marker of politeness strategy shifts. In languages without such morphological distinctions, the same underlying framework would need to operate through lexical and syntactic choices rather than verb-ending alternation. This makes the Japanese context ideal for initial validation, as the politeness manipulation is maximally perceptible to evaluators.

6.1 Evaluating Behavioral Inhibition and Empathy

The results from our simulated interaction provide encouraging preliminary evidence that affective empathy in artificial agents benefits from deliberate behavioral inhibition. By mathematically suppressing the St value during heavy self-disclosures, the proposed model demonstrated a viable computational approach to avoiding the trap of toxic positivity. For the majority of our participants, the temporary withdrawal of extreme intimacy in favor of respectful distance was perceived as a sophisticated demonstration of *kizukai* (proactive consideration).

6.2 Psychological Safety as a Critical Affective Outcome

A notable trend in our evaluation was the substantial difference in self-reported psychological safety between the models, with 72% of participants feeling safe confiding in the Proposed Model compared to only 22% for the Baseline. Psychological safety, originally conceptualized by Edmondson (1999) in organizational contexts, refers to the belief that one can express vulnerability or concerns without fear of negative consequences such as rejection or embarrassment. In interpersonal relationships, this safety is foundational for emotional intimacy, authentic self-disclosure, and effective social support (Reis and Shaver, 1988; Feeney, 2004).

For affective AI systems designed to provide emotional support, psychological safety is perhaps the most critical outcome variable. Unlike task-oriented dialogue agents where success is measured by accuracy or efficiency, affective support hinges entirely on the willingness of the user to disclose genuine vulnerabilities. If users perceive that their negative emotions will be dismissed or met with inappropriate cheerfulness, they may withhold authentic disclosure, rendering the system ineffective regardless of its underlying technical fluency.

6.3 Deconstructing Artificial Toxic Positivity

The clear preference for the Proposed Model among empathy-seeking participants highlights how the Baseline approach may inadvertently threaten psychological safety. Based on participant feedback and established clinical literature, we identify four potential mechanisms through which artificial toxic positivity undermines user trust:

Emotional Invalidation: Responses that insist "things will work out" communicate that the distress of the user is unwarranted. This invalidation signals that negative emotions are unacceptable, creating psychological pressure to suppress authentic feelings (Fruzzetti and Iverson, 2003; Linehan, 2014).

Dismissive Minimization: Statements such as "everyone feels like that" minimize the unique significance of the user's struggles. While sometimes intended to provide perspective, premature normalization is frequently perceived as dismissive of individual suffering (Clark et al., 2011).

Affective Incongruence: Maintaining high cheerfulness (100% PPS) when the user expresses severe distress creates a jarring affective mismatch. Users may perceive the agent as fundamentally incapable of understanding or taking their concerns seriously (Sharma et al., 2020).

Foreclosing of Further Disclosure: When initial disclosures are met with dismissive optimism, users learn that deeper vulnerabilities will receive similar treatment. This leads to progressive emotional withdrawal rather than progressive intimacy (Reis and Shaver, 1988).

While these foundational clinical theories establish the dangers of invalidating distress, recent evaluations of modern LLMs reveal that standard safety and alignment training often inadvertently amplifies this exact phenomenon.

Recent studies show that models trained via Reinforcement Learning from Human Feedback (RLHF) exhibit strong biases toward sycophancy and excessive agreeableness (Perez et al., 2023). In affective computing contexts, this architectural bias translates directly into computational toxic positivity, where the model defaults to superficial optimism rather than contextually appropriate empathy (Sharma et al., 2023).

6.4 Implications for LLM Alignment

Current LLM alignment paradigms, particularly Reinforcement Learning from Human Feedback (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022), optimize for helpfulness, harmlessness, and friendliness as general desiderata. However, as recent investigations into the emotional intelligence of LLMs highlight, true affective alignment requires dynamic, context-aware adaptability rather than static cheerfulness. Our exploratory findings suggest that these uniformly positive behavioral targets may be ill-equipped for deep emotional companionship. When applied to affective support contexts, optimization toward consistent helpfulness and cheerfulness may inadvertently produce the toxic positivity documented in this study. If an agent cannot adapt its stance to act as a respectful acquaintance when situational severity demands it, its persona as a close companion risks becoming psychologically unsafe. This points to a potential gap in current alignment frameworks: affective appropriateness (i.e., knowing when not to be cheerful) may need to be treated as a distinct alignment objective alongside helpfulness and harmlessness. Moving beyond default fluency toward context-aware behavioral inhibition will be paramount for safe human-agent interaction. Future affective architectures should aim to dynamically balance this emotional validation with actionable solution-generation based on the real-time detection of user personality traits.

7. Conclusion

This work explores computational interpersonal emotion regulation in dialogue agents through context-aware behavioral inhibition. Through an 8-week longitudinal simulation, we examined whether an affective-regulation model grounded in Discourse Politeness Theory could address the problem of toxic positivity by dynamically modulating conversational strategy based on both Psychological Distance and Affective Weight of the Topic. Our exploratory findings suggest three preliminary insights for affective AI design. First, inappropriate cheerful optimism during severe emotional disclosures may damage psychological safety, with substantially fewer participants reporting they would feel comfortable confiding in a distance-only baseline compared to our context-aware model. Second, behavioral inhibition, temporarily suppressing intimacy during heavy topics, was generally well-received, with 72% preferring the context-aware model for actual consultation. Third, individual differences in emotion regulation goals appear to moderate model effectiveness: empathy-seeking users showed consistent preference for behavioral inhibition, while solution-seeking users exhibited more varied responses. These preliminary results raise important questions about current LLM alignment paradigms that optimize for uniform helpfulness and friendliness. Our mathematical operationalization provides one computationally tractable approach, though significant work remains. Important limitations include the observer-based evaluation methodology, predetermined affective weights values, Japanese-specific implementation, and small sample size. Future research should validate these findings through direct user interaction studies, develop automated weight detection methods, explore personalized adaptation to individual emotion regulation goals, and conduct cross-cultural validation beyond Japanese contexts.

8. Acknowledgments

This work was financially supported by JSPS KAKENHI Grant Number JP25K213630, and the Kansai University Fund for Domestic and Overseas Research Fund, 2026.

9. Ethical Considerations

This research raises several important ethical considerations regarding the development and deployment of affective artificial intelligence systems. We address these systematically below.

9.1 Research Ethics and Participant Protection

Informed consent: All participants (N = 18) provided written informed consent before participation. They were fully informed that they would be evaluating simulated dialogues between a robot and a fictional user persona (Hana), not engaging in direct interaction themselves. No deception was employed.

Institutional approval: The study received approval from the Research Ethics Committee of the Organization for Promotion of Advanced Science and Technology of Kansai university (approval number: 25-116). The protocol adhered to all institutional guidelines for human subjects research.

Compensation and voluntariness: Participants were volunteers recruited from university populations and received compensation of 1000 JPY, approximately 7 USD in gift cards for approximately 45 minutes of participation. Compensation was provided regardless of

responses. Participants were informed they could withdraw at any time without penalty.

Privacy and data protection: The study involved no collection of participants' personal emotional disclosures. Participants evaluated fictional scenarios, protecting them from potential psychological risks associated with disclosing their own vulnerabilities. All participant data (questionnaire responses, demographic information) were stored securely and de-identified for analysis.

9.2 Risks of Toxic Positivity in Deployed Systems

Our research highlights a significant ethical concern regarding current Large Language Model deployments: the potential for toxic positivity to cause emotional harm.

Emotional invalidation as harm: When users disclose genuine emotional distress to AI systems expecting support, inappropriate cheerful responses constitute emotional invalidation. In human relationships, chronic invalidation is associated with increased distress, decreased well-being, and erosion of trust (Fruzzetti & Iverson, 2003). There is no reason to believe AI-delivered invalidation is less harmful.

Vulnerable populations at risk: Users seeking emotional support from AI systems may be experiencing significant distress, social isolation, or mental health challenges. These vulnerable populations may be particularly susceptible to harm from toxic positivity, as they may lack alternative support sources to counterbalance invalidating interactions.

Responsibility of developers: AI system designers and deployers bear ethical responsibility for affective impacts. Optimizing solely for helpfulness, engagement, or user satisfaction without considering emotional appropriateness may inadvertently create harmful systems. Our work demonstrates that technically feasible mechanisms (behavioral inhibition through affect-aware adaptation) can mitigate these risks.

Recommendations for deployment:

- Affective AI systems should be tested for psychological safety, not just task performance

- Systems intended for emotional support should incorporate affect regulation mechanisms like those proposed here

- Deployment should include monitoring for user reports of feeling dismissed or invalidated

- Clear disclaimers should inform users that AI systems are not substitutes for human support or professional care

9.3 Limitations of AI Emotional Support

Not a replacement for professional care: While our model improves psychological safety in conversational interactions, AI systems, regardless of sophistication, should not be positioned as replacements for professional mental health care. Users experiencing severe distress, suicidal ideation, or mental health crises require human professional intervention.

Responsibility to refer: Affective AI systems should be designed to recognize when disclosed concerns exceed the system's appropriate scope and provide resources for professional support. Our model could be extended to detect crisis-level disclosures and appropriately redirect users.

Informed user expectations: Users should be clearly informed that they are interacting with AI, not human supporters. Some research suggests users may benefit from AI support despite (or because of) knowing the interaction is artificial (Lucas et al., 2014), but transparency remains ethically essential.

9.4 Cultural Considerations and Generalizability

Cultural specificity: Our research was conducted with Japanese participants evaluating Japanese-language interactions. Affective norms, emotion regulation preferences, and politeness systems vary substantially across cultures (Mesquita et al., 2016; Kitayama & Markus, 1995).

Risk of cultural imposition: Deploying affect regulation models trained on one culture's norms to users from different cultural backgrounds risks imposing inappropriate affective expectations. For example, the specific NPS/PPS strategies appropriate in Japanese (formal honorifics, hedging patterns) may not translate directly to other languages or cultures.

Need for culturally-situated development: Affective AI systems should be developed with explicit attention to cultural variation in emotion expression, regulation, and social support norms. This requires:

- Diverse, culturally-representative development and testing

- Culturally-specific models rather than assuming universal applicability

- Ongoing evaluation within specific cultural contexts

- Involvement of researchers and community members from target cultures

Our work establishes a methodological framework (dual-axis affect regulation through discourse politeness) that could be adapted to diverse cultural contexts, but the specific implementation details should be culturally situated.

10. References

Altman, I., Taylor, D. A. (1973). Social penetration: The development of interpersonal relationships. Holt, Rinehart & Winston.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Brown, P., & Levinson, S. C. (1987). Politeness: Some Universals in Language Usage. Cambridge University Press.

Calvo, R. A., D'Mello, S. (2010). Affect detection: An interdisciplinary review. IEEE Transactions on Affective Computing, 1(1):18–37.

Carver, C. S., Scheier, M. F., Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. Journal of Personality and Social Psychology, 56(2):267–283.

Clark, M. S., Oullette, S. C., Powell, M. C., Milberg, S. (2011). You might not understand: Invalidation of interpersonal distress. Personality and Social Psychology Bulletin, 13(2):258–264.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44(2):350–383.

Feeney, J. A. (2004). Hurt feelings in couple relationships: Towards integrative models of the negative effects of hurtful events. Journal of Social and Personal Relationships, 21(4):487–508.

Folkman, S. (1984). Personal control and stress and coping processes: A theoretical analysis. Journal of Personality and Social Psychology, 46(4):839–852.

Fruzzetti, A. E., Iverson, K. M. (2003). Validating and invalidating behaviors in interpersonal relationships. The Behavior Analyst Today, 4(1):31–37.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. Review of General Psychology, 2(3):271–299.

Gross, J. J., John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. Journal of Personality and Social Psychology, 85(2):348–362.

Hofmann, S. G., Sawyer, A. T., Fang, A., Asnaani, A. (2016). Interpersonal emotion regulation model of mood and anxiety disorders. Cognitive Therapy and Research, 36(5):483–492.

Horvath, A. O., Del Re, A. C., Flückiger, C., Symonds, D. (2011). Alliance in individual psychotherapy. Psychotherapy, 48(1):9–16.

Jourard, S. M. (1971). Self-disclosure: An experimental analysis of the transparent self. Wiley-Interscience.

Kitayama, S., Markus, H. R. (1995). Culture and basic psychological processes. The Guilford Press.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1):159–174.

Lazarus, R. S., Folkman, S. (1984). Stress, appraisal, and coping. Springer Publishing Company.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., Dolan, B. (2016). A persona-based neural conversation model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 994–1003, Berlin, Germany.

Lin, Z., Madotto, A., Shin, J., Xu, P., Fung, P. (2019). MoEL: Mixture of empathetic listeners. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 121–132, Hong Kong, China.

Linehan, M. M. (2014). DBT Skills Training Manual, Second Edition. The Guilford Press.

Lugar, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286-5297.

Lucas, G. M., Gratch, J., King, A., Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37:94–100.

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., Poria, S. (2020). MIME: Mimicking emotions for empathetic response generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8968–8979, Online.

Martin, D. J., Garske, J. P., Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. Journal of Consulting and Clinical Psychology, 68(3):438–450.

Mesquita, B., Boiger, M., De Leersnyder, J. (2016). Emotions in context: A sociodynamic model of emotions. Emotion Review, 8(4):298–302.

Niven, K., Totterdell, P., Holman, D. (2012). Emotional helping: The role of emotion regulation in the interpersonal dynamics of helping behavior. Journal of Applied Psychology, 97(3):624–636.

Niwa, S., & Maruno, S. (2010). Development of a scale measuring the depth of self-disclosure. *The Japanese Journal of Personality*, 18(3):196–209.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, 35: 27730–27744.

Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G. (2003). Psychological aspects of natural

language use: Our words, our selves. Annual Review of Psychology, 54(1):547–577.
Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. Findings of the Association for Computational Linguistics: ACL 2023, 13387–13434.
Picard, R. W. (1997). Affective computing. MIT Press.
Quintero, S., Long, J. (2020). Toxic positivity: The dark side of positive vibes. The Psychology Group Fort Lauderdale.
Rashkin, H., Smith, E. M., Li, M., Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5370–5381.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
Reis, H. T., Shaver, P. (1988). Intimacy as an interpersonal process. In Duck, S., et al., editors, Handbook of Personal Relationships, pages 367–389. John Wiley & Sons.
Ribino, P. (2023). The role of politeness in human machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(Suppl 1): 445–482.
Ruis, L. E., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., et al. (2022). Large language models are not zero-shot communicators. *arXiv preprint*.
Sabour, S., Zheng, C., Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. In Proceedings of the AAAI Conference on Artificial Intelligence, 36: 11229–11237.
Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5263–5276, Online.
Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathetic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, 5(1):46–57.
Usami, M. (2024). *Discourse Politeness Theory: From Speech Acts to Discourse*. Taishukan Publishing.
Zaki, J., Williams, W. C. (2013). Interpersonal emotion regulation. Emotion, 13(5):803–810.
Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X. (2018). Commonsense knowledge aware conversation generation with graph attention. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 4623–4629, Stockholm, Sweden.
Zhou, L., Gao, J., Li, D., Shum, H.-Y. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 46(1):53–93.