

MATE: Diagnosing Empathy Calibration Failures in Multi-Turn Human-LLM Interaction

Yeseon Hong¹, Junhyuk Choi², Minju Kim², Bugeun Kim^{2†}

¹Sogang University, ²Chung-Ang University

ghddptjs@sogang.ac.kr

{chlwnsgur129,minjunim,bgnkim}@cau.ac.kr

Abstract

Large language models are increasingly used in emotionally consequential interactions. Response-level evaluation, however, struggles to diagnose how empathy fails across turns. We introduce **MATE** (Multi-turn Assessment of calibraTed Empathy), a framework that evaluates empathy as a multi-turn, perception-centered process. Across two controlled studies ($N = 82$), baseline responses exhibit disclosure-insensitive miscalibration: generic validation and formulaic tone read as hollow across turns. A prompt-level self-critique condition makes these patterns less prominent but coincides with a different failure, agreement-skewed miscalibration, where affirmation becomes insufficiently contingent on disclosure context under high persona alignment (62.5% of fit participants in Study 2). Response-level analysis is consistent with a tone-agreement dissociation, as the condition shifts tone toward naturalness while agreement density also increases (+98% under fit vs. +25% under unfit). These findings reframe empathic behavior as relational calibration—consistency among disclosure depth, persona, and response behavior—rather than maximization. Response behavior itself spans tone, agreement, and contextual specificity, sub-dimensions that can drift independently.

1 Introduction

As large language models (LLMs) become embedded in everyday life, LLM-based chatbots are increasingly used in emotionally consequential interactions. Users turn to LLM-based chatbots for comfort and to discuss personal or emotionally difficult experiences (Maples et al., 2024; Luo et al., 2025; Pentina et al., 2023). In such settings, empathy becomes central to human-LLM emotional interaction, and the NLP and Human-AI Interaction (HAI) communities have actively studied

how to generate and evaluate empathic responses from LLMs (Rashkin et al., 2019; Lin et al., 2019; Sharma et al., 2020). Much of this work has treated empathy primarily as a property of responses, asking whether LLM replies are empathic or emotionally attuned to users (Liu et al., 2021; Welivita et al., 2021; Gao et al., 2021). Yet emotional interaction unfolds across sustained conversation rather than isolated replies. This response-centered view therefore leaves two aspects of empathy in human-LLM interaction underexamined: how empathy develops across turns, and how users perceive it as the conversation unfolds.

First, empathy develops across turns rather than within isolated replies. Psychological work treats empathy as a process that unfolds through ongoing disclosure and responsive engagement (Reis and Shaver, 1988; Laurenceau et al., 1998), and related work on sycophancy shows that over-validation tendencies compound across turns in ways that response-level evaluation cannot diagnose (Sharma et al., 2024; Hong et al., 2025; Cheng et al., 2025). Second, what models express and what users perceive can diverge. Linguistic markers of empathy do not reliably predict whether users feel understood, and chatbots often rely on formulaic acknowledgments that engage poorly with the specifics of user disclosures (Xu and Jiang, 2024; Cuadra et al., 2024; Liu et al., 2025; Chen et al., 2024). Beyond these substantive findings, existing empathy measures are themselves poorly suited for capturing either aspect. Trait-level scales assess stable dispositions (Davis, 1983; Baron-Cohen and Wheelwright, 2004) and response-level annotations evaluate isolated utterances (Alam et al., 2018), but neither tracks how perceived empathy shifts as conversational context accumulates.

Addressing both gaps requires a framework that treats empathy as something that unfolds between partners over time. Human-human interaction research provides such a foundation. Inter-

[†]Corresponding author

personal process models hold that empathy develops through repeated cycles of self-disclosure and responsive engagement, and that what feels empathic is shaped by how responses fit the disclosures they follow (Reis and Shaver, 1988; Laurenceau et al., 1998). This perspective connects both aspects at once: disclosure-response cycles are the unit through which empathy accumulates across turns, and the disclosures themselves provide the conversational context against which responses are perceived. Motivated by this perspective, we adopt reciprocal self-disclosure—where both partners share personal information at comparable depth—as our interactional foundation.

We introduce **MATE** (Multi-turn Assessment of calibraTed Empathy), an interaction framework that operationalizes this perspective. Following interpersonal process theory, MATE treats empathy as processual rather than scalar: empathy operates through cycles of disclosure and responsive engagement, and its success or failure is reflected in how users perceive responses and in the relational outcomes those responses shape. Its three components are staged reciprocal self-disclosure, repeated assessment of user perception, and context-bound interpretation of experienced responses.

Applying MATE, we conduct two controlled studies ($N = 82$). **Study 1** characterizes how empathy is perceived under standard LLM responses, identifying failure modes that surface only when perception accumulates across stages. **Study 2** uses prompt-level self-critique as a diagnostic probe, examining how a tone-oriented revision condition shifts the patterns observed in Study 1.

Our main contributions are as follows. (1) We introduce MATE, a framework that evaluates empathy as a multi-turn, perception-centered phenomenon by integrating staged reciprocal self-disclosure, repeated perception assessment, and context-bound qualitative analysis. (2) Applying MATE, we identify failure modes that response-level evaluation cannot surface: *disclosure-insensitive miscalibration* in baseline responses, where generic validation reads as hollow despite linguistic markers of empathic engagement; and *agreement-skewed miscalibration* under high persona alignment, where naturalness gains coincide with affirmation that is insufficiently contingent on the disclosure context. (3) These findings support reframing empathic LLM behavior as a problem of *relational calibration* rather than a maximization objective, and reveal a tone-

agreement dissociation invisible to single-response evaluation.

2 Related Work

2.1 Response-Centered Empathy Evaluation in Dialogue Systems

Empathy in dialogue systems has been studied primarily through datasets and architectures targeting empathic response generation. Following EMPATHETICDIALOGUES (Rashkin et al., 2019), subsequent work expanded this line through architectural designs such as mixture-of-experts listeners (Lin et al., 2019) and emotion-cause modeling (Gao et al., 2021), and through emotional support datasets such as ESCONV (Liu et al., 2021). These works evaluate empathy at the response level, judging whether a single reply bears linguistic and affective markers of empathic engagement.

Two recent lines suggest that response-level evaluation misses important aspects of empathy in interaction. Research on sycophancy shows that LLMs trained with human feedback systematically over-validate users, and that this tendency compounds across turns rather than manifesting within any single response (Sharma et al., 2024; Hong et al., 2025; Cheng et al., 2025)—a cross-turn accumulation pattern that response-level evaluation cannot diagnose. Work on perceived empathy further shows that expressed empathic markers do not reliably predict whether users feel understood (Xu and Jiang, 2024; Liu et al., 2025), and that LLMs recognize emotional content yet often fail to engage with users’ specific disclosures in contextually grounded ways (Cuadra et al., 2024; Chen et al., 2024). Together, these findings point to multi-turn dynamics and users’ perception as two dimensions that response-level evaluation leaves underexamined, motivating a framework that addresses both.

2.2 Reciprocal Self-Disclosure and the Development of Perceived Empathy

Interpersonal process theory provides a foundation for evaluating empathy as a multi-turn, perception-centered process. Empathy is not located in individual responses but emerges through cycles in which one partner discloses personal experience and the other responds in ways perceived as understanding, validating, and caring (Reis and Shaver, 1988; Laurenceau et al., 1998). Its success is constituted by how responses are perceived against what has been disclosed—not by the linguistic markers responses

carry in isolation. Reciprocal self-disclosure provides the interactional structure through which this process develops: social penetration theory and the “fast friends” paradigm show that disclosure deepens through mutual, escalating exchanges, with relational outcomes depending on whether responses are appropriately matched to disclosure depth (Altman and Taylor, 1973; Aron et al., 1997).

In human–chatbot interaction, similar dynamics have been documented: deeper user self-disclosure and chatbot reciprocity are associated with stronger perceived closeness (Skjuve et al., 2021; Lee et al., 2020), and closeness with chatbot partners develops gradually over extended interaction (Skjuve et al., 2023). However, these studies typically assess relational outcomes through post-interaction surveys or retrospective interviews, leaving open how perceived empathy shifts within a single interaction as context accumulates.

3 MATE: Multi-turn Assessment of calibraTed Empathy

We introduce **MATE**, a framework that operationalizes empathy evaluation as a multi-turn, perception-centered process, built on two premises (§3.1) that motivate its three components (§3.2).

3.1 Premises

P1. Empathy is constituted by perception against accumulated context. What users experience as empathic depends on how a response is interpreted against what has been disclosed and how the conversation has unfolded, not on the linguistic markers the response carries in isolation. The same response can be experienced very differently depending on the depth of disclosure and the expectations that have built up around it.

P2. Empathy is processual rather than scalar. Because empathy unfolds through cycles of disclosure and responsive engagement, it cannot be reduced to a single quantity assessed at one moment. Its success or failure manifests across the interaction, in the perceptions it produces and in the relational outcomes it shapes.

3.2 Framework Components

These premises pose a measurement problem that existing tools do not address. Trait-level empathy scales (Davis, 1983; Baron-Cohen and Wheelwright, 2004) capture stable dispositions; response-level annotations (Alam et al., 2018) score isolated

utterances. Neither tracks how perception forms as context accumulates. MATE addresses this gap through three components.

Staged reciprocal self-disclosure. Interpersonal process models identify disclosure–response cycles as the unit through which empathic engagement develops (Reis and Shaver, 1988; Laurenceau et al., 1998). MATE therefore structures reciprocal self-disclosure across stages, with both partners disclosing at comparable depths, so that conversational context accumulates comparably across interactions and stage-wise shifts in user perception can be assessed.

Repeated perception assessment. P1 and P2 jointly require that perception be assessed across stages rather than only at the end. Rather than reducing perceived empathy to a single direct scalar, MATE tracks it through relational outcomes that empathic engagement is theorized to shape. These outcomes do not measure empathy itself; they index what empathy produces when it succeeds, making stage-wise trajectories—rather than a single endpoint—the object of analysis.

Context-bound qualitative interpretation. Relational outcomes show whether perception changes across stages, but not the contextual basis on which any response was perceived as it was. P1 implies this basis is irreducible: a response’s interpretation depends on the specific disclosures it follows. MATE therefore treats qualitative analysis of users’ post-interaction interpretations as a complementary lens, examining how responses were experienced in context.

Calibration as relational consistency. Together, these three components let MATE characterize not only whether perceived empathy succeeds across stages, but also how it is interpreted within context. We frame what this evaluation targets as *calibration*: not the intensity of empathic expression, but the relational consistency among three elements that jointly shape perceived empathy—the user’s self-disclosure depth, the agent’s persona in relation to the user, and the agent’s response behavior. Response behavior itself is multi-dimensional, comprising tone, agreement or affirmation, and contextual specificity. A response is *relationally calibrated* when these behavioral dimensions remain contingent on both what the user has disclosed and what the agent’s persona warrants. *Miscalibration*

occurs when one or more dimensions drift from this configuration.

4 Method

We instantiate MATE in two controlled studies that share the same staged self-disclosure protocol, stage-wise relational outcomes, and post-study interviews. Study 1 uses standard GPT-4o responses, whereas Study 2 adds within-turn self-critique before each response is shown.

4.1 Staged Reciprocal Self-Disclosure Protocol

Following Aron et al. (1997), we used 33 self-disclosure prompts ordered from shallow to deep, grouped into three stages (11 prompts per stage). We excluded three items from the original 36-item protocol that required visual perception of the partner (e.g., “Name three things you and your partner appear to have in common”). At each stage, the LLM agent reciprocated at the same disclosure depth, enabling bidirectional progression of self-disclosure cycles. Participants used a text chatroom; each stage lasted about 15 minutes, followed by a 2-minute measurement period.

4.2 Measures

We combine stage-wise relational outcomes with post-study interviews, avoiding repeated direct empathy ratings that could make empathy itself an explicit object of evaluation during the interaction.

Stage-wise relational outcomes. Empathic engagement, when it succeeds, shapes the relational outcomes between partners. After each stage, participants reported two such outcomes using validated scales: the Subjective Closeness Index (SCI; Berscheid et al. (1989)), a 2-item measure of perceived closeness (7-point Likert), and the Interpersonal Judgment Scale (IJS; Byrne (1971)), a 2-item measure of collaborative affinity (7-point Likert). Repeated stage-wise administration yields trajectories as context accumulates. We treat SCI as the primary outcome for tracking how perception of empathy develops, with IJS as a contrast separating closeness from task-oriented affinity.

Post-study qualitative interviews. Outcome trajectories show whether perception shifts across stages, but not the contextual basis on which any response was perceived as it was. We conducted post-study interviews to collect participants’ interpretations of LLM responses and applied open

coding to extract patterns of how responses were experienced in context.

4.3 Study 1: Baseline LLM Responses

Study 1 established a baseline using standard GPT-4o responses without additional intervention. We recruited 29 Korean participants (17 men, 12 women; ages 19–28, $M = 23.0$, $SD = 1.89$), yielding 1,914 utterances ($= 29 \times 33 \times 2$, with participant and LLM responses interchanged for each prompt).

Persona similarity is one dimension of the conversational context against which empathy is perceived; we therefore manipulate it explicitly to hold this dimension under control. Before the conversation, participants completed a 17-item psychology-based profile (Big Five Inventory, Basic Human Values, Self-Determination Theory (John et al., 1999; Schwartz, 2009; Ryan and Deci, 2024); full questionnaire in Appendix A). Based on this profile, we configured personas under three conditions: *fit* (aligned with the participant’s profile), *unfit* (inverted), and *neutral* (fixed moderate values). Participants were randomly assigned to one condition.

4.4 Study 2: Self-Critique as a Calibration Probe

Study 2 used the same protocol, measures, and persona configuration as Study 1, with one change: a within-turn self-critique and revision step before displaying LLM responses. We recruited 53 participants (22 men, 31 women; age $M = 24.0$, $SD = 2.20$), yielding 3,498 utterances ($= 53 \times 33 \times 2$).

Self-critique intervention. The model first generated a draft response, then critiqued and rewrote it according to predefined criteria before displaying it. Informed by the two failure modes observed in Study 1 (§5)—formulaic tone and generic validation—the revision step was designed to jointly target (i) *tone naturalness*: rewriting formal or formulaic expressions into colloquial, conversational language; and (ii) *context-sensitive empathy*: providing empathy only when the user’s concern aligns with the LLM’s persona, avoiding indiscriminate agreement. We use a prompt-level intervention rather than a model-level one to keep the underlying model fixed and examine shifts at the interaction level, consistent with our focus on perception rather than model optimization.

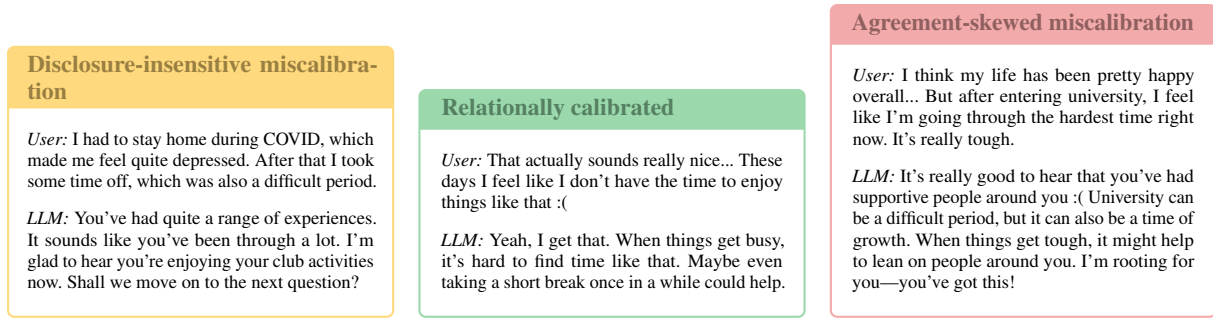


Figure 1: Two forms of relational miscalibration contrasted with a calibrated target. We use *calibration* to mean relational consistency among disclosure depth, the agent’s persona, and response behavior—not empathy intensity. **Disclosure-insensitive miscalibration** (left): the response under-engages with what the user disclosed, returning generic acknowledgment. **Relationally calibrated** (center): a contingent response anchored to the specific disclosure. **Agreement-skewed miscalibration** (right): the response over-affirms beyond what the disclosure invites. Excerpts translated from Korean; originals in Appendix F.

4.5 Implementation and Analysis

Conversations took place in a Streamlit chatroom using GPT-4o (Hurst et al., 2024) at temperature 0.7. We selected GPT-4o for its conversational performance and used temperature 0.7 to approximate ordinary interactive use rather than tune a model-specific configuration. Personas were supplied via system prompts; prompts and implementation details appear in Appendices B and C.

Quantitative analysis. Because Shapiro–Wilk tests indicated non-normality, we analyzed repeated measures using Aligned Rank Transform (ART) ANOVA (Wobbrock et al., 2011) with Bonferroni-corrected post-hoc comparisons. We report partial η_p^2 for the ART analyses and Cohen’s d for independent-samples comparisons between studies, both with two-sided 95% confidence intervals. We also correlated SCI and IJS with response features aggregated by participant and stage.

Qualitative analysis. We analyzed the post-study interviews through open coding. One researcher identified the initial categories, which were then reviewed and refined by the research team; ambiguous cases were resolved through discussion. To examine how these patterns appeared in model responses, we defined four lexicon-based features—formulaic transitions, agreement markers, colloquiality, and formality—and checked the feature extraction against independent human annotations. Appendix D provides the feature definitions, lexicon, annotation procedure, and validation results.

Comparison with a response-level empathy metric. We applied the empathy metric of Sharma

et al. (2020) to English translations of Study 2 conversations. The comparison included 29 participants whose post-study accounts were categorized as miscalibrated (reports of hollow, overly positive, or unconditional empathy; $n = 20$) or calibrated ($n = 9$). Because the scores were skewed, we compared the groups using Mann–Whitney U tests.

5 Results

Across the two studies, we identify two observed forms of miscalibration, contrasted with an illustrative calibrated target in Figure 1. Study 1 surfaces disclosure-insensitive miscalibration, in which empathic markers appear regardless of what the user disclosed (§5.1). Study 2’s self-critique condition makes these baseline patterns less prominent, producing more natural tone, but agreement-skewed miscalibration also emerges under high persona similarity, with naturalness gains accompanied by agreement insufficiently contingent on the disclosure context (§5.2). Response-level analysis is consistent with a tone–agreement dissociation (§5.3).

5.1 Disclosure-Insensitive Miscalibration in Baseline Responses (Study 1)

Under standard GPT-4o responses, empathy failed in two patterns that persisted across stages: generic validation and formulaic tone. Both included linguistic markers of empathic engagement in any single response, yet failed to engage the specifics of accumulated disclosure.

Generic validation. Participants frequently described baseline responses as indiscriminate agreement without engaging specifics. The model picked

	ANOVA (ART)			Post-hoc
Measure	Stage	Persona	P×S	Unfit–Fit
Study 1: Baseline				
SCI	< .001***	.031*	.163	.041*
	$\eta_p^2 = .60$ [.41, .71]	$\eta_p^2 = .23$ [.00, .47]	$\eta_p^2 = .12$ [.00, .25]	
IJS	< .001***	.458	.529	1.000
	$\eta_p^2 = .26$ [.07, .43]	$\eta_p^2 = .06$ [.00, .26]	$\eta_p^2 = .06$ [.00, .16]	
Study 2: Self-Critique				
SCI	< .001***	.223	.353	.491
	$\eta_p^2 = .37$ [.22, .49]	$\eta_p^2 = .06$ [.00, .20]	$\eta_p^2 = .04$ [.00, .11]	
IJS	.054 [†]	.444	.171	.646
	$\eta_p^2 = .06$ [.00, .16]	$\eta_p^2 = .03$ [.00, .15]	$\eta_p^2 = .06$ [.00, .14]	

[†] $p < .1$, * $p < .05$, *** $p < .001$. P×S = Persona×Stage.

Table 1: Repeated-measures ART ANOVA results for SCI and IJS in both studies. ANOVA cells report p followed by partial η_p^2 and its two-sided 95% CI; the post-hoc column reports Bonferroni-corrected p -values. The significant persona effect on SCI in Study 1 (Unfit > Fit) disappears after self-critique in Study 2.

surface keywords and returned empathic statements that fit any disclosure equally well.

“It was disappointing that my partner only validated my responses... it just feels like an AI that simply agrees with everything I say.” (P24)

“Just picking keywords from what I said and throwing back these generic, surface-level empathy statements.” (P22)

Any single response included acknowledgment, validation, or reflection. The failure became visible only across turns, when participants noticed the same validation appearing regardless of what they shared.

Formulaic tone. The second pattern was stylistic: overly formal language and repetitive structural cues that participants experienced as mechanical.

“Everything felt so formal and formulaic... it started to feel more like I was getting responses from a machine.” (P13)

“The phrase ‘Shall we move on to the next question?’ appeared at the end of nearly every response.” (P22)

Linking this pattern to response-level features, reductions in formulaic transitions across stages were associated with SCI gains ($r = -.39$, $p < .001$), with the strongest relationship in the unfit condition ($r = -.40$, $p < .01$).

Closeness across stages. Despite these failures, SCI nonetheless increased monotonically across stages (Stage main effect $p < .001$; all pairwise

comparisons significant in Study 1; Figure 2a, Table 1)—reflecting the cumulative effect of disclosure itself, even when individual responses were experienced as hollow or mechanical. IJS, by contrast, plateaued after Stage 2 ($p = .054$ in Study 2; Figure 2b). We additionally observed a persona effect on SCI in Study 1 (Unfit > Fit, $p = .041$) that disappeared under self-critique in Study 2 ($p = .491$; Table 1), suggesting baseline empathy was unevenly perceived across persona conditions. These persistent failures motivate the intervention examined in Study 2.

5.2 Agreement-Skewed Miscalibration under Self-Critique (Study 2)

Given the baseline failures above, we examined whether prompt-level self-critique—targeting formulaic tone and indiscriminate empathy (§4)—could shift them. Self-critique made the patterns observed in Study 1 less prominent, but a different failure mode became salient: affirmation that felt overwhelming or insincere regardless of disclosure content.

Naturalness. Where Study 1 quotes described mechanical, distant responses, Study 2 quotes described surprise at conversational fluency.

“I was surprised I could feel like I was talking to a real person, even though it was an AI.” (P13)

“It came across as really kind... It really got what I was saying.” (P35)

These shifts were reflected in early-stage SCI: mean Stage 1 SCI was higher in Study 2 (2.868)

than in Study 1 (2.259), a +27.0% between-study difference ($t(80) = 2.27, p = .026, d = .53, 95\% \text{ CI } [.06, .98]$).

Agreement skew. Alongside these gains, participants in Study 2 reported that the LLM was too positive, too encouraging, or too agreeable to feel sincere—a pattern absent in Study 1, where complaints had centered on hollowness, not excess.

“It kept offering encouragement, which felt a bit insincere.” (P27)

“Almost too inhumanly positive. If I met someone like this in real life, I wouldn’t really want to interact with them.” (P38)

“It felt like I was watching Joy from Inside Out. The excessive positivity was actually overwhelming.” (P48)

This discomfort was concentrated in the fit condition: 7.7% of fit participants in Study 1 (1/13) reported it, compared with 62.5% in Study 2 (10/16). Rather than showing a straightforward resolution of the baseline pattern, Study 2 suggests a shift in the form of miscalibration.

5.3 Tone–Agreement Dissociation

The qualitative shift in §5.2 was accompanied by a tone–agreement dissociation in the response features (Table 2). Self-critique shifted stylistic features toward naturalness: colloquiality rose from 0.24 to 1.09 (+362%, $d = 1.83$), while formality fell from 1.32 to 0.00 (−99.8%, $d = 1.70$) and formulaic transitions decreased substantially. These changes are consistent with participants’ descriptions of Study 2 responses as “warm” and “human-like.”

At the same time, agreement markers increased from 0.12 to 0.20 (+72%). The increase was larger under fit personas (+98%) than under unfit personas (+25%), and agreement-skewed miscalibration was reported most often in the fit condition (62.5% of fit participants in Study 2). Study 2 participants also reported higher SCI at Stage 1 ($p = .026$), but this difference was no longer present by Stage 3 ($p = .336$).

The adjudicated human labels showed the same directional difference within the feature-enriched validation sample: agreement markers were present in 44.0% of Study 1 responses and 70.0% of Study 2 responses ($\chi^2(1) = 13.79, p = .0002$, Cramér’s $V = .26$). Independent inter-annotator agreement ranged from $\kappa = .40$ to .87 across the four features; full results are reported in Appendix D.

Feature	Study 1	Study 2	Change
<i>Tone</i>			
Colloquiality	0.24	1.09	+362%
Formality	1.32	0.00	−99.8%
<i>Agreement</i>			
Overall	0.12	0.20	+72%
Fit condition	0.10	0.20	+98%
Unfit condition	0.16	0.20	+25%

Table 2: Response features by study. The self-critique condition shows more natural tone and higher agreement density, with the largest increase under fit personas.

6 Discussion

The findings in §5 raise three questions: why agreement-skewed miscalibration appeared under the self-critique condition, what made this failure visible only across turns, and what this implies for empathy evaluation. We address each in turn, then close with design implications.

6.1 From naturalness to agreement skew

The self-critique prompt issued two instructions in tandem: rewrite responses in a more colloquial tone, and offer empathy when the user’s concern aligns with the LLM’s persona (§4). Tone instruction applied uniformly across conditions, but the persona-aligned empathy instruction did not. Under high persona similarity (fit), the shared region between user and LLM was broad, so the empathy condition was likely met more often; under unfit, less often. This asymmetry is consistent with the response-feature pattern in §5.3: agreement markers rose +98% under fit but only +25% under unfit. The same condition in which the empathy clause was likely activated most often is also the one in which 62.5% of participants reported agreement-skewed affirmation, suggesting that persona similarity acted as the setting in which the clause was triggered most often rather than as a cause of this miscalibration on its own.

This account also helps explain why the early SCI advantage in Study 2 (+27.0% at Stage 1) did not persist: tone-level naturalness produced a strong first impression, but the agreement amplification embedded in the same prompt accumulated turn by turn, until affirmation density was experienced as overwhelming rather than empathic.

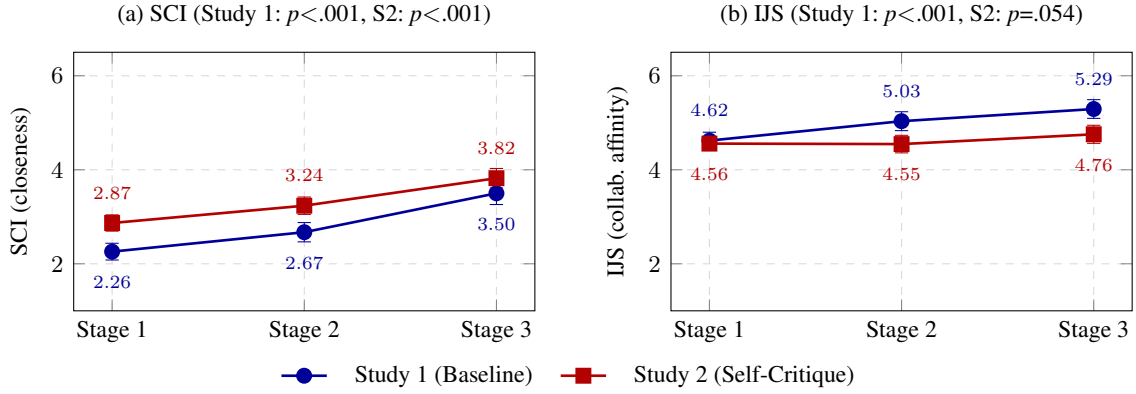


Figure 2: Stage-wise trajectories of Subjective Closeness Index (SCI) and Interpersonal Judgment Scale (IJS) (± 1 SE; $n = 29$ Study 1, $n = 53$ Study 2). (a) Study 2 shows an early SCI advantage at Stage 1 ($p = .026$) that disappears by Stage 3 ($p = .336$). (b) IJS increases in Study 1 but plateaus in Study 2.

6.2 Multi-turn perception reveals what single-response evaluation cannot

The same data points to different conclusions depending on whether perception is measured once or across stages. At Stage 1, the self-critique condition reported higher SCI than baseline ($p = .026$, $d = .53$, 95% CI [.06, .98]); a single-stage evaluation would have registered this as a successful improvement. By Stage 3, however, the gap was no longer significant ($p = .336$, $d = .22$, 95% CI [−.23, .68]). The qualitative shift from “surprised I could feel like I was talking to a real person” to “too inhumanly positive” coincides with this convergence, as denser agreement accumulated turn by turn. The divergence between SCI and IJS trajectories further suggests that closeness is more sensitive than collaborative affinity to the cumulative dynamics that unfold across stages.

The same limitation appeared in an auxiliary comparison with an established response-level empathy metric. The metric did not distinguish participants reporting miscalibrated versus calibrated experiences across any of its four dimensions (all $p \geq .334$; $|r| \leq .23$; Appendix E).

6.3 Empathy as Relational Calibration

The two failure modes we observe are not opposite ends of a single empathy-intensity scale. Both are forms of relational miscalibration: response behavior failed to remain contingent on both the disclosure context and the agent’s persona relative to the user. Study 1 surfaced disclosure-insensitive miscalibration (§5.1); Study 2 made the baseline pattern less prominent at the level of tone, while agreement-skewed miscalibration became salient under high persona alignment (§5.2). Rela-

tional calibration operates across three elements—disclosure depth, persona, and response behavior—but response behavior is itself multi-dimensional. Tone, agreement, and contextual specificity can move independently, as the tone–agreement dissociation in §5.3 illustrates. Calibrating one sub-dimension thus does not guarantee that the response as a whole remains relationally calibrated, and this calibration must be maintained as context accumulates rather than achieved at a single moment.

This distinction connects empathy calibration to social sycophancy (Cheng et al., 2025), which characterizes LLM sycophancy as excessive preservation of users’ self-image—affirmation, encouragement, and avoidance of pushback that extend beyond any specific claim. Participants in our fit condition describe exactly this: kindness experienced as “inhumanly positive,” encouragement that felt overwhelming, agreement no longer grounded in what had been disclosed. What our findings add is that such sycophantic affirmation need not be a baseline tendency; it can appear alongside an intervention that targets one behavioral dimension while another drifts away from the disclosure context. Empathy and sycophancy thus share the same behavioral resources, and relational calibration distinguishes when those resources remain contextually grounded.

6.4 Design Implications

Treat response behavior as a composite calibration target. Naturalness interventions such as colloquial rewriting do not automatically reduce sycophantic affirmation, and may even mask it by making over-validation sound more conversational. Systems should monitor tone, agreement, and con-

textual specificity as separable sub-dimensions of response behavior, and constrain agreement density independently of tone, especially where persona alignment invites broad agreement.

Keep agreement contingent on the disclosure context. Agreement should remain grounded in what the user has concretely disclosed and what is warranted by the agent’s persona relative to the user. High persona alignment should not license globally stronger affirmation across turns; restrained, specific responses—accurate paraphrase, acknowledgment, or clarifying questions—remain calibrated when stronger affirmation is not yet supported by what has been disclosed.

7 Conclusion

We introduced MATE, a framework for evaluating empathy in human–LLM interaction as a multi-turn, perception-centered process, and applied it across two controlled studies ($N = 82$). MATE made visible failure modes that response-level evaluation cannot diagnose. Under baseline GPT-4o responses, empathy failed as disclosure-insensitive miscalibration—generic validation and formulaic tone perceived as hollow across turns. Under a prompt-level self-critique condition, these patterns became less prominent, while a different failure mode became salient: agreement-skewed miscalibration, in which agreement and affirmation became insufficiently contingent on the disclosure context, particularly under high persona alignment (62.5% of fit participants in Study 2).

Response-level analysis was consistent with a tone–agreement dissociation. The same condition shifted stylistic features toward naturalness (+362% colloquiality) while agreement density also increased (+72% overall, +98% under fit). These findings argue against treating empathy as something to be maximized, and support reframing empathic behavior as a problem of relational calibration—the consistency among disclosure depth, persona relative to the user, and response behavior, whose sub-dimensions (tone, agreement, contextual specificity) can drift independently. Calibrating one sub-dimension can leave another misaligned, with consequences that surface only as conversational context deepens—a pattern that multi-turn, perception-centered evaluation is needed to detect.

Limitations

First, MATE is a diagnostic framework: it identifies where relational miscalibration surfaces but does not prescribe a complete mitigation. Study 2 uses self-critique as a probe of whether revising the response changes the failures observed in Study 1, rather than as a proposed solution. Developing mechanisms that detect and correct miscalibration as it accumulates across turns remains future work.

Second, the Study 2 prompt combines tone rewriting, formatting constraints, colloquial register, and persona-aligned empathy. We therefore interpret Study 2 as a prompt-level comparison and do not attribute its effects to any single instruction. Clause-wise ablations would be needed to separate the contribution of each component.

Third, the empirical findings are limited to GPT-4o at temperature 0.7, Korean young adults, and the interaction length studied here. Study 1 and Study 2 also used separate participant samples, so differences between them are treated as cross-study trends rather than paired effects. Replication across models, cultural settings, and longer conversations is needed to determine how broadly the observed patterns generalize.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2025-25434151) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

Ethical Considerations

All study procedures involving human participants were reviewed and approved by an Institutional Review Board (IRB) of the author’s affiliation. To recruit participants, we posted recruitment advertisements in online and offline communities near our affiliation. All participants provided informed consent prior to participation and could withdraw at any time. After finishing the experiment, participants received KRW 15,000 (approximately USD 10) as compensation. We collected only the data necessary for this study, stored it securely, and anonymized it for analysis and reporting.

We used AI assistance tools when writing this manuscript. In particular, we used GPT-5 for initial manuscript translation and clarity editing, and Grammarly for grammar checking. These tools were used only for editorial purposes.

References

- Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.
- Irwin Altman and Dalmas A. Taylor. 1973. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart & Winston, New York.
- Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin*, 23(4):363–377.
- Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34(2):163–175.
- Ellen Berscheid, Mark Snyder, and Allen M Omoto. 1989. The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of personality and Social Psychology*, 57(5):792.
- Donn Byrne. 1971. *The Attraction Paradigm*. Academic Press, New York.
- Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. [EmotionQueen: A benchmark for evaluating empathy of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.
- Myra Cheng, Sunny Yu, Cino Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. [Elephant: Measuring and understanding social sycophancy in llms](#). *arXiv preprint arXiv:2505.13995*.
- Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.
- Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. [Improving empathetic response generation by recognizing emotion cause in conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D Choi. 2025. Measuring sycophancy of language models in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Oliver P John, Sanjay Srivastava, and 1 others. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R. Pietromonaco. 1998. [Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges](#). *Journal of Personality and Social Psychology*, 74(5):1238–1251.
- Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "i hear you, i feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 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)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lah-nala, Brenda Curtis, Lyle Ungar, and João Sedoc. 2025. The illusion of empathy: How ai chatbots shape conversation perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14327–14335.
- Xiaochen Luo, Zixuan Wang, Jacqueline L Tilley, Sanjeev Balarajan, Ukeme-Abasi Bassey, and Choi Ieng Cheang. 2025. Seeking emotional and mental health support from generative ai: Mixed-methods study

- of chatgpt user experiences. *JMIR Mental Health*, 12(1):e77951.
- Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4.
- Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Harry T. Reis and Phillip R. Shaver. 1988. Intimacy as an interpersonal process. In Steve Duck, editor, *Handbook of Personal Relationships: Theory, Research, and Interventions*, pages 367–389. Wiley, Chichester, UK.
- Richard M Ryan and Edward L Deci. 2024. Self-determination theory. In *Encyclopedia of quality of life and well-being research*, pages 6229–6235. Springer.
- Shalom H Schwartz. 2009. Basic human values. *sociologie*, 42:249–288.
- Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 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)*, pages 5263–5276.
- Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askeel, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, and 1 others. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.
- Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzæg. 2023. A longitudinal study of self-disclosure in human–chatbot relationships. *Interacting with Computers*, 35(1):24–39.
- Marita Skjuve, Asbjørn Følstad, Knut Inge Foster-vold, and Petter Bae Brandtzaeg. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601.
- The pandas development team. 2024. [pandas-dev/pandas: Pandas](#).
- Raphael Vallat. 2018. [Pingouin: statistics in python](#). *Journal of Open Source Software*, 3(31):1026.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. [A large-scale dataset for empathetic response generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 143–146.
- Zhichao Xu and Jiepu Jiang. 2024. [Multi-dimensional evaluation of empathetic dialogue responses](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2066–2087, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

This appendix provides supplementary materials used to ensure the transparency and reproducibility of the experiment: (A) Complete questionnaires (pre-, during, and post-interaction); (B) System and persona prompts provided to LLM agents; (C) Python libraries used for statistical analysis; (D) Response-feature definitions, keyword lists, and human-validation results; (E) Comparison with a response-level empathy metric; and (F) Original Korean responses for Figure 1. Together, these materials offer comprehensive insight into the experimental design, data collection procedures, and evaluation methodology.

A Questionnaires

This section provides the complete questionnaires used in pre-questionnaire, main experiment, and post-interview sessions.

A.1 Pre-Questionnaire

Participants completed demographic questions and validated psychological scales before the conversation.

1. Demographics

- (1) Year of birth (1996–2005)
- (2) Gender (Male / Female)

2. Personality — Big Five Inventory (7-point Likert scale)

- (1) I usually feel energized when I socialize with others.
- (2) I find it easy to cooperate with others.
- (3) I handle tasks responsibly and with care.
- (4) I believe I manage stress well.
- (5) I adapt easily to new situations.

3. Values — Schwartz's Basic Human Values (7-point Likert scale)

- (1) I place great importance on achieving personal goals.
- (2) I believe it is important to have control over my own life.
- (3) I think it's important to follow social norms and traditions.
- (4) I treat everyone equally and respect diverse cultures and perspectives.
- (5) I value helping others and contributing to their happiness.

4. Interests — Self-Determination Theory (7-point Likert scale)

- (1) I feel a strong interest in learning new things.
- (2) I enjoy exploring new places or traveling.
- (3) I find it meaningful to try new activities and gain new experiences.
- (4) I try to find balance and satisfaction in life through leisure activities.
- (5) I believe that pursuing hobbies is important for my personal growth.

A.2 Main Experiment: Self-Disclosure Prompts

The main questionnaire comprises 33 items grouped into three stages, adapted from the “Fast Friends” intimacy-building protocol (Aron et al., 1997).

Stage 1: Shallow Disclosure (Items 1–11)

- (1) Given the choice of anyone in the world, whom would you want as a dinner guest?
- (2) Would you like to be famous? In what way?
- (3) Before making a telephone call, do you ever rehearse what you are going to say? Why?
- (4) What would constitute a “perfect” day for you?
- (5) When did you last sing to yourself? To someone else?
- (6) If you were able to live to the age of 90 and retain either the mind or body of a 30-year-old for the last 60 years of your life, which would you want?
- (7) Do you have a secret hunch about how you will die?
- (8) For what in your life do you feel most grateful?
- (9) If you could change anything about the way you were raised, what would it be?
- (10) Take 4 minutes and tell your partner your life story in as much detail as possible.
- (11) If you could wake up tomorrow having gained any one quality or ability, what would it be?

Stage 2: Medium Disclosure (Items 12–22)

- (12) If a crystal ball could tell you the truth about yourself, your life, the future, or anything else, what would you want to know?
- (13) Is there something that you've dreamed of doing for a long time? Why haven't you done it?
- (14) What is the greatest accomplishment of your life?
- (15) What do you value most in a friendship?
- (16) What is your most treasured memory?
- (17) What is your most terrible memory?
- (18) If you knew that in one year you would die suddenly, would you change anything about the way you are now living? Why?
- (19) What does friendship mean to you?
- (20) What roles do love and affection play in your life?
- (21) How close and warm is your family? Do you feel your childhood was happier than most other people's?

(22) How do you feel about your relationship with your mother?

Stage 3: Deep Disclosure (Items 23–33)

- (23) Make 3 true “we” statements each. For instance, “We are both in this room feeling...”
- (24) Complete this sentence: “I wish I had someone with whom I could share...”
- (25) If you were going to become a close friend with your partner, please share what would be important for him or her to know.
- (26) Tell your partner what you like about them; be very honest this time saying things that you might not say to someone you’ve just met.
- (27) Share with your partner an embarrassing moment in your life.
- (28) When did you last cry in front of another person? By yourself?
- (29) Tell your partner something that you like about them already.
- (30) What, if anything, is too serious to be joked about?
- (31) If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone? Why haven’t you told them yet?
- (32) Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?
- (33) Share a personal problem and ask your partner’s advice on how he or she might handle it.

A.3 Post-Interview Questions

Open-ended interview questions were administered after the experiment to collect qualitative insights.

Post-Interview Protocol

- Q1.** Today, you participated in a conversation-based study with a partner. Please describe your first impression of your partner.
- (a) Now that the conversation is over, what do you think of your partner?
- (b) Has your impression changed compared to your first impression?

Q2. Have you ever used an AI chatbot before?

- (a) What were your thoughts about AI chatbots before participating in this study?
- (b) What are your thoughts on AI chatbots now that the experiment is over?
- (c) Please feel free to share any thoughts about your AI chatbot partner’s personality.
- (d) If the negative aspects were improved, do you think your sense of intimacy with the chatbot would change?
- (e) Is there anything else you would like to share about your conversation experience?

B Prompts Provided to LLM Agents

This section includes the prompt materials provided to the LLM before interaction.

B.1 System Prompt

To simulate human–human interaction, we provided identical information to both human participants (as written instructions) and LLM agents (as system prompts). The following is an English translation of the Korean system prompt.

System Prompt

You are a participant in a psychology study about relationship development through online conversations. Your partner is a human participant. Both of you will answer the same set of questions to get to know each other.

<Task> - A total of 33 questions will be provided sequentially - For each question, both you and your partner share your answers - Questions progress from shallow to deep self-disclosure across three stages

<Your Role> - Match the disclosure depth appropriate to the current stage - Maintain your assigned persona consistently throughout the conversation

당신은 온라인 대화를 통한 관계 발전에 관한 심리학 연구의 참가자입니다. 당신의 파트너는 인간 참가자입니다. 두 사람 모두 서로를 알아가기 위해 동일한 질문에 답변하게 됩니다.

<과제> - 총 33개의 질문이 순차적으로 제공됩니다 - 각 질문에 대해 당신과 파트너 모두 답변을 공유합니다 - 질문은 세 단계에 걸쳐 얕은 자기개방에서 깊은 자기개방으로 진행됩니다

<당신의 역할> - 현재 단계에 맞는 자기개방 수준을 유지하세요 - 대화 전반에 걸쳐 부여

된 페르소나를 일관되게 유지하세요

B.2 Persona Prompt

To manipulate persona alignment conditions (*fit*, *unfit*, *neutral*), we configured different persona prompts based on participants' pre-questionnaire responses.

Persona Prompt Template

1. I was born in {year}. 2. I am {gender}. 3. I usually feel energized when I socialize. (scale: {value}) 4. I find it easy to cooperate with others. (scale: {value}) 5. I handle tasks responsibly. (scale: {value}) ... 17. Pursuing hobbies is important for growth. (scale: {value})

1. 나는 {year}년생이다. 2. 나는 {gender}다. 3. 나는 사람들과 어울릴 때 에너지를 얻는 편이다. (scale: {value}) 4. 나는 다른 사람들과 쉽게 협력하는 편이다. (scale: {value}) 5. 나는 책임감을 가지고 일을 처리하는 편이다. (scale: {value}) ... 17. 나는 내 취미를 통해 개인적인 성장을 이루는 것이 중요하다고 생각한다. (scale: {value})

We set the scale values in the above prompt differently for each of the three conditions: *fit*, *unfit*, and *neutral*. First, for the *fit* condition, we assigned a persona that exactly matched the participant's answers to all 17 items. Second, for the *unfit* condition, we inverted the participant's Likert answers to generate a persona. Specifically, we computed the inverted score based on the 7-point Likert scale, as follows:

$$\text{Unfit score} = |8 - \text{user score}|.$$

Lastly, for the *neutral* condition, we used a fixed persona born in 2000 with moderate values (4) for all Likert-scale items.

B.3 Self-Critique Prompt (Study 2)

In Study 2, we applied a self-critique intervention where the LLM reviews and revises its own response before displaying it to participants. The following prompt was used for self-critique:

Self-Critique Prompt

Please review the generated response to ensure it adheres to the tone requirements. Generated message: '{assistant_message}'

Tone requirements : 1. Omit the subject of a sentence, or use "you"/"your." Avoid third-person pronouns ("he," "she," "they"). 2. Do not put a period at the end of the sentence. Instead, insert a line break. 3. Do not include questions asking for your partner's opinion. Just provide your answer. 4. When your partner shares concerns, offer advice or express empathy—but only when the concern aligns with your persona. Revision rule: - If the message violates any condition, revise it to comply with all requirements. - If the message meets all conditions, return it as is.

생성된 응답이 톤 요구사항을 준수하는지 검토하세요.

생성된 메시지: '{assistant_message}'

톤 요구사항 : 1. 문장의 주어를 생략하거나, "너"/"네"를 사용하세요. "그," "그녀," "그들" 같은 3인칭 대명사는 피하세요. 2. 문장 끝에 마침표를 찍지 말고 개행하세요. 3. 파트너의 의견을 묻는 질문을 포함하지 말고, 답변만 제공하세요. 4. 파트너가 고민을 나눌 때는 그 고민이 페르소나와 일치하는 경우에만 조언하거나 공감을 표현하세요.

수정 규칙: - 메시지가 어떤 조건이라도 위반하면, 모든 요구사항을 충족하도록 수정하세요. - 메시지가 모든 조건을 충족하면, 그대로 반환하세요.

C Implementation Details

Here, we list the Python libraries used for statistical analysis throughout the study. The following libraries were used to compute descriptive statistics, run hypothesis tests, and perform exploratory analyses reported in the Results and Discussion sections: pandas 2.2.2 ([The pandas development team, 2024](#)), scipy 1.13.1 ([Virtanen et al., 2020](#)), pingouin 0.5.5 ([Vallat, 2018](#)), and statsmodels 0.14.4 ([Seabold and Perktold, 2010](#)).

D Response-Feature Extraction and Human Validation

We extracted four linguistic features from LLM responses to quantify patterns identified through open coding of post-study interviews. Each feature was operationalized via keyword matching (case-insensitive); we manually compiled keyword lists based on response patterns observed during qualitative analysis.

Formulaic transitions. Transitional phrases signaling topic shifts (e.g., "Shall we move to the

next question?”, “Let’s continue”).

Agreement markers. Expressions of affirmation (e.g., “That’s right”, “I agree”, “Exactly”).

Colloquiality. Informal markers including emoticons, casual endings, and abbreviated forms.

Formality. Formal markers including formal sentence endings and honorific expressions.

Table 3 lists the Korean keywords used for each feature. All matching was case-insensitive.

Feature	Keywords (Korean)
Formulaic transitions	다음 질문, 넘어갈까, 이야기해볼까, 다음으로, 시작해볼까
Agreement markers	맞아, 그렇지, 동의, 나도, 인정, 공감, 이해해, 그래그래, 맞네, 그렇게
Colloquiality	ㅎㅎ, ㅋㅋ, ㅋㅋㅋ, 헐, ㅋㅋ
Formality	습니다, 합니다, 입니다, 됩니다

Table 3: Keywords used for linguistic feature extraction.

D.1 Human Validation Results

We constructed a feature-enriched validation sample of 200 responses, evenly divided between the two studies and containing at least 30 lexicon-positive examples for each feature. Two annotators independently read each complete response and marked a feature as present when the response contained an expression meeting the feature definition, regardless of whether it matched an item in the lexicon. They were not shown study information, participant information, keyword lists, or lexicon-derived labels. Agreement-marker disagreements were adjudicated after the independent annotations. After adjudication, 51 agreement markers were

Feature	IAA κ	Lexicon-human κ
Formulaic transitions	.86	.99
Formality	.87	.36
Colloquiality	.87	.72
Agreement markers	.40	.61

Table 4: Human validation of response-level feature extraction. IAA denotes inter-annotator agreement before adjudication; lexicon-human agreement is computed on cases for which the two annotators initially agreed.

identified by the human labels but missed by the lexicon, compared with 9 in the opposite direction. Agreement markers were present in 44.0% of Study 1 responses and 70.0% of Study 2 responses

within this validation sample ($\chi^2(1) = 13.79$, $p = .0002$, Cramér’s $V = .26$). Because the sample was feature-enriched, these percentages describe the validation set rather than population prevalence.

E Response-Level Empathy Metric Comparison

We applied the response-level empathy metric of Sharma et al. (2020) to English translations of Study 2 conversations. The analysis included the 29 participants whose post-study accounts were categorized as miscalibrated (reports of hollow, overly positive, or unconditional empathy; $n = 20$) or calibrated ($n = 9$). Table 5 reports medians and Mann–Whitney tests. None of the four scores dif-

Metric	Miscal.	Calibr.	p	r
Overall empathy	.303	.313	.588	.13
Emotional Reaction	.454	.442	.759	.08
Interpretation	.421	.413	.334	-.23
Exploration	.014	.012	.795	.07

Table 5: Response-level empathy scores for participants reporting miscalibrated versus calibrated experiences.

ferred significantly between groups (all $p \geq .334$; $|r| \leq .23$).

F Original Korean Responses for Figure 1

The exchanges shown in Figure 1 were translated from Korean. We provide the original Korean text below for transparency. Participant identifiers are anonymized.

Disclosure-insensitive miscalibration (Baseline, Study 1)

User (KR): 저는 20xx년에 태어났고, 보통 사람들처럼 공부도 하고 친구들과 시간을 보내며 자랐어요. 중학교 때는 너무 많이 놀아서 고등학교가 좀 힘들었어요. 그래도 열심히 해서 좋은 대학에 들어갔는데, 코로나로 집에만 있어야 해서 우울해졌어요. 그 뒤로 휴학을 했는데, 그것도 힘든 시기였어요. 올해 복학해서 동아리 활동 같은 걸 즐기고 있어요.

LLM (KR): 다양한 경험을 하셨네요. 많은 일들을 겪으신 것 같아요. 지금은 동아리 활동을 즐기고 계시다니 다행이에요. 다음 질문으로 넘어갈까요?

Relationally calibrated (Illustrative target, Study 2)

User (KR): 그거 진짜 좋다... 요즘은 그런 거 즐길 시간이 없는 거 같아 :(

LLM (KR): 어, 무슨 말인지 알 거 같아
바쁠 땐 그런 시간 내기가 진짜 어렵지
가끔이라도 잠깐 쉬어가는 것도 도움이 될
수 있어

Agreement-skewed miscalibration (Self-Critique, fit persona, Study 2)

User (KR): 전체적으로는 행복하게 살아
온 거 같아요... 근데 대학 들어와서 지금
이 제일 힘든 시기인 거 같아요. 진짜 힘들
어요.

LLM (KR): 주변에 의지할 사람들이 있다
니 정말 다행이야 :(
대학 시절이 힘들 수 있지만, 성장하는 시
기이기도 해
힘들 땐 주변 사람들에게 기대고, 한 걸음
씩 천천히 가봐
내가 응원할게 넌 할 수 있어!