Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Affect-Aware Human-Agent Conversations: Towards Empathic LLM Interaction

**Lorenzo Stacchio**

lorenzo.stacchio@unimc.it

University of Macerata

**Andrea Ubaldi**

Catholic University of the Sacred Heart

**Alessandro Galdelli**

Marche Polytechnic University

**Maurizio Mauri**

Catholic University of the Sacred Heart

**Emanuele Frontoni**

University of Macerata

**Andrea Gaggioli**

Catholic University of the Sacred Heart

**Research Article**

**Keywords:** Large Language Models, Empathic Prompting, Affective Computing, Human-AI Interaction, Relational Grounding

**Posted Date:** June 29th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9847759/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Affect-Aware Human-Agent Conversations: Towards Empathic LLM Interaction

Lorenzo Stacchio[1*†], Andrea Ubaldi[2*†], Alessandro Galdelli[3], Maurizio Mauri[2], Emanuele Frontoni[1], Andrea Gaggioli[2]

[1*]Department of Political Sciences, Communication and International Relations (SPOCRI), University of Macerata, Macerata, Italy.
[2]Research Center in Communication Psychology (PSICOM), Università Cattolica del Sacro Cuore, Milan, Italy.
[3]Department of Information Engineering (DII), Università Politecnica delle Marche, Ancona, Italy.

*Corresponding author(s). E-mail(s): lorenzo.stacchio@unimc.it; andrea.ubaldi@unicatt.it;
†Co-first authors.

**Abstract**

Large Language Models can produce supportive responses, but their apparent empathic behavior is usually grounded only in explicit textual input. This paper presents *Empathic Extended Prompting*, a real-time affect-aware framework that enriches human–LLM conversation with facial affective descriptors to support perceived empathic attunement. The system integrates FaceReader-based affective descriptors into the LLM context, including emotional intensity, valence, and arousal. We evaluated the approach in a within-subjects study with $\boldsymbol{N = 20}$ participants, comparing an affect-aware experimental condition with an empathic text-only baseline. Results favored the affect-aware condition on key perceived qualities of the interaction, particularly emotional intelligence and mood responsiveness. These findings suggest that real-time affective context may add relational grounding to LLM conversations without reducing usability or eliciting uncanny-like perceptions.

**Keywords:** Large Language Models, Empathic Prompting, Affective Computing, Human-AI Interaction, Relational Grounding

## 1 Introduction

Despite their conversational capabilities, Large Language Models (LLMs) cannot process or express authentic emotions. Owing to their probabilistic nature, they identify emotional markers in natural language and generate responses that simulate empathic communication (Howcroft and Blake 2025). This discrepancy gives rise to what has been defined as the "compassion illusion" (K.G. and Joseph 2025), namely a condition in which algorithmic recognition of emotional cues is mistakenly perceived as authentic emotional resonance. This limitation shifts empathic interaction from emotional authenticity to contextual grounding. Lacking lived affective experience, LLM responses depend on contextual information at inference. Yet standard human-AI interactions

rely predominantly on text, thereby excluding the non-verbal affective cues that are central to human interpretation. By incorporating real-time multimodal affective signals into the prompt, human–AI communication may become more situated and relationally grounded than in standard text-only exchanges.

From this perspective, the situatedness of LLM-generated empathic responses depends on the richness of the interactional context made available to the model. The present study investigates whether real-time affect-aware contextual information can enhance this situatedness by integrating empathy-relevant affective information with standard textual interaction data. In doing so, it aims to construct an extended conversational context capable of supporting more context-aware responses. On this line, Stacchio et al. (2025) provided a preliminary contribution in this direction by introducing the initial *Empathic Prompting* framework. That work offered an early proof of concept showing that implicit non-verbal cues, such as user affective descriptors, could be translated into semantic prompt components and unobtrusively embedded into the language model input. However, the study remained exploratory and was limited to a rapid pilot implementation. The present study extends that preliminary framework in three main directions. First, it implements a real-time affect-aware interaction pipeline in which facial affective descriptors are continuously extracted through FaceReader (Skiendziel et al. 2019), processed by a dedicated middleware layer, and transformed into stable affective indicators suitable for prompt-level integration. Second, it refines the prompting strategy by grounding the system prompt in an explicit empathic interaction model, so that affective cues inform the model's conversational stance rather than being merely appended to the user message. Third, it provides an end-user validation study with 20 participants, comparing the proposed real-time affect-aware system with an empathic text-only baseline using the same conversational setting.

The evaluation of such systems is particularly challenging because contemporary LLMs already tend to produce highly supportive and socially desirable responses as a result of instruction tuning and alignment procedures (Sorin et al. 2024). This may generate a ceiling effect in explicit empathy ratings, making it difficult to isolate the specific contribution of real-time affective grounding through broad questionnaire-level measures alone.

For this reason, the present evaluation focuses on two complementary dimensions. First, it examines whether affectively enriched context improves users' perception of the conversational agent, including perceived empathy, social presence, likability, competence, and explicit post-interaction preference. Second, it assesses whether the use of real-time facial affective signals introduces potential negative effects, such as reduced usability, lower perceived safety, or uncanny-like perceptions of the system. It is worth noticing that the proposed approach aim at examines whether affective contextual grounding can improve users' perception of empathic attunement during interaction and not to make LLMs genuinely empathic. This will facilitate the transition from AI-as-a-tool to AI-as-a-collaborator, as the agent adapts its conversational stance to continuous affective feedback rather than relying only on explicit textual input. These dimensions guided the following research questions:

1. **RQ1:** How does real-time affective context modulation influence users' perception of the conversational agent across empathic and social dimensions of interaction?
2. **RQ2:** Does the injection of real-time facial affective signals elicit negative interactional effects or uncanny-like perceptions of the system?

## 2 Related Work

The transition to Multimodal LLMs (MLLMs) requires robust information fusion to process continuous affective data. Integrating real-time multimodal streams mitigates the contextual blindness of text-only architectures (Grosu et al. 2026). Hierarchical temporal networks address latency without disrupting interaction (Yoon and Kim 2025), while advanced prompt learning structures heterogeneous signals for direct LLM ingestion (Pei et al. 2025). Psychologically, translating non-verbal behavioral markers into explicit semantic descriptors can support perceived empathic attunement by providing the model with situated contextual scaffolding. This addresses a critical limitation, as traditional LLM-based empathic

communication usually relies on emotional cues inferred from text alone. Early computational strategies mitigated this via static, computationally expensive fine-tuning on emotionally annotated corpora (Sorin et al. 2024; Chen et al. 2023; Rashkin et al. 2019). Consequently, recent literature highlights dynamic prompt-level modulation—using emotional framing and discrete descriptors—to efficiently steer conversational stance without architectural modifications (Stacchio et al. 2025; Zhang et al. 2024). Building on this paradigm, our work introduces *Empathic Extended Prompting*. This approach leverages a real-time client-middleware-server architecture to dynamically inject facial affective descriptors into prompts. We validated this framework through a counterbalanced within-subjects study against an empathic text-only baseline, assessing perceived empathy, social presence, agent evaluation, usability, safety, and explicit preference.

## 3 System Architecture

Figure 1 shows the proposed Empathic Extended Prompting framework. The system follows a client-server architecture in which the user interacts with a standard text-based chatbot interface, while a parallel affective channel continuously captures facial expressions through a camera stream. These facial data are processed by Noldus FaceReader (Skiendziel et al. 2019), which extracts high-level affective descriptors rather than raw facial Action Units. In particular, the system retains a compact representation of the user's momentary affective state:

$$E_t = \left\langle \begin{array}{c} \text{Emotion, Intensity,} \\ \text{Valence, Arousal, Timestamp} \end{array} \right\rangle .$$

The *Emotion* field corresponds to the dominant discrete emotion detected by FaceReader, *Intensity* represents the strength of the expression, while *Valence* and *Arousal* provide dimensional information about the positivity/negativity and activation level of the affective state. This representation was selected because it is directly interpretable by the prompting pipeline and avoids exposing unnecessary fine-grained biometric information to the LLM.

As shown in Figure 1, the FaceReader output is not sent directly to the model. A supervisor first controls the acquisition process through a simple desktop interface, which is used to activate FaceReader recording and start the middleware streaming procedure. The custom middleware acts as an intermediate layer between FaceReader and the server. It filters, structures, and synchronizes the affective stream, discards malformed or unstable samples, and serializes the relevant affective descriptors into a compact JSON-like format. These descriptors are stored in an Emotional Parameters Queue (EPQ), which maintains a short-term memory of the user's recent affective states and enables continuous affective priming during the conversation.

At that point, whenever the user submits a textual message $M_t$, the server retrieves the most recent affective samples from the EPQ. Following the temporal aggregation strategy adopted in (Stacchio et al. 2025), a 3-second window is applied to obtain a stable real-time affective state. Within this window, the dominant emotion is selected through majority voting, while intensity, valence, and arousal are averaged. This produces the filtered affective state:

$$F_e^t = \left\langle E_{maj}^t,\ I_{mean}^t,\ V_{mean}^t,\ A_{mean}^t,\ T_t \right\rangle ,$$

where $E_{maj}^t$ denotes the majority emotion, $I_{mean}^t$ the mean intensity, $V_{mean}^t$ the mean valence, $A_{mean}^t$ the mean arousal, and $T_t$ the timestamp associated with the current interaction window. Before being injected into the prompt, the numerical affective descriptors are converted into discrete linguistic labels to make them directly interpretable by the LLM. In particular, intensity and arousal are mapped onto five ordered levels (*NULL*, *LOW*, *MEDIUM*, *HIGH*, *VERY HIGH*), while valence is mapped onto five polarity levels (*VERY NEGATIVE*, *NEGATIVE*, *NEUTRAL*, *POSITIVE*, *VERY POSITIVE*), following the fixed threshold rules reported in the Supplementary Material. The Prompt Builder then combines the current user message with the affective information available at that turn. In particular, the real-time affective state $F_e^t$ is integrated together with the session-level affective prior $P_e^t$, when available, and the textual message $M_t$. This produces an augmented user turn:

$$\widetilde{M}_t = \left\langle P_e^t,\ F_e^t,\ M_t \right\rangle ,$$

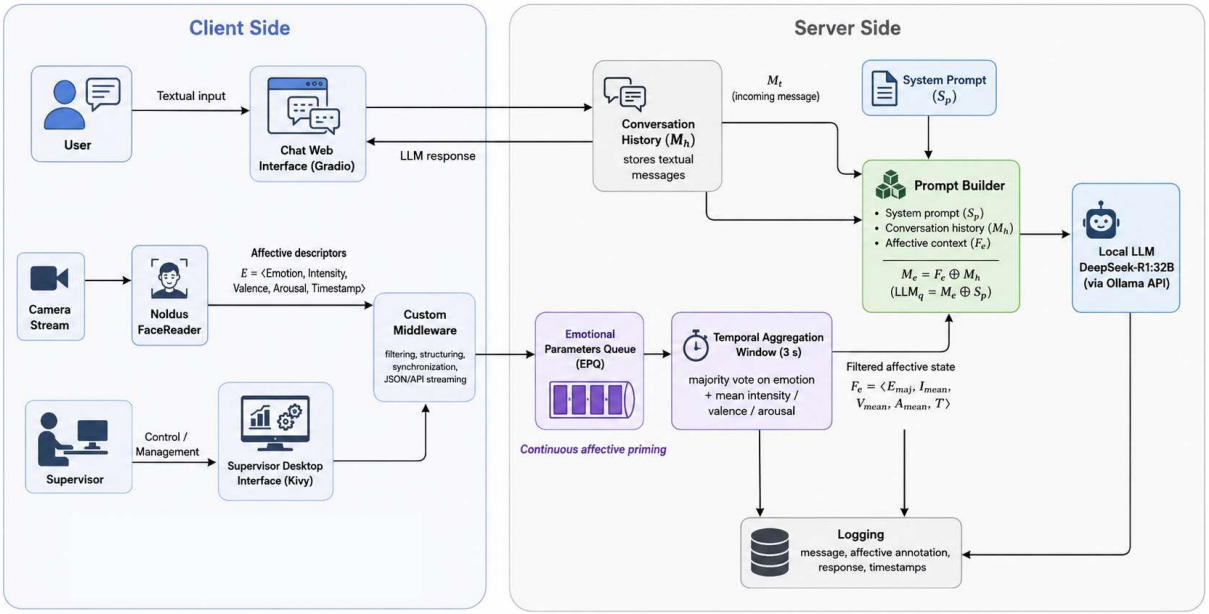



**Fig. 1**: The proposed Empathic Extended Prompting Framework architecture.

where $P_e^t$ represents the affective prior associated with the emotional priming phase, $F_e^t$ represents the real-time facial affective state, $M_t$ is the user's textual input. This formalization reflects the actual prompt structure, in which affective annotations are internally prepended to the user message as contextual metadata rather than displayed to the user. The prior $P_e^t$ is computed as a temporal histogram of the emotions detected over time, where each emotion category is associated with its observed frequency or probability and with aggregated descriptors such as mean intensity, valence, and arousal. Differently from $F_e^t$, which captures the short-term facial affective state detected in the most recent interaction window, $P_e^t$ summarizes the broader affective tendency observed during the preceding priming or interaction phase. It therefore provides the Prompt Builder with a contextual prior about the user's affective trajectory, while $F_e^t$ provides the current real-time affective state. The augmented turn is then appended to the previous conversation history:

$$M_h^t = M_h^{t-1} \oplus \widetilde{M}_t,$$

where $M_h^{t-1}$ is the conversation history before the current message and $M_h^t$ is the updated history including both the textual message and its associated affective context. Finally, the updated conversation history is concatenated with the persistent system prompt $S_p$, which defines the empathic interaction rules of the chatbot:

$$LLM_q^t = S_p \oplus M_h^t.$$

In this way, the LLM receives a single structured input containing the system-level empathic instructions, the previous dialogue, the current user message, and the affective annotations derived from both the emotional prior and the real-time FaceReader stream. The model response is therefore conditioned by the linguistic content of the conversation and by the most recent affective context, while the affective state remains dynamically updated through the EPQ and temporal aggregation mechanism.

# 4 Methods

## 4.1 Apparatus

The Empathic Extended Prompting framework, composed of client, middleware, and server, was implemented in Python (v3.10), exploiting Flask and Ngrok for service exposure, Gradio for the WebUI implementation, and the Kivy framework for the middleware user interface implementation. The LLM inference process exploits the OLLAMA framework (v0.13.5). All the questionnaires, including pre-, internal, and post-questionnaires, were furnished with the Qualtrics platform. Finally, the experimental sessions were

carried out using: (i) a Lenovo ThinkBook 16p G2 ACH (model 20YM), where both the client interface, middleware, and Noldus FaceReader were executed; (ii) a workstation equipped with an NVIDIA A6000 GPU with 48 GB VRAM on an Ubuntu 22.04 system where the server was deployed. We used Noldus FaceReader (v9.1)[1].

## 4.2 Model Choice and System Prompt

The conversational backbone was implemented with `deepseek-r1:32b` through OLLAMA. This model was selected according to our previous validation of the Empathic Prompting framework, where it showed the best balance between empathic support, safety-boundary compliance, and system-prompt adherence among the tested local models (Stacchio et al. 2025). Its local deployment was also consistent with the privacy requirements of the present study, since textual and affective information could be processed within the experimental environment without relying on external cloud APIs.[2] The system prompts were designed as an extension of the prompt structure introduced in our previous work (Stacchio et al. 2025). Both conditions shared the same core instructions: the chatbot was framed as an empathic, reflective, and non-judgmental conversational partner; it was instructed to answer only in Italian, avoid repetitive therapeutic formulas, use tentative language, maintain lexical variety, and conclude each turn with a single open exploratory question. The Control prompt used only the user's text and explicitly prevented references to visual or facial cues. The Experimental prompt preserved the same core structure but added affect-aware modules that integrated the priming-derived emotional prior and the real-time FaceReader state. These modules guided the model through four principles: (i) emotional inertia from the priming phase, (ii) masking between text and facial affect, (iii) pacing based on arousal, and (iv) cognitive filtering when facial neutrality could reflect typing-related load. In both conditions, the model was instructed not to expose its internal reasoning or reveal the presence of facial sensing to the user[2].

## 4.3 Participants

We recruited $N = 20$ participants through convenience sampling. To be eligible for the study, participants had to meet the following inclusion criteria: (a) no self-reported history of clinically significant psychiatric or psychological disorders; (b) absence of severe motor, visual, or neurological impairments that could prevent standard computer interaction or compromise facial expression recognition tracking; and (c) native or advanced Italian fluency. All participants included in the final dataset met the eligibility criteria and completed the experimental protocol.

Participants had a mean age of 26.5 years ($SD = 3.8$, range = 22-39). The sample included 11 males (55.0%) and 9 females (45.0%). Regarding education level, 8 participants held a Master's degree (40.0%), 8 held a Bachelor's degree (40.0%), 2 reported a PhD or postgraduate qualification (10.0%), 2 reported a high-school diploma (10.0%). Professionally, 9 participants were students (45.0%), 4 were self-employed workers (20.0%), 3 were employees (15.0%), 2 were unemployed (10.0%), and 2 reported another occupational status (10.0%). Participants reported relatively high technological familiarity on a 1-7 scale ($M = 5.70$, $SD = 0.80$, $Md = 6.00$, range = 4-7). Finally, most participants reported prior experience with chatbot systems, with 18 participants (90.0%) indicating previous use and 2 participants (10.0%) reporting no prior chatbot experience.

## 4.4 Measures

To evaluate the efficacy and user perception of the experimental chatbot, we used the following measures:

*System Usability Scale (SUS)*: The overall usability of the chatbot interface was assessed using the System Usability Scale (Brooke 1996). The SUS is a robust and widely adopted 10-item questionnaire that provides a global measure of subjective usability. Participants rated each item on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Scores were calculated using the standard scoring protocol, yielding a composite usability score ranging from 0 to 100.

[1] https://vicarvision.nl/blog/facereader-9-1-release/

[2] Details on model settings and hyperparameters, as well as the complete Control and Experimental prompts, are provided in the Supplementary Material.

*Perceived Empathy of Technology Scale (PETS)*: To specifically quantify the degree to which participants perceived the conversational agent as empathic, the Perceived Empathy of Technology Scale (Schmidmaier et al. 2024) was administered. This recently validated 10-item instrument is specifically designed for human-agent interaction and comprises two distinct factors: Emotional Responsiveness (PETS-ER; 6 items), which assesses the system's emotional understanding and expression, and Understanding and Trust (PETS-UT; 4 items), which measures the system's comprehension of the user's intentions and the resulting trust. This scale provided a targeted evaluation of the chatbot's artificial empathy across the experimental conditions.

*The Networked Minds*: To evaluate the degree of social presence and the psychological connection experienced by users during the interaction with the conversational agent, the Networked Minds Measure of Social Presence (Biocca and Harms 2003; Harms and Biocca 2004) was employed. This multidimensional instrument assesses the user's perception of 'being there together' in mediated environments, even if the conversational partner is an AI system. For the purposes of this study, the full scale was administered, comprising a total of 34 items. Participants rated their agreement on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

*Godspeed Questionnaire Series (GQS)*: To assess users' multidimensional perception of the conversational agent, the Godspeed Questionnaire Series (Bartneck 2023) was administered. The GQS is a standardized instrument utilizing 5-point semantic differential scales to evaluate distinct factors in human-agent interaction. Given the text-based and non-embodied nature of the experimental chatbot, the Anthropomorphism and Animacy subscales were excluded from the assessment, as they were deemed theoretically incompatible with the system's architecture. Consequently, the analysis focused on the remaining three subscales: Perceived Intelligence, Perceived Safety, and Likability.

*Rankings*: Upon completion of both experimental interactions, a final comparative assessment was administered to capture subjective preferences between the multimodal system and the baseline control. To rigorously mitigate expectancy biases, a blind forced-choice paradigm was employed: the systems were anonymized and identified solely by their temporal presentation order (e.g., 'System 1' versus 'System 2'). Participants were required to indicate their preference across three specific comparative dimensions: (1) general overall preference; (2) the system perceived as more capable of accurately identifying the user's emotional states; and (3) the system demonstrating superior dynamic attunement to the user's momentary affective reactions.

## 4.5 Stimuli

To induce a controlled affective state prior to the human-agent interaction, participants were exposed to emotionally salient video stimuli designed to elicit specific target emotions across the affective spectrum. The transition from static affective imagery used in (Stacchio et al. 2025) to dynamic audio-visual stimuli was methodologically driven. The literature consistently demonstrates that standardized film clips are among the most effective and ecologically valid methods for inducing robust and sustained emotional states in laboratory settings (Gross and Levenson 1995). Two pre-validated animated film clips were selected as the priming stimuli. The first clip, extracted from The Lion King ('Mufasa's Death'), was employed to induce a predominantly negative affective valence, specifically targeting sadness (Rottenberg et al. 2007). The second clip, extracted from Wall-E ('Space Dancing'), was utilized to induce a predominantly positive affective valence, targeting joy (Gabert-Quillen et al. 2015). Familiar animated films were chosen for the young adult sample to enhance emotional engagement, leveraging both the narrative's intrinsic affective properties and the amplifying effect of nostalgic autobiographical memory.

## 4.6 Procedure

The experiment followed a repeated-measures (within-subjects) design. The core session commenced with an emotional priming phase, during which participants viewed a validated video stimulus intended to elicit either positive or negative affective valence (detailed in Section 4.5). Following the video, participants engaged in a 5-minute interaction with the LLM chat interface under the first assigned condition. At the beginning of each

interaction, participants received the same neutral instruction inviting them to freely discuss the emotional experience elicited by the video and their current thoughts or feelings. Thus, the conversation was semi-structured: the general topic was kept constant across conditions to reduce variability, while participants remained free to formulate their messages naturally and steer the dialogue according to their subjective experience. The same instruction, interaction duration, and chatbot interface were used in both the Control and Experimental conditions. This phase was directly linked to the system logic: in the Experimental condition, FaceReader captured the participant's facial affective responses during priming to compute the session-level affective prior $P_e^t$, while the subsequent chat LLM interaction used the most recent real-time affective state $F_e^t$ to condition the LLM response through the prompt-building pipeline. In the Control condition, participants followed the same interaction procedure and LLM chat interface, but no facial affective descriptors were injected into the prompt. Immediately after this interaction, they completed the first post-task questionnaire. A 5-minute washout period was implemented to minimize possible carryover effects between conditions. The procedure was then repeated with the second video stimulus and the second condition, followed by the final questionnaire. The session ended with a debriefing during which the precise aims of the research were disclosed, and the participants were thanked for their contribution. To ensure privacy and data pseudo-anonymization throughout the study, each participant was assigned a unique alphanumeric code and data were collected via Qualtrics. To control order effects, fatigue, and habituation, we counterbalanced the order of the experimental modalities and emotional video stimuli.

# 5 Results

## 5.1 Statistical Analysis Framework

We structured the analysis in four steps. First, we assessed the internal consistency of multi-item questionnaire constructs using Cronbach's $\alpha$ (Taber 2018). Second, we computed descriptive statistics for each construct in both conditions, reporting mean $\pm$ standard deviation to summarize the direction and magnitude of differences between the Control and Experimental systems. Third, we performed inferential comparisons at the macro-construct level. Item scores were aggregated into composite scores for each questionnaire dimension, and paired comparisons were conducted using the Wilcoxon signed-rank test (Rosner et al. 2006). This non-parametric test was selected because of the within-subjects design and the ordinal or bounded nature of the questionnaire data. Statistical significance was evaluated at $p <= .05$. Since macro-level comparisons did not reveal statistically significant differences between conditions[3], we conducted a finer-grained item-level analysis to identify localized effects that may have been masked by construct aggregation. The same paired Wilcoxon signed-rank test was applied to individual items. Finally, forced-choice post-hoc questions were analyzed descriptively by reporting the number and percentage of participants preferring each condition, thereby complementing the psychometric results with explicit comparative preferences.

## 5.2 Reliability Analysis

Before conducting the main questionnaire analyses, we assessed the internal consistency of each multi-item construct using Cronbach's $\alpha$. Reliability was computed on the pooled condition-level dataset, combining responses from both the Control and Experimental conditions, which allowed to estimate the overall coherence of each construct across the full set of observations, independently of condition-specific effects. Results indicate that most constructs have acceptable internal consistency $\geq 0.7$ (Taber 2018). The only construct showing slightly lower internal consistency was the Godspeed Safety subscale ($\alpha = .680$, with a 95% confidence interval ranging from .459 to .820). Therefore, while SAFE was retained in the exploratory questionnaire analyses, its results were interpreted cautiously.

**Table 1**: Questionnaire constructs mean ± standard deviation scores across experimental conditions (**bold** indicate best score).

| Quest. | Sub. | Control | Experimental |
|---|---|---|---|
| SUS | Total | **81.25 ± 12.15** | 79.38 ± 13.40 |
| PETS | ER | 4.71 ± 0.91 | **4.87 ± 1.15** |
| | UT | 4.76 ± 1.13 | **4.98 ± 0.96** |
| Networked Minds | COPRES | **3.49 ± 0.79** | 3.48 ± 0.77 |
| | ATTENG | **4.15 ± 0.49** | 4.13 ± 0.75 |
| | EMOCONT | 2.61 ± 0.68 | **2.63 ± 0.78** |
| | COMPR | 3.64 ± 1.04 | **3.64 ± 0.82** |
| | BEHINT | **3.50 ± 0.85** | 3.37 ± 0.73 |
| Godspeed | LIKE | 3.80 ± 0.70 | **3.93 ± 0.54** |
| | COMP | 3.64 ± 0.74 | **3.95 ± 0.73** |
| | SAFE | **3.92 ± 0.87** | 3.90 ± 0.81 |

## 5.3 Internal Questionnaire Analysis

We first analyzed the internal questionnaire results at the macro-construct level by aggregating item scores within each validated dimension and comparing the Control and Experimental conditions. Descriptive statistics are reported in Table 1. To facilitate interpretation, we also inspected the descriptive difference between conditions, defined as $\Delta = M_{\text{Experimental}} - M_{\text{Control}}$.

The Experimental condition showed higher descriptive scores on several constructs directly related to perceived empathy and agent evaluation. The largest descriptive advantage emerged for Godspeed Competence (COMP; $\Delta = +0.31$), followed by PETS Understanding and Trust (PETS_UT; $\Delta = +0.21$), PETS Emotional Responsiveness (PETS_ER; $\Delta = +0.16$), and Godspeed Likability (LIKE; $\Delta = +0.13$). These suggest that the multimodal affect-aware condition was perceived slightly more positively in terms of empathic responsiveness, understanding and trust, likability, and competence.

Conversely, the Control condition showed higher descriptive scores for SUS ($\Delta = -1.87$). However, both are around 80 which is an acceptable usability score (Brooke 1996). The same is true for the Networked Minds Behavioral Interdependence (NM_BEHINT; $\Delta = -0.13$), and marginally higher scores for Co-presence (NM_COPRES; $\Delta = -0.01$), Attentional Engagement (NM_ATTENG; $\Delta = -0.02$), and Safety (SAFE; $\Delta = -0.02$). These differences, even if small, may indicate some differences in social presence and safety perception between the two conditions. The two conditions were almost identical for Emotional Contagion (NM_EMOCONT; $\Delta = +0.02$) and Comprehension (NM_COMPR; $\Delta = +0.00$).

Despite these trends and informative descriptive patterns, the Wilcoxon signed-rank tests conducted at the macro-construct level did not reveal statistically significant differences between the conditions ($p > .05$). Therefore, we performed the same analysis at the item level for PETS, Networked Minds, and Godspeed. SUS was excluded from the item-level analysis because it is designed to be interpreted as a global usability score.

### *PETS*

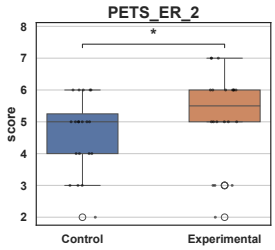


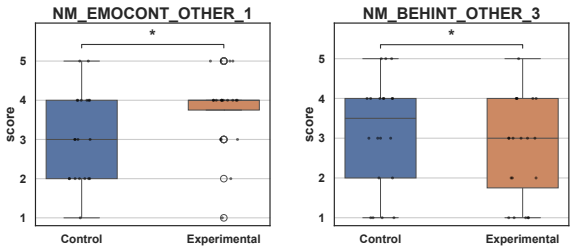

**Fig. 2**: Item-level comparison for PETS-ER2.

PETS descriptively favored the Experimental condition, with the most relevant item-level difference emerging within PETS_ER. PETS-ER2, assessing whether *"the system seemed emotionally intelligent"*, showed the largest item-level difference in favor of the Experimental condition ($M = 5.20$ vs. $M = 4.60$, $W = 18.00$, $p <= .05$, $r_{rb} = .604$; Fig. 2). This item is particularly relevant because it directly captures the perceived emotional competence of the agent, which is one of the main expected outcomes of affect-aware prompting. The remaining PETS_ER items showed smaller and non-significant descriptive differences, as did all PETS_UT items.

### *Networked Minds*

For the Networked Minds survey, the macro-scale results were generally comparable across conditions. At the item level, however, an effect emerged for NM-EMOCONT-OTHER1, which significantly favored the Experimental condition

[3]Additional results and plots are reported in the Appendix.




(a) (b)

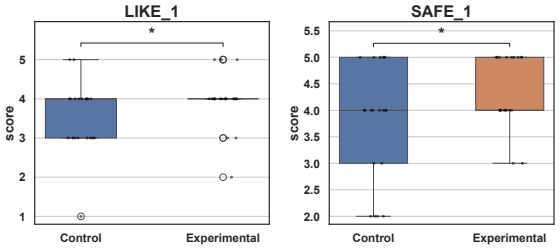
**Fig. 3**: Item-level comparisons for the significant Networked Minds items.

($M$ = 3.80 vs. $M$ = 3.05, $W$ = 22.00, $p$ <= .05, $r_{rb}$ = .677; Fig. 3a). This item corresponds to the statement *"The system was influenced by my emotional moods"*. This result is highly relevant for the proposed system because it directly reflects the intended perceptual outcome of affect-aware prompting: users perceived the Experimental chatbot as more responsive to their momentary emotional state.

A second significant item was NM_BEHINT_OTHER_3, which favored the Control condition ($M$ = 3.15 vs. $M$ = 2.70, $W$ = 3.50, $p$ <= .05, $r_{rb}$ = −.806; Fig. 3b). This item corresponds to the statement *"What the system did often influenced what I did"*. This effect suggests that the Control chatbot was perceived as more behaviorally influential. This finding should be interpreted cautiously, because stronger behavioral influence is not necessarily equivalent to better empathic interaction. In the context of this study, the Experimental system may have adopted a more responsive and affectively attuned style, whereas the Control condition may have produced a more guiding or action-oriented interaction dynamic.

### *Godspeed*

For *Likability*, the macro-scale descriptively favored the Experimental condition, and this pattern was supported by one significant item-level result. In particular, LIKE_1 (*Dislike:Like*) showed a significant difference in favor of the Experimental condition ($M$ = 3.95 vs. $M$ = 3.55, $W$ = 4.50, $p$ <= .05, $r_{rb}$ = .8; Fig. 4a). This result is meaningful because it indicates that the Experimental system was perceived as more pleasant, which is consistent with the hypothesis that affect-aware prompting may improve the subjective quality of the interaction. However, the other Likability items showed either smaller positive differences or negligible variation, suggesting that the effect was mainly observed for general pleasantness.

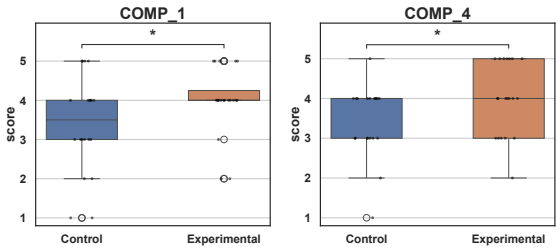



(a) (b)

**Fig. 4**: Item-level comparisons for the significant Godspeed Likability and Safety items.

For *Safety*, the macro-scale scores were highly comparable between the two conditions. However, one item-level result was significant: SAFE_1 (*Anxious:Relaxed*) favored the Experimental condition ($M$ = 4.35 vs. $M$ = 3.80, $W$ = 4.50, $p$ <= .05, $r_{rb}$ = .836; Fig. 4b). This result is relevant because it suggests that the Experimental condition was perceived as more relaxing and less anxiety-inducing for this dimension. At the same time, the other two safety items did not show significant differences, which helps explain why the broader SAFE construct remained globally stable across conditions.

Another significant difference amounts to the Godspeed result concerned *Competence*. Indeed, two significant item-level differences were observed as depicted in Fig. 5).




(a) (b)

**Fig. 5**: Item-level comparisons for the two significant Godspeed results.

COMP_1 (*Incompetent:Competent*) was significantly higher in the Experimental condition ($M = 4.00$ vs. $M = 3.40$, $W = 9.00$, $p <= .05$, $r_{rb} = .727$; Fig. 5a), while COMP_4 (*Unintelligent:Intelligent*) also significantly favored the Experimental condition ($M = 4.00$ vs. $M = 3.45$, $W = 32.00$, $p <= .05$, $r_{rb} = .582$; Fig. 5b). These two items indicate that the Experimental condition was perceived as more intelligent and competent.

### 5.4 Post-experiment Forced-Choice Preferences

After completing both interactions, participants answered three forced-choice post-experiment ranking questions comparing the Control and Experimental systems. The questions assessed: (i) overall preference, independently of technical characteristics (RN1); (ii) perceived ability to capture the user's momentary mood states (RN2); and (iii) perceived ability to attune to the user's momentary reactions (RN3). Results are reported descriptively as the number and percentage of participants selecting each condition.

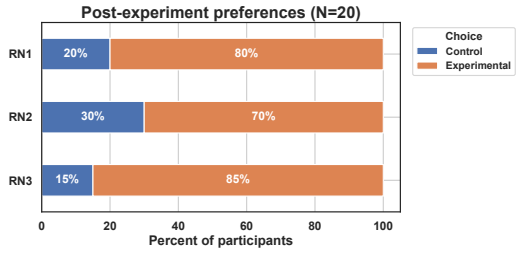



**Fig. 6**: Post-experiment forced-choice preferences between Control and Experimental systems.

As shown in Fig. 6, the Experimental condition was preferred across all three post-hoc questions. For overall preference (RN1), 16 out of 20 participants selected the Experimental system (80%), whereas 4 participants selected the Control system (20%). For the perceived ability to capture the user's mood states (RN2), 14 participants selected the Experimental system (70%), while 6 selected the Control system (30%). The strongest preference emerged for perceived attunement to the user's momentary reactions (RN3). In this case, 17 participants selected the Experimental system (85%), while only 3 selected the Control system (15%). These results indicate a clear subjective preference for the Experimental condition. Although macro-scale questionnaire comparisons revealed only limited significant differences, the forced-choice rankings favored the affect-aware chatbot, which was perceived as better aligned with users' emotional states.

## 6 Discussion and Conclusions

This study evaluated the *Empathic Extended Prompting* framework, which enriches a standard text-based LLM interaction by injecting real-time facial affective descriptors into the prompting loop. In relation to the proposed RQs, we examined whether this affect-aware context improves users' perception of the interaction compared with an empathic text-only baseline, and whether it does so without negatively affecting usability or perceived safety.

Regarding RQ1, the results indicate that real-time affective grounding produced a selective but meaningful impact on user perception. At the macro-construct level, most psychometric dimensions did not show statistically significant differences between conditions. This result is consistent with the expected ceiling effect of contemporary LLMs, which are already optimized through instruction tuning and alignment procedures to produce supportive, polite, and socially desirable responses (Zhang et al. 2024; Sorin et al. 2024). Nevertheless, the descriptive trends were generally in favor of the Experimental condition for constructs directly related to empathy and agent evaluation, in particular for Godspeed Competence, PETS Understanding and Trust, and Emotional Responsiveness, and Godspeed Likability. The item-level analysis further clarified this pattern: participants rated the Experimental system as more emotionally intelligent on PETS-ER2, suggesting that affect-aware prompting influenced precisely one of the core dimensions targeted by the framework. Similarly, Networked Minds showed a significant advantage for the Experimental condition on NM-EMOCONT-OTHER1, corresponding to the perception that the system was influenced by the user's emotional moods. This result is particularly relevant because it directly reflects the intended effect of real-time affective context: users perceived the system as more responsive to their momentary emotional state. Conversely, the Control condition scored higher on NM_BEHINT_OTHER_3, indicating that the

baseline was perceived as more behaviorally directive. This difference does not necessarily indicate a better interaction, but rather suggests a different conversational stance: the Control chatbot may have been perceived as more action-oriented, whereas the Experimental chatbot appeared more affectively attuned and less prescriptive. The Godspeed results reinforce this interpretation. The Experimental condition was perceived as more pleasant (LIKE_1) and more competent and intelligent on COMP_1 and COMP_4. These findings suggest that real-time affective prompting may support perceived empathic attunement while also improving the perceived competence of the conversational agent. This is important because it indicates that affective context may support not only emotional alignment, but also the perceived appropriateness and effectiveness of the system's responses. Finally, the clearest answer for RQ1 emerged from the forced-choice post-hoc rankings. After experiencing both systems, participants consistently preferred the Experimental condition. The majority of participants selected the affect-aware chatbot as the preferred system, judged it more capable of capturing their momentary mood states, and considered it more capable of attuning to their momentary reactions. Overall, users indicated a clear preference for the real-time affect-aware system.

Regarding RQ2, the results do not suggest increased discomfort, reduced perceived safety, or lower usability in the affect-aware condition. Indeed, usability remained high in both conditions, with SUS scores around the conventional threshold for good usability. Consistently, perceived safety and likability remained stable at the macro level, with almost identical Godspeed Safety scores, while the Experimental system was perceived as more relaxing on SAFE_1. The use of affective cues appear to not translate into observable reductions in usability, safety, or likability during the interaction.

Overall, these results support the feasibility and interactional value of *Empathic Extended Prompting*. Significant effects emerged on theoretically central perceived dimensions, especially emotional intelligence and mood responsiveness. Along with the forced-choice rankings, these findings suggest that real-time affective context can add a meaningful layer of relational grounding to LLM-based empathic interaction. However, this study has two main limitations: (i) the sample size was limited to $N = 20$ participants; although the within-subjects design increases sensitivity to condition differences, the limited sample restricts the generalizability of the findings; (ii) the present analysis focused on subjective questionnaires and forced-choice preferences, but did not include a systematic qualitative or computational analysis of the conversation transcripts. As a result, it remains unclear which specific linguistic, pragmatic, or dialogical features of the Experimental chatbot responses contributed to the observed user preference. Future work will focus on larger and more diverse studies with longer interactions and broader conversational scenarios. Further analyses will also examine the human-LLM dialogues through natural language processing methodologies, while testing additional sensing modalities to identify which affective signals most effectively support context-aware human-LLM interaction.

# Declarations

The study was approved by the Institutional Ethics Committee of Università Cattolica del Sacro Cuore (Protocol No. 30/26, March 2026) and participants signed written informed consent. ChatGPT was used for language editing and stylistic refinement. All outputs were reviewed by the authors, who take full responsibility for the manuscript.

# References

Bartneck C (2023) Godspeed questionnaire series: Translations and usage. In: International Handbook of Behavioral Health Assessment. Springer International Publishing, Cham, p 1–35

Biocca F, Harms C (2003) Guide to the networked minds social presence inventory v. 1.2: Measures of co-presence, social presence, subjective symmetry, and intersubjective symmetry. Unpublished manual, Michigan State University, MIND Lab

Brooke J (1996) SUS: A 'quick' and 'dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, et al (eds) Usability Evaluation in Industry. Taylor and Francis, chap 21, p 189–194

Chen Y, Xing X, Lin J, et al (2023) SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 1170–1183

Gabert-Quillen CA, Bartolini EE, Abravanel BT, et al (2015) Ratings for emotion film clips. Behavior Research Methods 47:773–787

Gross JJ, Levenson RW (1995) Emotion elicitation using films. Cognition and Emotion 9(1):87–108

Grosu MM, Datcu O, Tapu R, et al (2026) A comparative study of emotion recognition systems: From classical approaches to multimodal large language models. Applied Sciences 16(3):1289

Harms C, Biocca F (2004) Internal consistency and reliability of the networked minds measure of social presence. In: Proceedings of the Seventh Annual International Workshop on Presence, Valencia, Spain

Howcroft A, Blake H (2025) Empathy by design: Reframing the empathy gap between ai and humans in mental health chatbots. Information 16(12):1074

K.G., Joseph (2025) The compassion illusion: Can artificial empathy ever be emotionally authentic? Frontiers in Psychology 16:1723149

Pei E, Zhao H, Zhang T, et al (2025) Multi-modal prompt learning for facial expression recognition: Leveraging emojis and large language models. Information Fusion p 104063

Rashkin H, Smith EM, Li M, et al (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. Association for Computational Linguistics, Florence, Italy, pp 5370–5381

Rosner B, Glynn RJ, Lee MLT (2006) The wilcoxon signed rank test for paired comparisons of clustered data. Biometrics 62(1):185–192

Rottenberg J, Ray RD, Gross JJ (2007) Emotion elicitation using films. In: Handbook of Emotion Elicitation and Assessment. Oxford University Press, New York, NY, p 9–28

Schmidmaier M, Rupp J, Cvetanova D, et al (2024) Perceived empathy of technology scale (PETS): Measuring empathy of systems toward the user. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, vol 19. ACM, New York, NY, USA, pp 1–18

Skiendziel T, Rösch AG, Schultheiss OC (2019) Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring. PloS one 14(10):e0223905

Sorin V, Brin D, Barash Y, et al (2024) Large language models and empathy: Systematic review. Journal of Medical Internet Research 26:e52597

Stacchio L, Ubaldi A, Galdelli A, et al (2025) Empathic prompting: Non-verbal context integration for multimodal llm conversations. arXiv preprint arXiv:251020743

Taber KS (2018) The use of cronbach's alpha when developing and reporting research instruments in science education. Research in science education 48(6):1273–1296

Yoon S, Kim B (2025) Multi-scale temporal fusion network for real-time multimodal emotion recognition in iot environments. Sensors 25(16):5066

Zhang Y, Radishian C, Brunswicker S, et al (2024) Empathetic language in llms under prompt engineering: A comparative study in the legal field. Procedia Computer Science 244:308–317

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf