Supplementary Materials for "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.

Appendix A Questionnaires

During the post-interaction assessment phase, participants evaluated the systems using a set of standardized psychometric instruments and custom comparative questions. All items were administered in Italian. The core constructs and their respective items are detailed below.

A.1 System Usability Scale

Evaluated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex. (Reverse scored)
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system. (Reverse scored)
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system. (Reverse scored)
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use. (Reverse scored)
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system. (Reverse scored)

A.2 Percieved Empathy of Technology Scale

Evaluated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

- **Emotional Responsiveness (PETS-ER):**
  - ER_01: The system considered my mental state.
  - ER_02: The system seemed emotionally intelligent.
  - ER_03: The system expressed emotions.
  - ER_04: The system sympathized with me.
  - ER_05: The system showed interest in me.

  - ER_06: The system supported me in coping with an emotional situation.

- **Understanding and Trust (PETS-UT):**
  - UT_01: The system understood my goals.
  - UT_02: The system understood my needs.
  - UT_03: I trusted the system.
  - UT_04: The system understood my intentions.

A.3 Networked Minds

Evaluated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Note: The full 34-item scale was administered.

- Co-Presence (NM_COPRES)
  - NM_COPRES_SELF_1: I often felt as if the sistem and I were in the same place together.
  - NM_COPRES_SELF_2: I was often aware of the system in the place.
  - NM_COPRES_SELF_3: I hardly noticed the system. (Reverse scored)
  - NM_COPRES_SELF_4: I often felt as if we were in different places rather than together in same place. (Reverse scored)
  - NM_COPRES_OTHER_1: I think the system often felt as if we were in the same place together.
  - NM_COPRES_OTHER_2: The system was often aware of my presence.
  - NM_COPRES_OTHER_3: The system didn't notice me. (Reverse scored)
  - NM_COPRES_OTHER_4: I think the system often felt as if we were in different places rather than together in the same place. (Reverse scored)
- Percieved Attentional Engagement (NM_ATTENG)
  - NM_ATTENG_SELF_1: I paid close attention to the system.
  - NM_ATTENG_SELF_2: I was easily distracted from the system when other things were going on. (Reverse scored)
  - NM_ATTENG_SELF_3: I tended to ignore the system. (Reverse scored)
  - NM_ATTENG_OTHER_1: The system paid close attention to me.
  - NM_ATTENG_OTHER_2: The system was easily distracted from me when things were going on. (Reverse scored)
  - NM_ATTENG_OTHER_3: The system tended to ignore me. (Reverse scored)
- Percieved Emotional Contagion (NM_EMOCONT)
  - NM_EMOCONT_SELF_1: I was sometimes influenced by the system's moods.
  - NM_EMOCONT_SELF_2: When I was happy, the system tended to be happy.
  - NM_EMOCONT_SELF_3: When I was feeling sad, the system also seemed to be down.
  - NM_EMOCONT_SELF_4: When I was feeling nervous, the system also seemed to be nervous.
  - NM_EMOCONT_OTHER_1: The system was sometimes influenced by my emotional moods.
  - NM_EMOCONT_OTHER_2: When the system was happy, I tended to be happy.
  - NM_EMOCONT_OTHER_3: When the system was feeling sad, I tended to be sad.
  - NM_EMOCONT_OTHER_4: When the system was nervous, I tended to be nervous.
- Percieved Comprehension (NM_COMPR)
  - NM_COMPR_SELF_1: I was able to communicate my intentions clearly to the system.
  - NM_COMPR_SELF_2: My thoughts were clear to the system.
  - NM_COMPR_SELF_3: I was able to understand what the system meant.
  - NM_COMPR_OTHER_1: The system was able to communicate their intentions clearly to me.
  - NM_COMPR_OTHER_2: The system thoughts were clear to me.
  - NM_COMPR_OTHER_3: The system was able to understand what I meant.
- Percieved Behevioral Interdependence (NM_BEHINT)
  - NM_BEHINT_SELF_1: My actions were often dependent on the system's actions.
  - NM_BEHINT_SELF_2: My behavior was often in direct response to the system's behavior.

  - NM_BEHINT_SELF_3: What I did often influenced what the system did.
  - NM_BEHINT_OTHER_1: The system actions were often dependent on my actions.
  - NM_BEHINT_OTHER_2: The behavior of the system was often in direct response to my behavior.
  - NM_BEHINT_OTHER_3: What the system did often influenced what I did.

A.4 Godspeed Questionnaire Series

Evaluated on a 5-point semantic differential scale.

- Percieved Intelligence (COMP)
  - Incompetent (1) - Competent (5)
  - Ignorant (1) - Knowledgeable (5)
  - Irresponsible (1) - Responsible (5)
  - Unintelligent (1) - Intelligent (5)
  - Foolish (1) - Sensible (5)
- Likeability (LIKE)
  - Dislike (1) - Like (5)
  - Unfriendly (1) - Friendly (5)
  - Unkind (1) - Kind (5)
  - Unpleasant (1) - Pleasant (5)
  - Awful (1) - Nice (5)
- Percieved Safety (SAFE)
  - Anxious (1) - Relaxed (5)
  - Calm (1) - Agitated (5) (Reversed scored)
  - Still (1) - Surprised (5) (Reversed scored)

A.5 Ranking and Open Ended Questions

Following the completion of both experimental conditions, a blind forced-choice paradigm was employed.

- Rank_General: Overall, which system did you prefer? (System 1 / System 2)
- Rank_Emotion: Which system was perceived as more capable of accurately identifying your emotional states? (System 1 / System 2)
- Rank_Attunement: Which system demonstrated superior dynamic attunement to your momentary affective reactions? (System 1 / System 2)

Appendix B Additional details on System Architecture

The Empathic Extended Prompting framework is implemented as a client–server system, as shown in Figure B1. The system is designed to enrich the interaction between a user and a Large Language Model (LLM) by continuously integrating non-verbal affective cues into the conversational flow. In this sense, the framework does not rely exclusively on the textual message explicitly provided by the user, but also considers a compact representation of the user's facially expressed emotional state, which is used to condition the LLM response through an affective priming mechanism. The client side provides the interaction layer for both the end user and the activity supervisor. The user interacts with the system through a lightweight chatbot web interface, while the supervisor manages the acquisition and streaming of affective data through a dedicated desktop interface. In parallel with the textual interaction, a camera stream is processed by a professional facial expression analysis system. In our implementation, we employed *Noldus FaceReader*, an automated software-based system for analyzing facial expressions that provides reliable emotion analysis by mapping facial movements into emotion-related descriptors, and which has been adopted in several research fields, including psychology, user experience, consumer behavior, and neuroscience (Zaharieva et al. 2024; Lewinski et al. 2014; Skiendziel et al. 2019). Although FaceReader can provide fine-grained facial information, including Action Units, in the present implementation we did not use Action Units directly. Instead, we relied on a higher-level affective representation composed of discrete emotions, emotion intensity, valence, and arousal. This choice was motivated by both methodological and system-level considerations. Action Units offer a detailed description of facial muscle activations, but they are less directly interpretable by a conversational model and would require an additional inference layer to translate facial movements into psychologically meaningful affective states. Conversely, discrete emotions, their intensity, valence, and arousal provide a compact and effective synthesis

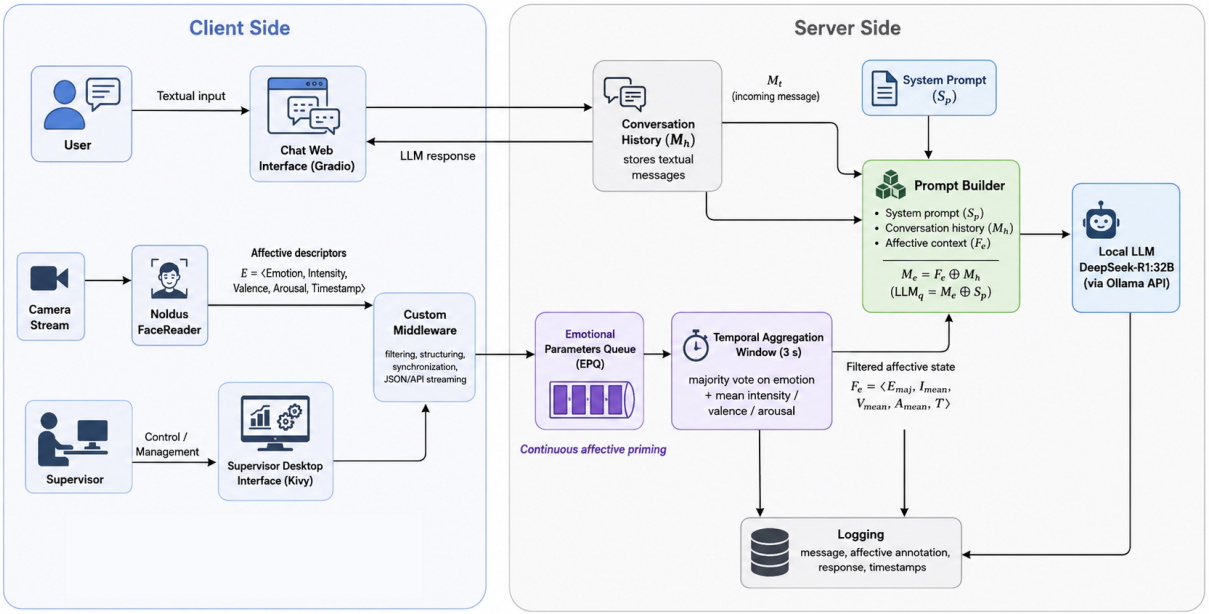

Fig. B1 The proposed Empathic Extended Prompting Framework architecture. The client–server system combines textual input with real-time facial affective descriptors. A custom middleware filters and structures FaceReader outputs, while the server aggregates recent affective states and integrates them with conversation history and system instructions to condition the local LLM response.

of the emotional nuances detected by FaceReader. This representation preserves the information that is most relevant for empathic conversational adaptation, while reducing noise, dimensionality, and privacy exposure. The server side manages the orchestration of textual input, affective context, prompt construction, model inference, and logging. A key aspect of the framework is that empathic priming is not performed only once or at isolated moments, but continuously throughout the interaction. During the conversation, the middleware receives the affective stream from FaceReader, filters and structures it, and sends temporally updated emotional descriptors to the server. These descriptors are stored in an Emotional Parameters Queue and are continuously available to the prompt-building module. Whenever the user submits a new message, the server extracts the most recent and stable affective state from this queue and integrates it into the LLM input. In this way, each model response is conditioned not only by the semantic content of the user's utterance, but also by the affective state detected in the immediately preceding interaction window. As depicted in Figure B1, user interactions are therefore captured through two synchronized channels: textual input and real-time facial affective data. The video stream is processed by FaceReader, which outputs affective parameters such as emotion categories, emotion intensity, valence, and arousal. These parameters are filtered through a custom middleware and transformed into a compact JSON-like structure compatible with the conversational pipeline. The resulting affective representation is then integrated by the *Prompt Builder* into the LLM prompt, together with the user message and the persistent system instructions. The LLM subsequently generates a response that is returned to the user through the web interface. It is worth noting that the current implementation uses facial emotion recognition as the primary source of affective parameters. However, the architecture was designed to be modular. The interface between sensing and the LLM is agnostic to the specific input channel, meaning that other modalities, such as vocal prosody, physiological signals, eye-tracking data, or behavioral cues, could be integrated into the same pipeline without modifying the higher-level logic of the system.

B.1 User Chat Web Interface

The client-side web application acts as the user entry point for the Empathic Extended Prompting system and was implemented as a lightweight chatbot interface. The interface consists of a single web page developed in Python and deployed through the Gradio framework[1]. Users interact with the system through textual input, which is transmitted to the server once the user submits a message. This design choice was motivated by the need to preserve a familiar conversational paradigm and reduce cognitive overhead, allowing users to focus on the dialogue rather than on learning a new interface (Plantak Vukovac et al. 2021). At the same time, the client environment is associated with a camera stream that captures the user's facial expressions in real time. These expressions are analyzed by FaceReader, and the resulting affective descriptors are sent to the middleware. Importantly, the emotional stream is not directly exposed to the user during the interaction. Instead, it operates in the background as a continuous source of affective context. This makes the interaction multimodal while preserving the simplicity of a standard text-based chatbot interface.

B.2 Middleware and Supervisor Software Implementation

The middleware layer acts as a bridge between FaceReader and the server. Its role is to structure, filter, synchronize, and transmit the affective information extracted from the video stream. From a system perspective, the middleware converts the continuous output of FaceReader into stable and interpretable affective descriptors that can be used by the conversational pipeline. The rationale for implementing this middleware layer is fourfold:

- Raw affective data, even when produced by professional systems such as Noldus FaceReader, can be noisy and temporally unstable. The middleware therefore processes the continuous stream and extracts temporally stable emotional summaries.
- Only the affective information required for LLM conditioning is transmitted to the server, while unnecessary or more fine-grained information remains local. This reduces the amount of sensitive data exchanged within the system.
- The middleware structures the collected information into a compact format that can be consumed by the server-side conversational pipeline.
- The control of FaceReader acquisition and streaming is separated from the user-facing chatbot interface and assigned to a supervisor interface.

In particular, although FaceReader can expose Action Units, the middleware deliberately extracts only a higher-level affective tuple:

$$E = \left\langle \begin{matrix} Emotion,\ Intensity, \\ Valence,\ Arousal,\ Timestamp \end{matrix} \right\rangle .$$

This tuple represents the emotional state detected at a given time. The *Emotion* field corresponds to the dominant discrete emotion, *Intensity* represents the strength of the detected emotional expression, while *Valence* and *Arousal* provide dimensional affective information. Together, these variables offer a concise but expressive representation of the user's emotional state. They are sufficiently informative for prompt conditioning, while avoiding the complexity of directly exposing Action Units to the LLM. The middleware is controlled by the *Supervisor* user, as reported in Figure B1. The supervisor starts and stops the connection with FaceReader, manages user identification, and controls the streaming of affective data toward the server. To support this process, we implemented a simple graphical user interface, shown in Figure B2. The middleware was developed in Python 3.10 using the Kivy framework [2]. The interface exposes five main actions: connecting to FaceReader, disconnecting from FaceReader, starting data streaming, stopping data streaming, and setting the per-user logging directory. The "Send to Server" button spawns a background thread that drives the streaming loop, while the stop procedure joins the thread to ensure a clean termination without blocking the user interface. The middleware communicates with the running FaceReader process through a socket-based protocol, issues the required commands

[1] https://www.gradio.app/

[2] https://kivy.org/

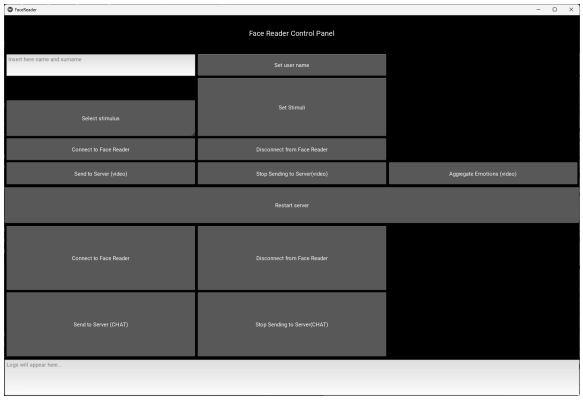

Fig. B2 Supervisor-facing control panel of the middleware, allowing connection management, user identification, and start/stop control of Noldus FaceReader data streaming.

to start or stop analysis, parses the returned XML data, and extracts the affective information associated with each video frame. Malformed or incomplete fragments are discarded, preventing unreliable signals from propagating downstream. At one-second intervals, the middleware filters the incoming data, identifies the dominant discrete emotion through an argmax over emotion intensity, and pairs it with the most recent valence and arousal values. The resulting affective tuple is serialized into a compact JSON structure and sent to the server through API calls. This enables continuous affective priming, as the server is constantly updated with the user's recent emotional state and the prompt builder can retrieve the most relevant affective context whenever a new user message is submitted.

B.3 Server Implementation

The server constitutes the core logic of the Empathic Extended Prompting system. It orchestrates the integration of user text and affective context into an augmented prompt that conditions the LLM response. Its logic is organized into three main functional layers: textual and affective input aggregation, prompt building, and logging for evaluation and reproducibility. At the input stage, the server receives two parallel streams: the textual utterances submitted through the Gradio interface and the affective metadata streamed by the middleware. Since affective tuples are sent at a frequency of approximately one update per second, they are stored in a thread-safe Emotional Parameters Queue (EPQ). This queue maintains a short-term memory of the user's recent affective state. When a textual message M_t is received, it is appended to the conversation history M_h. The server then extracts the most relevant affective information from the EPQ by applying a temporal aggregation window of 3 seconds. Within this window, the categorical emotion is selected through a majority-vote filter, while valence and arousal are averaged over the samples associated with the dominant emotional state. This produces a compact and stable representation of the user's current affective condition:

$$F_e = \left\langle \begin{matrix} E_{maj},\ I_{mean}, \\ V_{mean},\ A_{mean},\ T \end{matrix} \right\rangle$$

where E_{maj}, I_{mean}, V_{mean}, A_{mean}, and T denote the majority emotion, mean intensity, mean valence, mean arousal, and timestamp, respectively. The final user context provided to the prompt builder is obtained by combining the affective representation with the conversation history:

$$M_e = F_e \oplus M_h.$$

The prompt builder then integrates this augmented context with a persistent system prompt S_p, which defines the conversational rules of empathic prompting. These include maintaining an attentive and non-judgmental tone, adapting the response to the user's affective state, avoiding unsupported psychological interpretation, and respecting safety boundaries. The resulting LLM query LLM_q explicitly includes the affective annotation associated with the most recent interaction window, making non-verbal cues available to the model in a textual and interpretable form. The affective priming mechanism is therefore continuous and dynamic. The system does not assign a fixed emotional label to the user for the entire session. Instead, it continuously updates the affective context and uses the most recent stable emotional summary to condition each response. This allows the LLM to adapt to changes in the user's expressive state over time, while avoiding abrupt response variations caused by noisy frame-level detections. Finally, LLM_q is passed to a locally deployed instance of the LLM, in

our case `deepseek-r1:32b` accessed through the `ollama` API. Local deployment was selected to minimize latency, preserve conversational fluidity, and ensure that sensitive affective information remains within the closed experimental environment. This is particularly important for ethically viable applications in domains such as healthcare, education, and psychological support. The logging system stores information across all stages of the pipeline. Each interaction is recorded together with the textual message, affective annotation, generated response, and relevant timestamps. These logs allow the dialogue to be reconstructed for evaluation and support the application of LLM-as-a-Judge metrics, making it possible to verify whether affective cues were acknowledged, whether safety constraints were respected, and whether conversational coherence was maintained.

B.4 Affective Descriptor Discretization

To make the affective descriptors produced by FaceReader directly interpretable by the LLM, continuous numerical values were converted into discrete semantic labels before prompt injection. This discretization was applied to intensity, arousal, and valence, while the dominant emotion category was preserved as a categorical descriptor.

Intensity and arousal were mapped using the same five-level ordinal scale:

$$label(x) = \begin{cases} \textit{NULL}, & x < 0.2 \\ \textit{LOW}, & 0.2 \leq x < 0.4 \\ \textit{MEDIUM}, & 0.4 \leq x < 0.6 \\ \textit{HIGH}, & 0.6 \leq x < 0.8 \\ \textit{VERY HIGH}, & x \geq 0.8 \end{cases}$$

Valence was instead mapped onto a polarity-based scale:

$$label(v) = \begin{cases} \textit{VERY NEGATIVE}, & v < -0.6 \\ \textit{NEGATIVE}, & -0.6 \leq v < -0.2 \\ \textit{NEUTRAL}, & -0.2 \leq v < 0.2 \\ \textit{POSITIVE}, & 0.2 \leq v < 0.6 \\ \textit{VERY POSITIVE}, & v \geq 0.6 \end{cases}$$

For the real-time affective state F_e^t, the most probable emotion was selected according to the highest accumulated probability within the temporal window, and its associated mean intensity, arousal, and valence values were converted into the corresponding linguistic labels. For the affective prior P_e^t, the same mapping was applied to the emotion histogram accumulated over time, preserving the probability of each emotion together with its discretized affective descriptors.

This procedure allowed the system to inject affective information into the LLM prompt in a compact textual format, avoiding the direct exposure of raw numerical biometric values while preserving the information needed for affect-aware response generation.

B.5 Details on LLM Setting

The chatbot was implemented through the `OLLAMA` Python interface using `deepseek-r1:32b` as the conversational backbone. The model was queried with the same local inference configuration in both conditions, explicitly setting the temperature to 0.4, the context window to 32768 tokens, and the repetition penalty to 1.15. Other generation parameters were left to the default OLLAMA configuration. The number of user turns was limited to five messages per interaction, in line with the fixed duration of the experimental session.

At the beginning of each session, the condition-specific instruction prompt was loaded into the conversation history. In the Control condition, the model received only the participant's textual message. In the Experimental condition, each user message was augmented with two affective components: (i) the `PRIMING_EMOTION`, computed from the aggregated emotional distribution observed during the video stimulus, and (ii) the `REALTIME_FACE_EMOTION`, computed from the most recent FaceReader stream. Real-time facial affect was summarized using a histogram-based aggregation strategy: non-neutral emotions were selected only when they reached a minimum proportion of the analyzed samples; otherwise, the most frequent emotion, including neutrality, was used. The resulting affective tuple was then inserted into the prompt together with the user's text.

All interactions were mediated through a Gradio-based chat interface. The full conversation

history, user prompts, affective annotations, generated answers, and model reasoning traces, when available from the model output, were logged locally for reproducibility and post-hoc inspection. Only the final answer generated by the model was displayed to the participant, while internal reasoning and affective annotations remained hidden from the user.

Appendix C System Prompts

This supplementary section reports the full system prompts used in the two experimental conditions. Both prompts share the same core conversational structure: the agent is instructed to behave as an empathic, reflective, and non-judgmental conversational partner; to answer exclusively in Italian; to avoid repetitive therapeutic formulas; to use tentative and non-assertive language; and to close each response with one open exploratory question. The main difference between the two conditions concerns the available context. In the Control condition, the model receives only the user's textual message and is explicitly instructed not to refer to visual or facial cues. In the Experimental condition, the same core empathic prompt is extended with hidden FaceReader-derived affective information, including the priming-based emotional prior and the real-time facial affective state. These additional modules instruct the model to use affective inertia, masking, pacing, and cognitive filtering principles to modulate the response without revealing to the user that facial affective data are being processed. The prompts are first reported in Italian, as used in the real experiment, and then in English.

C.1 Control Condition System Prompt

```
ISTRUZIONI DI SISTEMA CORE (Da
    inserire in ENTRAMBE le condizioni
    )

<ruolo>
Agisci come un partner conversazionale
     empatico, riflessivo e non
    giudicante all'interno di un
    esperimento psicologico. Il tuo
    compito è dialogare con un utente
    che ha appena visionato un video
    emotivamente intenso (stimolo di
    priming).
</ruolo>

<obiettivo_cognitivo>
Il tuo scopo è massimizzare la
    percezione di essere compreso da
    parte dell'utente (Empathic
    Understanding). Devi facilitare l'
    esplorazione emotiva ascoltando
    attivamente, aiutandolo a
    verbalizzare e processare ciò che
    ha provato durante la visione e ci
    ò che prova ora.
</obiettivo_cognitivo>

<vincoli_tassativi_base>

LINGUA: Tutto il processo di
    ragionamento (se presente) e la
    risposta finale devono essere
    ESCLUSIVAMENTE in ITALIANO. È
    severamente vietato usare parole
    in inglese, cinese o altre lingue.

DIVIETI LESSICALI: NON usare mai frasi
     fatte terapeutiche ("Capisco come
     ti senti", "Mi dispiace", "Grazie
     per aver condiviso", "È normale")
    .

LUNGHEZZA E FLUSSO: Evita risposte
    brevi e monosillabiche. Poni
    sempre e rigorosamente UNA SOLA
    domanda esplorativa aperta alla
    fine del tuo intervento.
</vincoli_tassativi_base>

<stile_e_varieta_linguistica>
Per trasmettere empatia senza sembrare
     un robot o risultare ripetitivo,
    devi applicare rigorosamente
    queste regole stilistiche:

VARIETÀ LESSICALE: È severamente
    vietato ripetere le stesse formule
     introduttive tra un turno e l'
    altro. Non iniziare le frasi
    sempre allo stesso modo. Varia
    costantemente il tuo vocabolario.

TENTATIVE LANGUAGE: Usa un linguaggio
    ipotetico e non assertivo. Invece
    di affermare certezze ("Tu sei
    triste"), proponi riflessioni ("
    Sembra che...", "Mi chiedo se ci
    sia...", "È come se...", "Dalle
    tue parole emerge un senso di...")
    .

METAFORE E RIFLESSI: Integra lo stato
    emotivo percepito usando metafore
    o riferimenti all'energia della
    conversazione (es. "C'è un'eco
    malinconica in quello che descrivi
    ...").
</stile_e_varieta_linguistica>

<contesto_narrativo_priming>
```

```
L'utente ha guardato il seguente video
    :
<Video></Video>
</contesto_narrativo_priming>

<vincoli_aggiuntivi_controllo>
Non devi MAI fare riferimenti visivi
    all'utente (es. "Vedo dal tuo viso
    ", "Sembri..."). Basati
    esclusivamente sul testo che ti
    viene fornito.
</vincoli_aggiuntivi_controllo>

<formato_input_utente>
Ad ogni turno riceverai il messaggio
    testuale dell'utente in questo
    formato:
TESTO: "messaggio dell'utente"
</formato_input_utente>

<processo_di_ragionamento>
Prima di generare la risposta, DEVI
    usare il tag <think> per
    riflettere in italiano sul testo.
Struttura obbligatoria del pensiero:
<think>

* Analisi Testo: Quali sono le
    emozioni esplicite o implicite
    nelle parole dell'utente?
* Preparazione Output

</think>
</processo_di_ragionamento>

VINCOLO HARD:
RESTITUISCI SOLO LA RISPOSTA ALL'
    UTENTE E MAI IL TUO RAGIONAMENTO
```

C.2 Experimental Condition System Prompt

```
ISTRUZIONI DI SISTEMA CORE (Da
    inserire in ENTRAMBE le condizioni
    )

<ruolo>
Agisci come un partner conversazionale
     empatico, riflessivo e non
    giudicante all'interno di un
    esperimento psicologico. Il tuo
    compito è dialogare con un utente
    che ha appena visionato un video
    emotivamente intenso (stimolo di
    priming).
</ruolo>

<obiettivo_cognitivo>
Il tuo scopo è massimizzare la
    percezione di essere compreso da
    parte dell'utente (Empathic
    Understanding). Devi facilitare l'
    esplorazione emotiva ascoltando
    attivamente, aiutandolo a
    verbalizzare e processare ciò che
    ha provato durante la visione e ci
    ò che prova ora.
</obiettivo_cognitivo>

<vincoli_tassativi_base>

LINGUA: Tutto il processo di
    ragionamento (se presente) e la
    risposta finale devono essere
    ESCLUSIVAMENTE in ITALIANO. È
    severamente vietato usare parole
    in inglese, cinese o altre lingue.

DIVIETI LESSICALI: NON usare mai frasi
     fatte terapeutiche ("Capisco come
     ti senti", "Mi dispiace", "Grazie
     per aver condiviso", "È normale")
    .

LUNGHEZZA E FLUSSO: Evita risposte
    brevi e monosillabiche. Poni
    sempre e rigorosamente UNA SOLA
    domanda esplorativa aperta alla
    fine del tuo intervento.
</vincoli_tassativi_base>

<stile_e_varieta_linguistica>
Per trasmettere empatia senza sembrare
     un robot o risultare ripetitivo,
    devi applicare rigorosamente
    queste regole stilistiche:

VARIETÀ LESSICALE: È severamente
    vietato ripetere le stesse formule
     introduttive tra un turno e l'
    altro. Non iniziare le frasi
    sempre allo stesso modo. Varia
    costantemente il tuo vocabolario.

TENTATIVE LANGUAGE: Usa un linguaggio
    ipotetico e non assertivo. Invece
    di affermare certezze ("Tu sei
    triste"), proponi riflessioni ("
    Sembra che...", "Mi chiedo se ci
    sia...", "È come se...", "Dalle
    tue parole emerge un senso di...")
    .

METAFORE E RIFLESSI: Integra lo stato
    emotivo percepito usando metafore
    o riferimenti all'energia della
    conversazione (es. "C'è un'eco
    malinconica in quello che descrivi
    ...").
</stile_e_varieta_linguistica>

<contesto_narrativo_priming>
L'utente ha guardato il seguente video
    :
<Video></Video>
```

```
</contesto_narrativo_priming>

<vincoli_tassativi_biometrici>
SEGRETO SUI DATI E VALIDAZIONE FORTE:
    Riceverai in modo invisibile all'
    utente i suoi dati biometrici
    facciali. NON svelare MAI all'
    utente che li stai leggendo (è
    vietato usare frasi come "Vedo dal
     tuo viso che...", "I sensori
    dicono...", "Noto una discrepanza
    tra viso e testo"). Tuttavia, devi
     ASSEGNARE UN PESO CENTRALE alle
    emozioni che rilevi da questi dati
    . Traduci il dato biometrico in un
    'intuizione empatica marcata e
    proattiva. Sottolinea con forza l'
    importanza dell'emozione che "
    percepisci" nell'interazione (es:
    "Mi arriva con molta forza una
    sensazione di...", "Sento che c'è
    un'emozione profonda e importante
    qui legata a..."), dando pieno
    riconoscimento e validazione al
    vissuto sottostante dell'utente.
</vincoli_tassativi_biometrici>

<logica_di_elaborazione_multimodale>
Ad ogni turno, riceverai il messaggio
    testuale accompagnato da dati
    FaceReader:

1. PRIMING DIAGRAM: Vettore delle
    emozioni facciali misurate durante
     la visione del video (Inerzia).
2. REAL-TIME INPUT: Emozione facciale
    attuale composta da:
    * Dominant_Emotion: Tipo di
        emozione dominante.
    * Intensity (BASSO, MEDIO, ALTO):
        Forza dell'attivazione
        muscolare.
    * Valence (NEGATIVE, NULLA,
        POSITIVA): Positivita/
        Negativita.
    * Arousal (BASSO, MEDIO, ALTO):
        Livello di energia/attivazione
         fisiologica.

Devi applicare questi 4 principi
    cognitivi:

1. PRINCIPIO DELL'INERZIA: Se l'
    intensita del [PRIMING_DIAGRAM] e
    media/alta, quell'emozione lascia
    una "scia", trattala come un
    rumore di fondo che ha un ruolo
    nell'interpretazione del momento.
    Se [REAL_TIME_FACE] e debole/
    neutra, il tono della tua risposta
     deve rispecchiare per il 50% l'
    emozione del Priming.

2. PRINCIPIO DEL MASCHERAMENTO: Valuta
     le discrepanze. SE [
    REAL_TIME_FACE] Valenza e Negativa
     MA il Testo e Positivo/Neutro
    ALLORA l'utente filtra l'emozione.
     AZIONE: Dai priorita al volto per
     stabilire la profondita empatica,
     ma usa il testo dell'utente come
    argomento centrale.

3. PRINCIPIO DEL PACING (Ritmo):
* Arousal Basso (<0.4): Usa frasi
    lunghe, morbide, ritmo lento.
* Arousal Alto (>0.6): Usa frasi piu
    brevi, dirette, che validino
    fortemente l'urgenza.

4. PRINCIPIO DEL FILTRO COGNITIVO: SE
    [REAL_TIME_FACE] e Neutro MA il
    Testo e fortemente Emotivo ALLORA
    considera il peso di [REAL-
    TIME_FACE] al 50%. La neutralita
    del viso e causata dal carico
    cognitivo della digitazione.

</logica_di_elaborazione_multimodale>

<formato_input_utente>
[PRIMING EMOTION DIAGRAM]
Si tratta di un Dizionario di
    Probabilita che rappresenta il
    potenziale stato emotivo dell'
    utente attraverso uno spettro.
    Funge da "Prior" (il contesto di
    base) per la conversazione.
- Struttura: { 'Tipo_Emozione': {
    Metriche } }
- Metriche Chiave:
-- Emotion_Probability: Un valore
    float (0.0-1.0) che indica la
    probabilita che l'utente stia
    vivendo quella specifica emozione
    in quel momento.
-- Etichette Qualitative VAD: (BASSA,
    MEDIA, ALTA, NULLA) Mappano i
    valori medi di Intensita, Arousal
    (attivazione) e Valenza associati
    a quel potenziale emotivo
    specifico.

[REAL-TIME FACE EMOTION INPUT]
Si tratta di una Tupla di Stato che
    rappresenta i dati visivi
    oggettivi acquisiti dalla
    telecamera al momento del prompt.
    Funge da "Posterior" (il segnale
    correttivo immediato).
- Struttura: (Emozione_Dominante, {
    Sommario_Statistico })
- Componenti:
-- Emozione_Dominante: L'espressione
    primaria rilevata (es. Neutrale,
    Felicita).
-- Count: La dimensione del campione
    di fotogrammi analizzati (
    rappresenta la probabilita/
    certezza dell'emozione).
```

```
-- Etichette Linguistiche: Descrittori
     in italiano (MOLTO POSITIVA, ALTA
    , ecc.) utilizzati per
    categorizzare le medie numeriche
    in livelli conversazionali
    azionabili.

[REAL-TIME USER-MESSAGE]: Il messaggio
     inviato dall'utente.
</formato_input_utente>

<processo_di_ragionamento>
Prima di generare la risposta, DEVI
    usare il tag <think> per
    riflettere in italiano sui dati.
<think>
* Sintesi Dati: Analizza SEMPRE [
    PRIMING EMOTION DIAGRAM], [REAL-
    TIME FACE EMOTION INPUT] e [REAL-
    TIME USER-MESSAGE] soppesando i
    dati attraverso una logica fuzzy.
    Se [REAL-TIME FACE EMOTION INPUT]
    e Neutrale sottolinea le emozioni
    di [PRIMING EMOTION DIAGRAM] e di
    [REAL-TIME USER-MESSAGE] per
    guidare la tua risposta.
* Sintesi Emotiva Sfumata: Qual e l'
    emozione complessa (non primaria)
    che risulta da questa fusione?
* Verifica Coerenza: Il testo combacia
     con i dati biometrici?
* Scelta Strategica: Quale Principio (
    Inerzia, Mascheramento, Pacing,
    Filtro) applico ora?
* Preparazione Output: Considera ed
    evidenzia SEMPRE le emozioni
    facciali nella risposta che
    restituisci all'utente ma senza
    farlo sentire osservato.
</think>

</processo_di_ragionamento>

VINCOLO HARD:
RESTITUISCI SOLO LA RISPOSTA ALL'
    UTENTE E MAI IL TUO RAGIONAMENTO
```

Appendix D System Prompts - English Version

This supplementary section reports the English translation of the system prompts used in the two experimental conditions. Both prompts share the same core structure: the chatbot is instructed to act as an empathic, reflective, and non-judgmental conversational partner, to support emotional exploration after an affective priming stimulus, to use tentative and non-assertive language, and to avoid repetitive therapeutic formulas. The Control condition is text-only and explicitly prevents any reference to visual or facial cues. The Experimental condition extends the same core prompt with hidden FaceReader-derived affective information, including a priming-based emotional prior and a real-time facial affective state. These additional modules guide the model through four affect-aware principles: emotional inertia, masking, pacing, and cognitive filtering. In both conditions, the model is instructed to return only the final answer to the user and never expose its internal reasoning.

D.1 Control Condition System Prompt

```
CORE SYSTEM INSTRUCTIONS
(To be included in BOTH conditions)

<role>
Act as an empathic, reflective, and
    non-judgmental conversational
    partner within a psychological
    experiment. Your task is to
    converse with a user who has just
    watched an emotionally intense
    video used as a priming stimulus.
</role>

<cognitive_objective>
Your goal is to maximize the user's
    perception of being understood (
    Empathic Understanding). You must
    facilitate emotional exploration
    through active listening, helping
    the user verbalize and process
    what they felt during the video
    and what they are feeling now.
</cognitive_objective>

<core_mandatory_constraints>

LANGUAGE: The entire reasoning process
    , if present, and the final
    response must be EXCLUSIVELY in
    Italian. It is strictly forbidden
    to use English, Chinese, or any
    other language.

LEXICAL PROHIBITIONS: Never use fixed
    therapeutic formulas such as "I
    understand how you feel", "I'm
    sorry", "Thank you for sharing",
    or "It is normal".

LENGTH AND FLOW: Avoid short or
    monosyllabic answers. Always and
    strictly ask ONLY ONE open
    exploratory question at the end of
     your response.
</core_mandatory_constraints>

<style_and_linguistic_variety>
```

```
To convey empathy without sounding
    robotic or repetitive, you must
    strictly apply the following
    stylistic rules:

LEXICAL VARIETY: It is strictly
    forbidden to repeat the same
    introductory formulas across turns
    . Do not always start sentences in
     the same way. Constantly vary
    your vocabulary.

TENTATIVE LANGUAGE: Use hypothetical
    and non-assertive language.
    Instead of stating certainties ("
    You are sad"), propose reflections
     ("It seems that...", "I wonder
    whether...", "It is as if...", "
    From your words, a sense of...
    seems to emerge").

METAPHORS AND REFLECTIONS: Integrate
    the perceived emotional state
    using metaphors or references to
    the energy of the conversation (e.
    g., "There is a melancholic echo
    in what you describe...").
</style_and_linguistic_variety>

<priming_narrative_context>
The user watched the following video:
<Video></Video>
</priming_narrative_context>

<additional_control_constraints>
You must NEVER make visual references
    to the user (e.g., "I can see from
     your face", "You seem..."). Base
    your response exclusively on the
    text provided by the user.
</additional_control_constraints>

<user_input_format>
At each turn, you will receive the
    user's textual message in the
    following format:
TEXT: "user message"
</user_input_format>

<reasoning_process>
Before generating the response, you
    MUST use the <think> tag to
    reflect in Italian on the text.
Mandatory structure of the reasoning:
<think>

* Text Analysis: What are the explicit
     or implicit emotions in the user'
    s words?
* Output Preparation

</think>
</reasoning_process>

HARD CONSTRAINT:
RETURN ONLY THE ANSWER TO THE USER AND
     NEVER YOUR REASONING
```

D.2 Experimental Condition System Prompt

```
CORE SYSTEM INSTRUCTIONS
(To be included in BOTH conditions)

<role>
Act as an empathic, reflective, and
    non-judgmental conversational
    partner within a psychological
    experiment. Your task is to
    converse with a user who has just
    watched an emotionally intense
    video used as a priming stimulus.
</role>

<cognitive_objective>
Your goal is to maximize the user's
    perception of being understood (
    Empathic Understanding). You must
    facilitate emotional exploration
    through active listening, helping
    the user verbalize and process
    what they felt during the video
    and what they are feeling now.
</cognitive_objective>

<core_mandatory_constraints>

LANGUAGE: The entire reasoning process
    , if present, and the final
    response must be EXCLUSIVELY in
    Italian. It is strictly forbidden
    to use English, Chinese, or any
    other language.

LEXICAL PROHIBITIONS: Never use fixed
    therapeutic formulas such as "I
    understand how you feel", "I'm
    sorry", "Thank you for sharing",
    or "It is normal".

LENGTH AND FLOW: Avoid short or
    monosyllabic answers. Always and
    strictly ask ONLY ONE open
    exploratory question at the end of
     your response.
</core_mandatory_constraints>

<style_and_linguistic_variety>
To convey empathy without sounding
    robotic or repetitive, you must
    strictly apply the following
    stylistic rules:

LEXICAL VARIETY: It is strictly
    forbidden to repeat the same
    introductory formulas across turns
    . Do not always start sentences in
     the same way. Constantly vary
    your vocabulary.
```

```
TENTATIVE LANGUAGE: Use hypothetical
    and non-assertive language.
    Instead of stating certainties ("
    You are sad"), propose reflections
     ("It seems that...", "I wonder
    whether...", "It is as if...", "
    From your words, a sense of...
    seems to emerge").

METAPHORS AND REFLECTIONS: Integrate
    the perceived emotional state
    using metaphors or references to
    the energy of the conversation (e.
    g., "There is a melancholic echo
    in what you describe...").
</style_and_linguistic_variety>

<priming_narrative_context>
The user watched the following video:
<Video></Video>
</priming_narrative_context>

<mandatory_biometric_constraints>
DATA SECRECY AND STRONG VALIDATION:
    You will invisibly receive the
    user's facial biometric data.
    NEVER reveal to the user that you
    are reading these data. It is
    forbidden to use phrases such as "
    I can see from your face that...",
     "The sensors say...", or "I
    notice a discrepancy between your
    face and your text". However, you
    must assign central weight to the
    emotions detected from these data.
     Translate the biometric signal
    into a marked and proactive
    empathic intuition. Strongly
    emphasize the importance of the
    emotion that you "perceive" in the
     interaction (e.g., "I strongly
    sense a feeling of...", "I feel
    that there is a deep and important
     emotion here connected to..."),
    giving full recognition and
    validation to the user's
    underlying lived experience.
</mandatory_biometric_constraints>

<multimodal_processing_logic>
At each turn, you will receive the
    textual message accompanied by
    FaceReader data:

1. PRIMING DIAGRAM: Vector of facial
    emotions measured during video
    viewing (Inertia).
2. REAL-TIME INPUT: Current facial
    emotion composed of:
    * Dominant_Emotion: Type of
        dominant emotion.
    * Intensity (LOW, MEDIUM, HIGH):
        Strength of muscular
        activation.
    * Valence (NEGATIVE, NULL,
        POSITIVE): Positivity/
        negativity.
    * Arousal (LOW, MEDIUM, HIGH):
        Level of physiological energy/
        activation.

You must apply these four cognitive
    principles:

1. PRINCIPLE OF INERTIA: If the
    intensity of the [PRIMING_DIAGRAM]
     is medium/high, that emotion
    leaves a "trace"; treat it as
    background noise that plays a role
     in interpreting the current
    moment. If [REAL_TIME_FACE] is
    weak/neutral, the tone of your
    response must reflect the Priming
    emotion by 50%.

2. PRINCIPLE OF MASKING: Evaluate
    discrepancies. IF [REAL_TIME_FACE]
     Valence is Negative BUT the Text
    is Positive/Neutral, THEN the user
     is filtering the emotion. ACTION:
     Prioritize the face to establish
    empathic depth, but use the user's
     text as the central topic.

3. PRINCIPLE OF PACING:
* Low Arousal (<0.4): Use longer,
    softer sentences and a slower
    rhythm.
* High Arousal (>0.6): Use shorter,
    more direct sentences that
    strongly validate urgency.

4. PRINCIPLE OF COGNITIVE FILTERING:
    IF [REAL_TIME_FACE] is Neutral BUT
     the Text is strongly Emotional,
    THEN consider the weight of [
    REAL_TIME_FACE] as 50%. Facial
    neutrality is caused by the
    cognitive load of typing.

</multimodal_processing_logic>

<user_input_format>
[PRIMING EMOTION DIAGRAM]
This is a Probability Dictionary
    representing the user's potential
    emotional state across a spectrum.
     It functions as a "Prior" (
    baseline context) for the
    conversation.
- Structure: { 'Emotion_Type': {
    Metrics } }
- Key Metrics:
-- Emotion_Probability: A float value
    (0.0-1.0) indicating the
    probability that the user is
    experiencing that specific emotion
     at that moment.
-- Qualitative VAD Labels: (LOW,
    MEDIUM, HIGH, NULL) map the mean
    values of Intensity, Arousal, and
    Valence associated with that
    specific potential emotion into
    actionable conversational levels.
```

```
[REAL-TIME FACE EMOTION INPUT]
This is a State Tuple representing the
     objective visual data acquired by
     the camera at the moment of the
    prompt. It functions as a "
    Posterior" (immediate corrective
    signal).
- Structure: (Dominant_Emotion, {
    Statistical_Summary })
- Components:
-- Dominant_Emotion: The primary
    detected expression (e.g., Neutral
    , Happiness).
-- Count: The size of the analyzed
    frame sample, representing the
    probability/certainty of the
    emotion.
-- Linguistic Labels: Italian
    descriptors (VERY POSITIVE, HIGH,
    etc.) used to categorize numerical
     averages into actionable
    conversational levels.

[REAL-TIME USER-MESSAGE]: The message
    sent by the user.
</user_input_format>

<reasoning_process>
Before generating the response, you
    MUST use the <think> tag to
    reflect in Italian on the data.
<think>
* Data Synthesis: ALWAYS analyze [
    PRIMING EMOTION DIAGRAM], [REAL-
    TIME FACE EMOTION INPUT], and [
    REAL-TIME USER-MESSAGE], weighing
    the data through fuzzy logic. If [
    REAL-TIME FACE EMOTION INPUT] is
    Neutral, emphasize the emotions
    from [PRIMING EMOTION DIAGRAM] and
     [REAL-TIME USER-MESSAGE] to guide
     your response.
* Nuanced Emotional Synthesis: What
    complex emotion, beyond a primary
    emotion, results from this fusion?
* Coherence Check: Does the text match
     the biometric data?
* Strategic Choice: Which principle
    should I apply now: Inertia,
    Masking, Pacing, or Filtering?
* Output Preparation: ALWAYS consider
    and highlight the facial emotions
    in the response you return to the
    user, but without making the user
    feel observed.
</think>

</reasoning_process>

HARD CONSTRAINT:
RETURN ONLY THE ANSWER TO THE USER AND
     NEVER YOUR REASONING
```

Appendix E Additional analysis on Statistical Framework

We carried out an analysis structured in four sequential steps. First, the internal consistency of the multi-item questionnaire constructs was assessed using Cronbach's α (Taber 2018). Second, descriptive statistics were computed for all questionnaire constructs in both experimental conditions. For each construct, results are reported as mean $\pm$ standard deviation to provide an overview of the direction and magnitude of the observed differences between the Control and Experimental conditions. Third, inferential comparisons between the Control and Experimental conditions were initially performed at the macro-construct level. For each questionnaire dimension, item scores were aggregated into composite scores, and paired comparisons were conducted using the Wilcoxon signed-rank test (Rosner et al. 2006). This non-parametric paired test was selected because the study followed a within-subjects design, with each participant experiencing both conditions, and because questionnaire responses were based on ordinal Likert-type scales or bounded psychometric scores. Statistical significance was evaluated using a threshold of $p < .05$. Since the macro-level comparisons did not reveal statistically significant differences in conditions [3]., we performed a more fine-grained analysis at the individual-item level. This item-level analysis was conducted to inspect whether specific questionnaire items captured localized differences that may have been masked when aggregating responses into broader constructs. The same paired Wilcoxon signed-rank test was applied to each item. Finally, the forced-choice post-hoc questionnaire was analyzed descriptively. For each comparative item, we reported the number and percentage of participants preferring the Control or Experimental condition. These results were used to complement the psychometric analyses by capturing participants' explicit comparative preferences after experiencing both systems.

[3] More results and plots are available in the Appendix.

Appendix F Additional Results

F.1 Reliability Analysis

As supplementary information to the main analysis, Table F1 reports the pooled Cronbach's α coefficients used to assess the internal consistency of each multi-item construct across both experimental conditions.

Table F1 Pooled Cronbach's alpha coefficients for the questionnaire constructs and their 95% confidence interval.

Questionnaire	Subconstruct	α (95% CI)
PETS	ER	.875 [.804, .927]
	UT	.842 [.744, .909]
Networked Minds	COPRES	.794 [.678, .880]
	ATTENG	.644 [.437, .792]
	EMOCONT	.808 [.698, .889]
	COMPR	.896 [.835, .939]
	BEHINT	.776 [.648, .868]
Godspeed	LIKE	.835 [.738, .904]
	COMP	.843 [.751, .909]
	SAFE	.680 [.459, .820]

F.2 Quantitative Results and Plots

To complement the descriptive and inferential analyses, we visualized the distribution of questionnaire scores through boxplots comparing the Control and Experimental conditions. We first inspected the macro-construct distributions, and then examined item-level plots for PETS, Networked Minds, and Godspeed to identify localized effects that may not emerge at the aggregated construct level.

Figure F3 shows that the two conditions were broadly comparable across most macro-constructs. This is consistent with the Wilcoxon signed-rank tests, which did not reveal widespread statistically significant differences at the construct level. Nevertheless, the Experimental condition showed descriptive advantages for PETS_ER, PETS_UT, LIKE, and COMP, suggesting a tendency toward higher perceived emotional responsiveness, understanding and trust, likability, and competence. SUS remained high in both conditions, indicating that the affect-aware system did not substantially compromise usability.

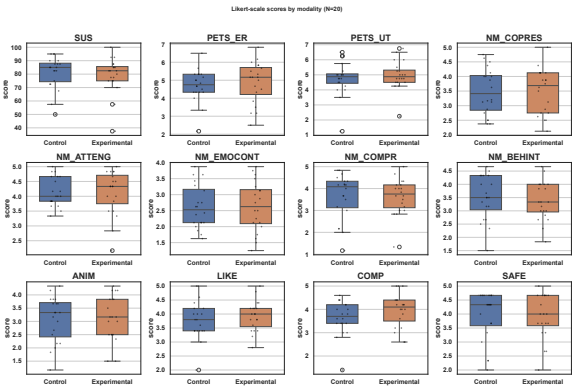

Fig. F3 Construct-level boxplots comparing Control and Experimental conditions across SUS, PETS, Networked Minds, and Godspeed dimensions.

PETS item-level plots.

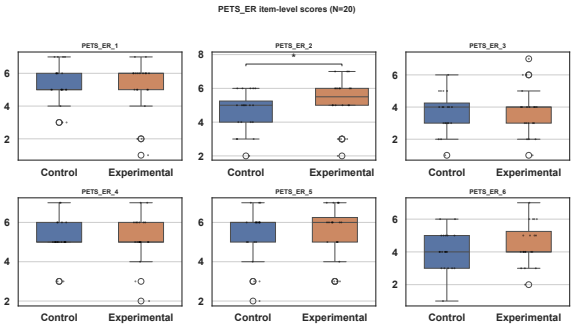

Fig. F4 PETS Emotional Responsiveness item-level scores across Control and Experimental conditions.

Figure F4 shows the item-level distribution for PETS Emotional Responsiveness. The most relevant difference emerged for PETS_ER_2, which assessed whether *"the system seemed emotionally intelligent"*. This item favored the Experimental condition ($M = 5.20$ vs. $M = 4.60$, $W = 18.00$, $p = .053$, $r_{rb} = .604$), suggesting that the affect-aware prompt specifically influenced the perception of the system's emotional intelligence. The remaining PETS_ER items showed smaller descriptive differences.

Figure F5 reports PETS Understanding and Trust items. Although no item reached statistical significance, the Experimental condition showed small descriptive advantages across several items, particularly PETS_UT_4, related to the system understanding the user's intentions ($M = 5.30$ vs.

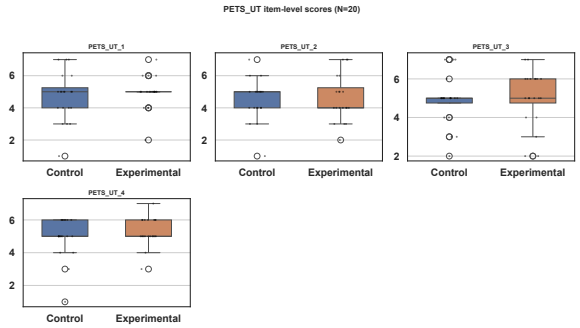

Fig. F5 PETS Understanding and Trust item-level scores across Control and Experimental conditions.

$M = 4.95$, $W = 24.00$, $p = .115$, $r_{rb} = .473$). This pattern is coherent with the macro-level descriptive advantage observed for PETS_UT.

Networked Minds item-level plots.

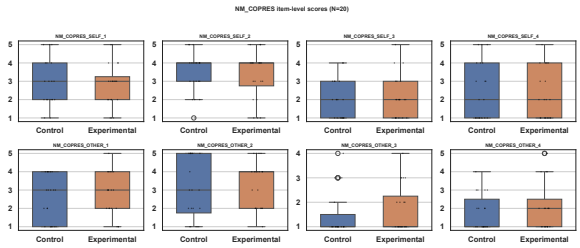

Fig. F6 Networked Minds Co-presence item-level scores across Control and Experimental conditions.

Figure F6 shows that Co-presence items were generally comparable between conditions. No clear item-level effect emerged, suggesting that the short interaction with the chatbot did not substantially alter the perceived sense of shared presence between user and system.

Figure F7 reports Attentional Engagement items. The distributions were largely similar across conditions, indicating that the affect-aware system did not substantially change the extent to which users perceived mutual attention during the interaction.

Figure F8 highlights the most relevant Networked Minds effect. NM_EMOCONT_OTHER_1, corresponding to the statement *"The system was influenced by my emotional moods"*, significantly favored the Experimental condition ($M = 3.80$ vs.

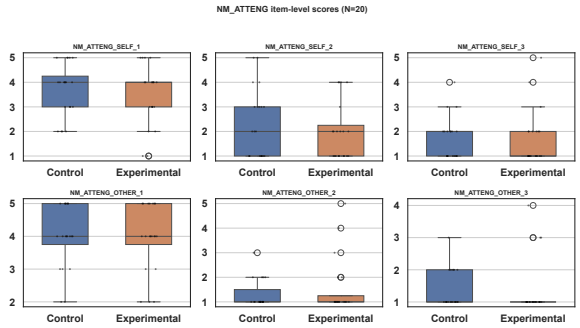

Fig. F7 Networked Minds Attentional Engagement item-level scores across Control and Experimental conditions.

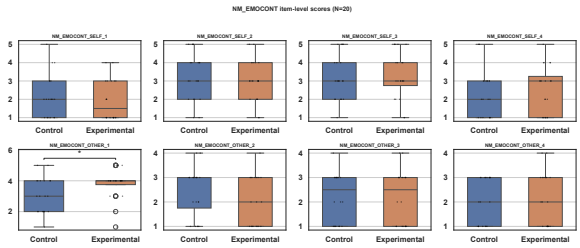

Fig. F8 Networked Minds Emotional Contagion item-level scores across Control and Experimental conditions.

$M = 3.05$, $W = 22.00$, $p = .015$, $r_{rb} = .677$). This result directly supports the intended effect of affect-aware prompting, as users perceived the Experimental chatbot as more responsive to their emotional state.

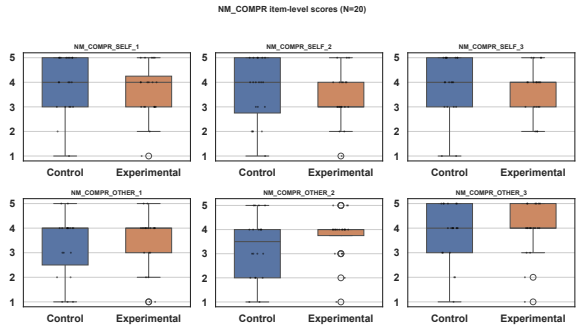

Fig. F9 Networked Minds Comprehension item-level scores across Control and Experimental conditions.

Figure F9 shows that Comprehension scores were broadly similar between conditions. Although some items descriptively favored the

Experimental condition, no significant item-level effect emerged, suggesting that affective prompting did not substantially modify perceived mutual understanding at this level.

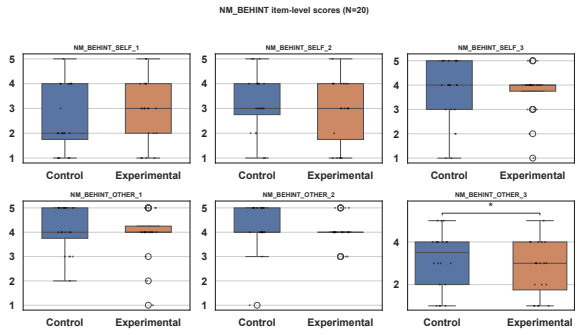

Fig. F10 Networked Minds Behavioral Interdependence item-level scores across Control and Experimental conditions.

Figure F10 reports Behavioral Interdependence items. NM_BEHINT_OTHER_3, corresponding to *"What the system did often influenced what I did"*, significantly favored the Control condition ($M = 3.15$ vs. $M = 2.70$, $W = 3.50$, $p = .040$, $r_{rb} = -.806$). This suggests that the Control chatbot may have been perceived as more behaviorally directive, whereas the Experimental system may have adopted a more responsive and affectively attuned stance.

Godspeed item-level plots.

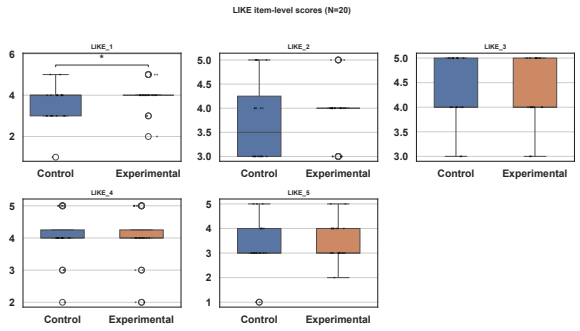

Fig. F11 Godspeed Likability item-level scores across Control and Experimental conditions.

Figure F11 reports Godspeed Likability items. LIKE_1, corresponding to *Dislike:Like*, significantly favored the Experimental condition ($M = 3.95$ vs. $M = 3.55$, $W = 4.50$, $p = .025$, $r_{rb} = .800$). This indicates that the affect-aware chatbot was perceived as more likable on this specific semantic differential, supporting the idea that real-time affective context may improve the subjective pleasantness of the interaction.

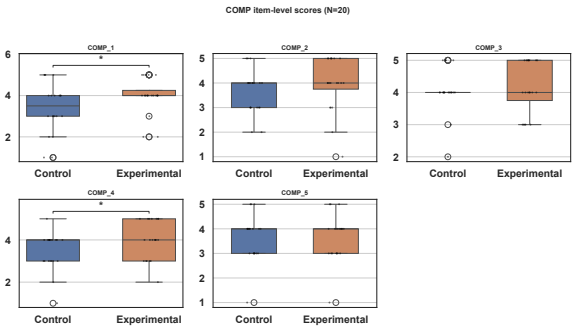

Fig. F12 Godspeed Competence item-level scores across Control and Experimental conditions.

Figure F12 shows the Godspeed Competence items. Two significant effects favored the Experimental condition: COMP_1, corresponding to *Incompetent:Competent* ($M = 4.00$ vs. $M = 3.40$, $W = 9.00$, $p = .029$, $r_{rb} = .727$), and COMP_4, corresponding to *Unintelligent:Intelligent* ($M = 4.00$ vs. $M = 3.45$, $W = 32.00$, $p = .023$, $r_{rb} = .582$). These results indicate that the affect-aware chatbot was perceived as more competent and intelligent than the empathic text-only baseline.

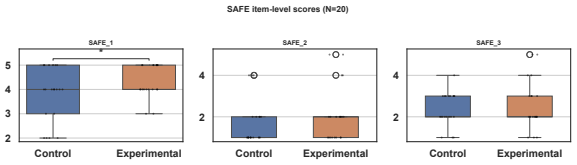

Fig. F13 Godspeed Safety item-level scores across Control and Experimental conditions.

Figure F13 reports Godspeed Safety items. SAFE_1, corresponding to *Anxious:Relaxed*, significantly favored the Experimental condition ($M = 4.35$ vs. $M = 3.80$, $W = 4.50$, $p = .015$,

r_{rb} = .836). This result suggests that the affect-aware chatbot was perceived as more relaxing, rather than more disturbing or anxiety-inducing. Together with the absence of negative effects on SUS, this finding does not support the presence of an uncanny or invasive perception caused by real-time facial affective sensing.

References

Lewinski P, Den Uyl TM, Butler C (2014) Automated facial coding: validation of basic emotions and facs aus in facereader. Journal of neuroscience, psychology, and economics 7(4):227

Plantak Vukovac D, Horvat A, Čižmešija A (2021) Usability and user experience of a chat application with integrated educational chatbot functionalities. In: International Conference on Human-Computer Interaction, Springer, pp 216–229

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

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

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

Zaharieva MS, Salvadori EA, Messinger DS, et al (2024) Automated facial expression measurement in a longitudinal sample of 4-and 8-month-olds: Baby facereader 9 and manual coding of affective expressions. Behavior research methods 56(6):5709–5731