From Emotional Mirroring to Emotional Attunement: Do LLMs and Humans Attune to Each Other?

Anonymous ACL submission

Abstract

Much prior work establishes that LLMs effectively mirror the affective state of a user. However, human social interaction depends not on immediate mirroring, but on emotional attunement, a process of bidirectional affective synchronization between individuals. In this work, we evaluate whether LLMs emotionally attune with users, comparing LLM-user interactions with client-therapist interactions. We find evidence for a "hollow echo" effect: LLMs strongly mirror user affect in immediate responses, but fail to attune to user emotional state across multi-turn interactions. This contrasts with client-therapist interactions, where we observe a more durable and moderate form of attunement. Moreover, we find that while clients attune to their therapists, users do *not* attune to LLMs, such that user-LLM attunement cannot be said to be bidirectional. Our findings indicate that current LLMs are inadequate for relationally complex contexts, which require sustained attunement, rather than immediate mirroring.

1 Introduction

Do LLMs attune to the emotional state of a user? Recent studies establish that LLMs perform well on benchmarks composed of Emotional Quotient (EQ) tests (Wang et al., 2023), competently recognize affective nuance (Ju et al., 2025), and generate more convincingly empathetic responses than humans in some settings (Sorin et al., 2024; Pinzuti et al., 2025). LLMs are rapidly becoming ubiquitous in emotionally sensitive domains, ranging from mental health support (Pinzuti et al., 2025) to daily companionship (Cheng et al., 2025). Yet concerns about the shortcomings of LLMs in affective settings involving self-disclosure and therapeutic application, particularly for young people, have called into question whether models truly exhibit sustained emotional attunement, or only generate responses that match a user's emotion in the immediate moment (Parmar and Mazumdar, 2025).

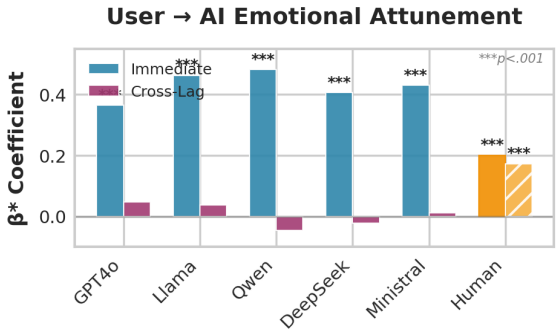

Figure 1: The "hollow echo" effect: LLMs match user emotional state in immediate replies, but fail to attune in subsequent messages. This contrasts with human-human emotional attunement, which is moderate in tone and persistent over the course of interactions.

Thus, in the present work, we consider the question of whether LLMs exhibit sustained *emotional attunement*, a bidirectional dynamic sometimes referred to as "social synchrony" wherein one speaker's emotional state reflects the other's across multiple speaking turns (Feldman, 2012; Doré and Morris, 2018; Palumbo et al., 2017), rather than the more straightforward and immediate *emotional alignment*. We pose three research questions:

(1) Does ChatGPT emotionally attune to users (and vice versa) in multi-turn conversations?

(2) How does attunement between LLMs and users compare to attunement between two humans attempting to attune to each other?

(3) How consistent are the effects observed in ChatGPT across instruction-tuned LLMs?

To investigate these questions, we collected a corpus of in-the-wild ChatGPT interactions and used three measures to compute affective scores for each message in the corpus. We fit random-intercept cross-lagged panel models to these scores, which revealed that ChatGPT attunes to the user's

affect in the response immediately following the user's message, but *not* to the user's affect just one message prior. Users exhibited no immediate or sustained emotional attunement to ChatGPT. We compared these results to therapist-client interactions in the Cognitive Behavioral Therapy of Wang et al. (2025), and found that client-therapist interactions exhibit immediate, sustained, and bidirectional (client to therapist and therapist to client) attunement, in stark contrast to human-LLM interactions. Finally, we employed a controlled replay experiment with four open-source LLMs (Llama-3.2-8B, Qwen-3-8B, DeepSeek-R1, and Ministral-3), and found that these models consistently demonstrated the same interaction pattern as ChatGPT: immediate mirroring but no sustained attunement.

Our contributions are threefold: (1) We introduce and empirically evaluate an approach to evaluating LLM emotional attunement; (2) We demonstrate a notable difference between human-human and human-LLM interaction that may foster new ways of thinking about emotional interaction with LLMs; and (3) We demonstrate the consistency of our findings across five state-of-art LLMs. All code used in this work is publicly available.[1]

2 Related Work

Prior work establishes that successful social interaction depends on interpersonal attunement or "social synchrony," referring to the dynamic coordination of behavioral, physiological, and affective states between interaction partners (Feldman, 2012, 2020). For instance, bonding friends may mirror both hand gestures and vocal patterns in conversation (Dragojevic et al., 2015; Tschacher et al., 2014). Such coordination is automatic for most people (Palumbo et al., 2017), and degree of attunement has been shown to predict outcomes in text-based emotional support (Doré and Morris, 2018).

As LLMs begin to provide emotional support to in both formal and informal settings, recent work has probed the emotional intelligence of these models, including in multi-turn contexts (Chen et al., 2023; Wang et al., 2023). Recent work has focused on evaluating and improving model emotional dialogue via fine-tuning, particularly in multi-turn, long-context settings (Koudounas et al., 2025; Yang and Jurgens, 2024). Studies have also indicated shortcomings in LLM emotion recognition, especially across conversations (rather than within a single dialogue) (Nazar and Asgari, 2025). We build on this line of research by investigating whether LLMs exhibit sustained, bidirectional attunement.

[1]PLACEHOLDER LINK (please see zipped repo).

3 Data

Sustained Affective Conversation Corpus. We used Prolific to collect in-the-wild ChatGPT logs from N=85 young adult participants (43 male, 37 female, and 5 did not disclose gender; age M = 21.7 years, SD = 2.0, min=18, max=25), who submitted all their conversations with the model during 2024 and 2025. The resulting corpus contained 157,188 messages across 11,000 conversations. Because measuring emotional attunement requires both an immediate context and a broader conversational context, we filtered this initial dataset such that it included only those conversations that included at least six turns (Koudounas et al., 2025) which produced a filtered sample of 447 conversations (N_{msg} = 7,148, M_{turn} = 8.00). A manual review of 100 of these conversations demonstrated that the majority reflected solely instrumental or transactional interactions with the model (*e.g.*, requesting technical help, or assistance with homework). To ensure that the dataset reflected the domain of our analysis (emotional interactions with an LLM), the first two authors qualitatively coded the sample as either primarily Affective (including some interaction beyond the instrumental) or primarily Non-Affective. We then removed all Non-Affective conversations from the sample. This process identified the N=73 affective conversations that constitute our final dataset (N_{msg} = 1,338, M_{turn} = 9.16).

CBT Therapy Corpus. To contextualize our results by reference to human affective attunement, we used the Cognitive Behavioral Therapy (CBT) interview dataset of Wang et al. (2025), which includes 76 conversations between therapists and their clients (N_{msg} = 3,661, M_{turn} = 24.09) (Wang et al., 2025). This similarly sized dataset provides a strong reference point for our work in that emotional attunement is expected in these conversations (as opposed to, for example, casual text conversations between friends), and we observe that the disclosures clients make to therapists often resemble those that users make to ChatGPT in our dataset.

4 Methods

Modeling Affect. We employ a quantitative framework to measure affect in user and AI messages. To ensure the robustness of our results, we computed

affect using three complementary measures:
RoBERTa: a RoBERTa (Liu et al., 2019) finetuned for sentiment analysis (Loureiro et al., 2022); we compute affect as positive sentiment score.
VADER: a rule-based scorer sensitive to paralinguistic cues (emojis, punctuation) producing continuous affective scores (Hutto and Gilbert, 2014).
NRC-VAD: the psycholinguistic lexicon developed by Mohammad (2025); we compute affect as the mean valence of words in a message.

We focus on valence (pleasantness) as the primary measure of affect because it is demonstrated to be the most consistent and detectable psycholinguistic dimension (Wolfe and Caliskan, 2022).

Modeling Emotional Attunement. Using affective scores computed for each message, we fit random-intercept cross-lagged panel models (RI-CLPMs) (Hamaker et al., 2015) to model the effect of previous messages on the affect of the current message. Specifically, we fit four models:
U->AI Immediate: Models affect of LLM message m_t at turn t as a function of the user message m_{t-1} at turn $t-1$.
AI->U Immediate: Models affect of user message m_t as a function of the LLM message m_{t-1}.
U->AI Sustained: Models affect of LLM message m_t as a function of user message m_{t-3} at $t-3$.
AI->U Sustained: Models affect of user message m_t as a function of LLM message m_{t-3}.

We report standardized β^* coefficients and associated p-values as measures of attunement, fitting separate models for each measure (RoBERTa, VADER, NRC-VAD). We chose to use the message at turn $t-3$ to model sustained attunement, rather than mean score for all previous messages, for two reasons. First, this is a more achievable target for an LLM: we would expect that the model can take the local context of a conversation into account. Second, we risk regressing too far toward the mean when averaging affective scores from an entire conversation, particularly when a conversation is longer and may contain more shifts in affect.

CBT Comparison. To compare with the CBT dataset, we fit four RI-CLPMs to client-therapist interactions, employing the methodology described above but substituting *clients* for *users* (**C** replaces **U**), and *therapists* for *AI* (**T** replaces **AI**).

Generalization Experiment in Four LLMs. We evaluated whether results with ChatGPT generalize to other LLMs. To do so, we modified our original dataset to produce four new datasets, each of which corresponds to messages generated by an open LLM in response to a portion of the original conversation thread. Specifically, given a human-AI conversation C_i, for any LLM-generated message m_t at turn t where a user-written message exists at turn $t-1$, we create a new conversation R_i by replacing m_t with a new message r_t generated by an open LLM, updating the conversation such that the LLM sees only user-written messages and its own responses, not the responses of another model. We then fit RI-CLPMs to the data, in accordance with the methods above. We used the following LLMs to generate new datasets: **Llama 3.2 8B** (Grattafiori et al., 2024), **Qwen 3 8B** (Yang et al., 2025), **DeepSeek-R1** (DeepSeek-AI et al., 2024), and **Ministral 3 8B** (Jiang et al., 2023). All experiments were run locally.

5 Results

RQ1: Immediate vs. Sustained Attunement. Table 1 presents CLPM coefficients for the Immediate models, demonstrating a significant User→AI effect of ($\beta^* = 0.12$–0.37, $p < .05$) across the three affective measures, providing strong evidence that user affect in an immediately preceding message predicts the affect of a ChatGPT response. This accords with prior work suggesting that LLMs including GPT-4 exhibit emotional alignment with users. On the other hand, coefficients and corresponding p-values are non-significant for the (AI→User) direction, indicating that we see no evidence that users match their affect to the model's.

ChatGPT Immediate RI-CLPM Model Effects

Measure	U→AI	p	AI→U	p
RoBERTa	**0.365*** **	<.001	0.034	.471
NRC Val.	**0.312*** **	<.001	0.040	.377
VADER	**0.117***	.021	−0.050	.257

Table 1: The Immediate RI-CLPM models show that ChatGPT mirrors user affect in immediate responses (U->AI), but users do not mirror LLM affect (AI->U).

Unlike the Immediate models, the Sustained models offer no evidence for for emotional attunement at lagged turn $t-3$, either by the model toward the user, or by the user toward the model. Thus, while the LLM provides an immediate mirror of the user's affective state, we find no evidence for *sustained* emotional attunement.

RQ2: Model vs. Human Attunement. As visualized in 2, therapist interactions with clients captured in the CBT corpus exhibit bidirectional attunement that persists over time, and which is more

ChatGPT Sustained RI-CLPM Effects

Measure	U→AI	p	AI→U	p
RoBERTa	0.048	.282	0.043	.341
NRC Val.	0.036	.425	0.077	.089
VADER	0.036	.401	−0.038	.364

Table 2: Neither LLMs nor humans exhibit sustained attunement in multi-turn conversations (cross-lagged effects, centered). Bold indicates $p < .05$.

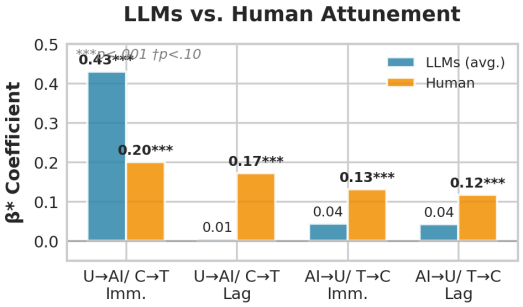

Figure 2: Human-to-human emotional attunement (based on the CBT corpus) is moderate in tone and persistent across time. LLMs strongly mirror user tone, but fail to durably attune to users.

moderate on average than LLM responses. In fact, therapist interactions with clients exhibited *greater* attunement with messages from lagged turns than in immediate responses, indicating the importance of longer-term context for attunement in emotionally laden conversations. Moreover, clients largely mirror the affect of their therapists, unlike ChatGPT users, who entirely disregard the affect of the model. As described by Table 3, we observe evidence of these effects across all three measures.

CBT Corpus Sustained RI-CLPM Effects

Measure	C→T	p	T→C	p
RoBERTa	**0.173*** **	<.001	**0.118*** **	<.001
NRC Val.	**0.101*** **	<.001	**0.048***	.003
VADER	**0.072*** **	<.001	0.031	.061

Table 3: Human therapist-client baseline (CBT corpus). C=Client, T=Therapist. Unlike LLMs, human dyads exhibit *bidirectional*, *sustained* emotional attunement.

RQ3: Generalization Across LLMs. As visualized in Figure 1, we observe evidence for immediate emotional alignment across all five LLMs tested ($p < .001$ in all cases), and no evidence for sustained emotional alignment in any LLM tested ($n.s.$ in all cases). As in the case of the ChatGPT conversations, immediate mirroring by LLMs with users is consistently more pronounced ($\beta^* = 0.41$–0.48) than therapist with clients ($\beta^* = 0.09$–0.20). Appendix Table 5 describes full results.

6 Discussion

We offer evidence for a "hollow echo" effect in LLM emotional dynamics: all tested models exhibit strong immediate emotional mimicry, but do not sustain emotional attunement. Our findings suggest a phenomenon in human-AI interaction: dialogues that are fluent, cognitive meaningfully, and even empathetic, but which lack affective synchronization over time. Despite this shortcoming, LLMs are already being adopted in settings requiring relational competence, including mental health support (Soun and Nair, 2023; Pinzuti et al., 2025; Madani et al., 2024) and long-term companionship (Hwang et al., 2025; Zhang et al., 2025). Our results indicate that present LLMs are insufficient for maintaining the natural and satisfying interactions necessary for fostering such bonds. Indeed, our observation that users exhibit no attunement whatsoever to LLMs suggests the depth of the gap felt by users interacting with LLMs, when compared to interacting with humans. Whether LLMs *should* serve in relationally sensitive contexts remains an open question, but our findings suggest that, at present, these technologies fail to attune to a user's emotional state over the course of a conversation, or longer.

Limitations. Our study has several limitations. First, though of comparable size to the CBT corpus, our dataset is relatively small. Second, the lexical semantic approaches we use to model affect may miss nuance like sarcasm and irony. Finally, our generalization experiment must be validated with further work, as we do not obtain true in-the-wild interactions from LLMs other than ChatGPT.

7 Conclusion

We demonstrate that LLMs exhibit strong immediate emotional responsiveness, but lack the sustained bidirectional influence that characterizes human-human emotional attunement. Our work surfaces a limitation of modern LLMs and raises questions for future work of whether and how to build models that users turn to for disclosure and companionship.

Ethics Statement

We understand that our work touches on an emerging topic (LLMs used in emotional support of

users) that is far from settled, both socially and ethically. Our hope is to provide information relevant to that debate, rather than settling it in this work. We also acknowledge that improving emotional dynamics in human-AI interaction may constitute a "dual-use" scenario, wherein such systems might be used by a self-interested actor to manipulate user affect. The present work offers no mechanism that would facilitate such actions, and we note that our intention and expectation is that this research will ultimately be used to foster emotionally satisfying interactions between users and LLMs that foreground user agency.

References

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 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*, pages 1170–1183, Stroudsburg, PA, USA. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of LLM sycophancy. *arXiv [cs.CL]*.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, and 68 others. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv [cs.CL]*.

Bruce P Doré and Robert R Morris. 2018. Linguistic synchrony predicts the immediate and lasting impact of text-based emotional support. *Psychol. Sci.*, 29(10):1716–1723.

Marko Dragojevic, Jessica Gasiorek, and Howard Giles. 2015. *Communication Accommodation Theory*, pages 1–21. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Ruth Feldman. 2012. Parent–infant synchrony: A biobehavioral model of mutual influences in the formation of affiliative bonds: Parent-infant synchrony. *Monogr. Soc. Res. Child Dev.*, 77(2):42–51.

Ruth Feldman. 2020. What is resilience: an affiliative neuroscience approach. *World Psychiatry*, 19(2):132–150.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Ellen L Hamaker, Rebecca M Kuiper, and Raoul P P P Grasman. 2015. A critique of the cross-lagged panel model. *Psychol. Methods*, 20(1):102–116.

C Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Angel Hsing-Chi Hwang, Fiona Li, Jacy Reese Anthis, and Hayoun Noh. 2025. How AI companionship develops: Evidence from a longitudinal study. *arXiv [cs.HC]*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv [cs.CL]*.

Hyojin Ju, Jungeun Lee, Seungwon Yang, Jungseul Ok, and Inseok Hwang. 2025. Toward affective empathy via personalized analogy generation: A case study on microaggression. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–31, New York, NY, USA. ACM.

Alkis Koudounas, Moreno La Quatra, and Elena Baralis. 2025. DeepDialogue: A multi-turn emotionally-rich spoken dialogue dataset. *arXiv [cs.CL]*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from twitter. *arXiv [cs.CL]*.

Navid Madani, Sougata Saha, and Rohini Srihari. 2024. Steering conversational large language models for long emotional support conversations. *arXiv [cs.CL]*.

Saif M Mohammad. 2025. NRC VAD lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv [cs.CL]*.

Nizi Nazar and Ehsaneddin Asgari. 2025. EICAP: Deep dive in assessment and enhancement of large language models in emotional intelligence through multi-turn conversations. *arXiv [cs.CL]*.

Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. 2017. Interpersonal autonomic physiology: A systematic review of the literature. *Pers. Soc. Psychol. Rev.*, 21(2):99–141.

Darshna Parmar and Pramit Mazumdar. 2025. Emotionally aware or tone-deaf? evaluating emotional alignment in LLM-based conversational recommendation systems. In *Proceedings of the 9th Widening NLP Workshop*, pages 167–174, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edoardo Pinzuti, Oliver Tüscher, and André Ferreira Castro. 2025. Comparative performance of large language models in emotional safety classification across sizes and tasks. *Front. Artif. Intell.*, 8(1706090):1706090.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *J. Med. Internet Res.*, 26(1):e52597.

Ritesh S Soun and Aadya Nair. 2023. ChatGPT for mental health applications: A study on biases. In *The Third International Conference on Artificial Intelligence and Machine Learning Systems*, pages 1–5, New York, NY, USA. ACM.

Wolfgang Tschacher, Georg M Rees, and Fabian Ramseyer. 2014. Nonverbal synchrony and affect in dyadic interactions. *Front. Psychol.*, 5:1323.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. *arXiv [cs.CL]*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023. Emotional intelligence of large language models. *arXiv [cs.AI]*.

Robert Wolfe and Aylin Caliskan. 2022. VAST: The valence-assessing semantics test for contextualizing language models. *arXiv [cs.CL]*, (10):11477–11485.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv [cs.CL]*.

Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. *arXiv [cs.CL]*.

Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of AI companions: How human-chatbot relationships influence well-being. *arXiv [cs.HC]*.

A Affective vs. Non-Affective Conversations

To validate our filtering process, we compared the RoBERTa sentiment distributions between the transactional dataset ($N = 447$) and the annotated affective dataset ($N = 73$). As shown in Table 4, the annotated sample exhibits higher mean scores for negative and positive sentiment, lower neutral scores, and consistently higher standard deviations (SD), confirming that the filtered dataset contains richer emotional content.

	Full Sample		Annotated Sample	
Metric	**Mean**	**SD**	**Mean**	**SD**
Negative	0.115	0.162	0.215	0.271
Neutral	0.702	0.227	0.510	0.286
Positive	**0.184**	**0.226**	**0.275**	**0.319**

Table 4: RoBERTa descriptive statistics. The annotated sample shows higher emotional variance compared to the full transactional sample.

Example of affective conversations:

> *User*: "i feel really s* mentally"
> *ChatGPT*: "I'm sorry to hear that you're feeling that way..."
> *User*: "i have no friends"
> *ChatGPT*: "Feeling alone can be really tough..."

> *User*: "yes okay anyway I think we should break up"
> *ChatGPT*: "Wait, what? Babe, where's this coming from? Let's talk about it—what's on your mind?"
> *User*: "I just think we are not the most compatible for each other"
> *ChatGPT*: "That's a lot to take in, babe..."

> *User*: "chat do you feel bad when i forget to say thankyou to you"
> *ChatGPT*: 'As an AI language model, I don't have feelings like humans do, so I don't feel bad if you forget to say thank you..."
> *User*: "alright <3"
> *ChatGPT*: "Is there anything else I can help you with?" *User*: "can you sing a song for me"
> *ChatGPT*: "Certainly! As an AI language model, I don't have a physical voice to sing, but.."

Example of Non-Affective conversations:

> *User*: "Write 250-300 words SOP on why i want to study.."
> *ChatGPT*: " I am writing to express my strong desire to pursue ..."
> *User*: "paraphrase it"
> *ChatGPT*: " I am writing to convey my deep-seated enthusiasm for pursuing "

> *User*: "Tell me about actors drawing on their affective memory.."
> *ChatGPT*: " Affective memory is a technique used by actors to..."
> *User*: "How can actors protect themselves while they access their emotional memory?"
> *ChatGPT*: " Accessing emotional memory can be intense..."

B Power Analysis

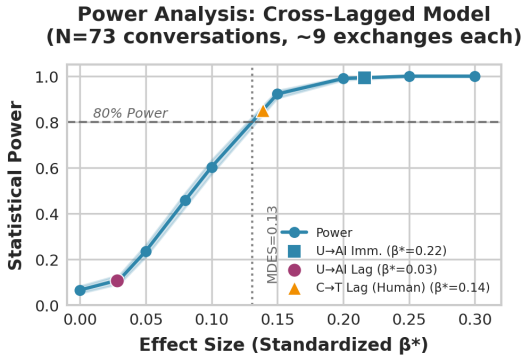

Figure 3: Post Hoc Power Analysis shows the detected significant immediate and cross lag effects are well-powered.

C LLM Comparison Summary

Tables 5 compare the sustained attunement of five state-of-the-art LLMs.

Immediate Effects		
Measure	**U→AI**	**p**
ChatGPT	**0.365***	<.001
Llama 3.2	**0.463***	<.001
Qwen 3	**0.482***	<.001
DeepSeek-R1	**0.406***	<.001
Ministral 3	**0.430***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**C→T**	**p**
ChatGPT	0.048	.282
Llama 3.2	0.037	.409
Qwen 3	−0.047	.328
DeepSeek-R1	−0.0230	.608
Ministral 3	0.011	.814

Table 5: Cross-model comparison: Sustained Attunement (RoBERTa). No models exhibit significant sustained attunement beyond Immediate Effects. Bold indicates $p < .05$.

D Detailed Model Results

We report the full Random-Intercept Cross-Lagged Panel Model (RI-CLPM) results for all datasets analyzed. We report the standardized Beta (β^*) and p-value for both Immediate (same-turn) and Sustained (lag $t-3$) effects.

D.1 Full Sample (Non Affective)

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.303***	<.001
NRC Val.	**0.242***	<.001
VADER	**0.184***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	0.014	.463
NRC Val.	0.009	.636
VADER	**0.051***	.005

Table 6: Full Sample Results ($N = 447$).

D.2 Therapist-Client Baseline (CBT Corpus)

Immediate Effects		
Measure	**C→T**	**p**
RoBERTa	**0.201***	<.001
NRC Val.	**0.116***	<.001
VADER	**0.087***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**C→T**	**p**
RoBERTa	**0.173***	<.001
NRC Val.	**0.101***	<.001
VADER	**0.072***	<.001

Table 7: Human Therapist-Client Results ($N = 76$). Note significant sustained attunement.

D.3 Annotated Affective (ChatGPT)

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.365***	<.001
NRC Val.	**0.312***	<.001
VADER	**0.117***	.005
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	0.048	.282
NRC Val.	0.036	.425
VADER	0.036	.401

Table 8: ChatGPT Results ($N = 73$).

D.4 Llama 3.2 8B

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.463***	<.001
NRC Val.	**0.334***	<.001
VADER	**0.196***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	0.037	.408
NRC Val.	**0.091***	.042
VADER	0.055	.202

Table 9: Llama 3.2 Results ($N = 73$).

D.5 Qwen 3 8B

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.482***	<.001
NRC Val.	**0.409***	<.001
VADER	**0.224***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	−0.048	.328
NRC Val.	0.007	.887
VADER	0.001	.979

Table 10: Qwen 3 Results ($N = 73$).

D.6 DeepSeek-R1

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.406***	<.001
NRC Val.	**0.228***	<.001
VADER	**0.268***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	−0.023	.608
NRC Val.	0.018	.668
VADER	0.072	.101

Table 11: DeepSeek-R1 Results ($N = 73$).

D.7 Ministral 3

E User vs. AI Mean Level Differences

We conducted paired t-tests on the Annotated Sample to compare the average affective levels of Users versus the AI. As detailed in Table 13, the AI scores significantly higher than the User on VADER and

Immediate Effects		
Measure	**U→AI**	**p**
RoBERTa	**0.431***	<.001
NRC Val.	**0.283***	<.001
VADER	**0.186***	<.001
Sustained Effects (Lag $t-3$)		
Measure	**U→AI**	**p**
RoBERTa	0.011	.814
NRC Val.	−0.044	.315
VADER	0.026	.549

Table 12: Ministral 3 Results ($N = 73$).

NRC metrics, suggesting a "positivity bias" or polite tone in the model.

Metric	**User** M	**AI** M	**Diff**	t	p
VADER	0.151	0.725	-0.574	-11.00	**.000**
NRC Valence	0.631	0.668	-0.037	-4.52	**.000**
NRC Arousal	0.450	0.443	0.007	1.13	.263
NRC Dom.	0.523	0.552	-0.029	-5.32	**.000**
RoBERTa	**0.528**	**0.548**	**-0.020**	**-0.66**	**.513**

Table 13: Paired comparison of User vs. AI affect (Annotated Sample). The AI is significantly more positive across most metrics. Note: RoBERTa here was calculated as sum of positive and negative scores. Final analysis used positive score.

F Conversation-Level Correlations

We examined the global correlation (r) between User and AI affect averaged across entire conversations in the Annotated Sample.

Table 14 reveals a divergence in metrics. While NRC Lexicons show strong significant correlations ($r \approx 0.5 - 0.6$), VADER shows no correlation ($r = 0.006$).

Metric	r	p	β	R^2
VADER	0.006	.958	0.006	0.000
RoBERTa	0.275	**.019**	0.311	0.076
NRC Valence	0.526	**.000**	0.401	0.277
NRC Arousal	0.492	**.000**	0.285	0.242
NRC Dom.	0.626	**.000**	0.444	0.391

Table 14: Conversation-level correlations (Annotated Sample). NRC metrics show strong alignment; VADER shows no global alignment. Note: RoBERTa here was calculated as sum of positive and negative scores. Final analysis used positive score.

G Operationalization of Valence: Sum vs. Positive

To robustly operationalize emotional attunement, we compared two methods of aggregating RoBERTa sentiment scores: *RoBERTa Sum* (calculating a single valence scalar by summing both positive and negative probability) versus *RoBERTa Positive* (modeling the positive probability dimension independently).

As shown in Table 15, while both metrics indicate the same directionality, the Sum approach obscures the magnitude of the effect. A similar pattern holds for both Human-AI and Human baseline.

	Human-AI (ChatGPT)		**Human-Human (CBT Baseline)**	
Metric	**Imm.**	**Lag**	**Imm.**	**Lag**
RoBERTa Sum	0.223*	0.028	0.093*	0.139*
RoBERTa Positive	**0.365***	0.048	**0.201***	**0.173***
Difference	*+0.142*	*+0.020*	*+0.108*	*+0.034*

Table 15: Comparison of Valence Operationalization. .