LLM-MC-Affect: LLM-Based Monte Carlo Modeling of Affective Trajectories and Latent Ambiguity for Interpersonal Dynamic Insight

Yu-Zheng Lin[1]*, Bono Po-Jen Shih[2], John Paul Martin Encinas[1], Elizabeth Victoria Abraham Achom[1], Karan Himanshu Patel[1], Jesus Horacio Pacheco[3], Sicong Shao[4], Jyotikrishna Dass[1], Soheil Salehi[1], Pratik Satam[1]

[1]University of Arizona [2]Pennsylvania State University
[3]Universidad de Sonora [4]University of North Dakota
***Correspondence:** yuzhenglin@arizona.edu

Abstract

Emotional coordination is a fundamental aspect of human interaction that shapes how relational meaning unfolds across interaction. While text-based affect inference has become increasingly feasible, prior approaches often treat sentiment as a deterministic point estimate for individual speakers, failing to capture the inherent subjectivity, latent ambiguity, and sequential coupling found in mutual exchanges. We introduce LLM-MC-Affect, a probabilistic framework that characterizes emotion not as a static label, but as a continuous latent probability distribution defined over an affective space. By leveraging stochastic LLM decoding and Monte Carlo (MC) estimation, the methodology approximates these distributions to derive high-fidelity sentiment trajectories that explicitly quantify both central affective tendencies and perceptual ambiguity. These trajectories enable a structured analysis of interpersonal coupling through sequential cross-correlation and slope-based indicators, identifying leading or lagging influences between interlocutors. To validate the interpretive capacity of this approach, we utilize teacher-student instructional dialogues as a representative case study, where our quantitative indicators successfully distill high-level interaction insights such as effective scaffolding. This work establishes a scalable and deployable pathway for understanding interpersonal dynamics, offering a generalizable solution that extends beyond education to broader social and behavioral research.

1 Introduction

Emotional coordination is a core property of human interaction that shapes how relational meaning is constructed in real time, particularly in instructional settings where rapport, engagement, and frustration can propagate across conversational turns. One related concept is affective synchrony (Wood et al., 2021), which broadly refers to the alignment or co-evolution of affective states among interacting individuals over time. Empirical studies of affective synchrony have traditionally relied on biometric coupling signals (Qi et al., 2024; Nguyen et al., 2021; Bevilacqua et al., 2019), such as autonomic responses or neural alignment, which provide high sequential fidelity but require specialized instrumentation and tightly controlled protocols that are difficult to scale to large, naturalistic learning environments. In contrast, conversational text is ubiquitous and inherently longitudinal, and prior studies have demonstrated that it encodes a recoverable affective structure that can be summarized as sentiment trajectories (Liu and Chen, 2024). However, many text-based sentiment analysis pipelines continue to represent affect as a deterministic point estimate at the level of individual utterances or speakers (Liu et al., 2023a; Gao et al., 2022). Although effective for coarse-grained evaluation, this representation tends to compress subjective variability, offers limited visibility into latent affective ambiguity, and often treats affect as an individual attribute rather than an explicitly coupled dyadic process in interaction analysis (Tsakalidis et al., 2022; Wemmer et al., 2024).

This paper introduces LLM-MC-Affect, a probabilistic framework that treats utterance-level emotion not as a fixed label, but as a latent affective distribution over a continuous affective space. Recent advances in large language models have demonstrated strong zero-shot affective perception capabilities, enabling them to infer nuanced emotional states from natural language without task-specific fine-tuning (Lin et al., 2025; Zhang et al., 2025; Liu et al., 2024). Rather than approximating this latent distribution through labels collected from multiple human raters, which can be costly and difficult to scale, we

*Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27984–28000

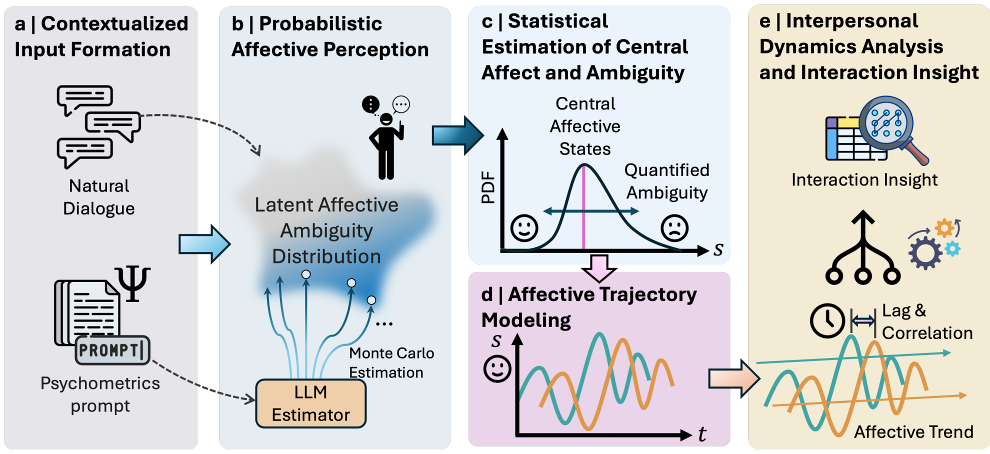


Figure 1: Overview of the LLM-MC-Affect framework for probabilistic affect modeling and dyadic interaction analysis. (a) Contextualized input formation integrates natural dialogue with a psychometric prompt to guide affect estimation. (b) Probabilistic affective perception leverages stochastic LLM decoding with Monte Carlo sampling to estimate a latent distribution over affective states. (c) Statistical estimation derives the central affective tendency and quantifies ambiguity from the inferred distribution. (d) Affective trajectory modeling tracks the temporal evolution of affective states, capturing both trends and uncertainty over time. (e) Interpersonal dynamics analysis jointly interprets lag, correlation, and trajectory trends to extract interpretable dyadic interaction patterns, linking low-level probabilistic affect measurements to higher-level interaction insights.

leverage the stochastic nature of large language models under non-zero decoding temperature (Renze, 2024; Liu et al., 2023b; Salinas et al., 2025) and perform LLM-based Monte Carlo (MC) estimation: repeated, independently sampled affective evaluations are treated as draws from an implicit model-induced proxy for the latent distribution, enabling estimation of both (1) the central affective tendency and (2) dispersion as a quantitative indicator of latent affective ambiguity. This yields stable, uncertainty-aware sentiment trajectories that explicitly retain the variance information discarded by deterministic scoring.

Building on these trajectories, LLM-MC-Affect characterizes interaction-level synchrony by modeling how two affective signals align across conversational turns. Specifically, we compute a normalized sequential cross-correlation function between teacher and student trajectories, estimate an influence lag that indicates leading or lagging influence, and combine these sequential alignment measures with slope-based trend indicators that summarize long-horizon affective development. By jointly interpreting lag, correlation sign, and trajectory slope coupling, the framework yields interpretable dyadic patterns (e.g., teacher-led positive scaffolding, synchronous decline, or dynamic compensation) derived from low-level probabilistic affect measurements, thereby linking fine-grained affective perception to higher-level interpersonal interpretation.

We validate the method on teacher-student instructional dialogues as a representative case study, because education provides a structured yet socially rich domain in which affective coupling is meaningful and can be qualitatively examined. Under a unified, prompt-defined psychometric rubric, the proposed approach supports zero-shot affect estimation without task-specific fine-tuning, while enabling controlled sensitivity analyses over decoding stochasticity and cross-model comparisons that reveal differences in affective resolution and ambiguity across LLMs. Taken together, the results demonstrate that affective synchrony and directional influence can be derived from linguistic affect alone, providing a deployable pathway for affect-aware instructional analytics and, more broadly, for scalable research on interpersonal dynamics without biometric sensing.

Our contributions are summarized as follows: (1) We propose LLM-MC-Affect, a probabilistic pipeline that uses stochastic LLM decoding with Monte Carlo estimation to model utterance-level affect as a latent distribution and quantify ambiguity via variance, rather than relying on deterministic sentiment labels. (2) We introduce an interaction analysis layer that extracts interpretable dyadic indicators, including sequential cross-correlation, estimated influence lag, and slope-based trajectory signals, yielding a structured typology of interpersonal affective dynamics. (3) We demonstrate feasibility in instructional dialogues, showing that uncertainty-aware trajectories remain robust under decoding stochasticity and can support interpretable affective synchrony insights in a scalable, text-only setting.

2 Background

2.1 Statistical Formulation of Affective Ambiguity

Affective modeling in natural language interaction must account for the inherent subjectivity and multifaceted ambiguity found in emotional expressions (Davani et al., 2022; Kuhn et al., 2023). Drawing upon the conceptual foundation established by Wu et al. (2025), we characterize the emotion associated with a specific utterance u_t not as a deterministic point estimate, but as a latent probability distribution $p(e|u_t)$ defined over a continuous affective space $\mathcal{E}$. This probabilistic paradigm acknowledges that the actual underlying continuous distribution $p(e|u_t)$ is a latent construct that remains inaccessible for direct observation. Consequently, conventional methodologies approximate this hidden continuous density by collecting a finite set of independent inferences, such as human ratings, where each individual's discrete score is a sample discretized from the continuous distribution $p(e|u_t)$. This statistical premise establishes that inter-rater variability and perceived ambiguity are not merely noise to be averaged out; instead, they represent essential emotional nuances that characterize the underlying affective state.

2.2 Emotion as a Temporally Coupled Interpersonal System

Emotion in social interaction is increasingly understood as a temporally structured process that emerges through ongoing interpersonal coordination rather than as an isolated individual attribute. From this perspective, Butler (2011) conceptualizes emotions as Temporal Interpersonal Emotion Systems (TIES), in which social partners function as a dynamically coupled, self-organizing system. In this framework, emotional subcomponents interact across individual boundaries over time, giving rise to integrated dyadic patterns. From an emotion dynamics perspective, cross-correlation functions provide an effective and well-established tool for capturing time-lagged relationships between affective processes, enabling the characterization of temporal coordination and relative lag in dyadic interactions after accounting for overall trends and level effects (Feldman, 2003; Krone et al., 2018; Main et al., 2016).

3 Methodology

We introduce LLM-MC-Affect, a probabilistic framework for modeling affective trajectories from conversational text using LLMs. The method treats utterance-level emotion not as a deterministic label, but as a latent distribution estimated via Monte Carlo sampling under stochastic LLM decoding, enabling robust affective trajectory construction while explicitly capturing affective ambiguity. These trajectories form a continuous sequential representation that supports downstream interpersonal analysis, including sequential alignment, directional influence, and long-term affective progression. By aggregating low-level probabilistic sentiment estimates into interpretable interaction indicators, LLM-MC-Affect bridges fine-grained affect perception with higher-level explanatory insights. The overall processing pipeline is illustrated in Figure 1.

3.1 LLM-Based Affective Trajectory Modeling

This subsection formulates affective trajectory modeling as a probabilistic inference problem over latent emotion distributions. Given a dialogue context and a psychometric prompt, we approximate the latent affective distribution at each utterance via Monte Carlo sampling under stochastic LLM decoding. The resulting samples are used to estimate the mean affective state and variance, capturing central tendency and ambiguity, respectively. These estimates are subsequently standardized and assembled into longitudinal trajectories for each interlocutor. This pipeline enables uncertainty-aware trajectory representations that support downstream analysis of interpersonal dynamics.

3.1.1 Probabilistic Affective Perception: Monte Carlo Estimation via Stochastic LLM Decoding

Our methodology leverages the stochastic properties of Large Language Models (LLMs) to approximate the parameters of the latent affective distribution. We formalize the inference process as a mapping $\mathcal{M}$ from the textual input space to the sentiment trajectory space. The model is conditioned on a domain-specific psychometric prompt ρ_{emo} and the conversational context window $\mathcal{W} = ((u_t^{\text{T}}, u_t^{\text{S}}))_{t=1}^{N}$, where u_t^{T} and u_t^{S} represent the utterances from the teacher and the student at turn t, respectively. The model input $\mathcal{I}$ is constructed via the sequence concatenation of the instruction and the dialogue history: $\mathcal{I} = [\rho_{\text{emo}}; \mathcal{W}]$. To estimate the mean affective state while accounting for linguistic subjectivity, we employ Monte Carlo estimation via stochastic LLM decoding. Specifically, for each turn and each interlocutor $(\cdot) \in \{\text{T}, \text{S}\}$, we perform K independent inference trials under non-zero temperature settings, yielding a set of affective state estimates $\{\hat{s}_{t,k}^{(\cdot)}\}_{k=1}^{K}$. This framework assumes that the LLM response distribution, when conditioned on ρ_{emo}, serves as a computational proxy for the inaccessible latent distribution $p(e|u_t)$. Consequently, each trial provides a realization of the latent affective state, denoted as: $\hat{s}_{t,k}^{(\cdot)} \sim \mathcal{M}(\mathcal{I})$, which facilitates the quantification of affective fluctuations and ambiguity without relying on a single, potentially biased, deterministic output.

3.1.2 Estimation of Mean Affective State and Ambiguity

The primary objective of the sampling process is to estimate the fundamental statistical moments of the underlying distribution. In accordance with the framework of Wu et al. (2025), the mean of the distribution represents the central emotional tendency, while the variance quantifies the degree of affective ambiguity inherent in the utterance. We define the aggregated raw affective signal s_t as the sample mean derived from the K trials, serving as an empirical estimate of the mean affective state: $s_t^{(\cdot)} = \frac{1}{K}\sum_{k=1}^{K} \hat{s}_{t,k}^{(\cdot)}$. Furthermore, the variance across these trials, σ_t^2, is utilized to represent the uncertainty in the model's affective inference: $\sigma_t^{2(\cdot)} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{s}_{t,k}^{(\cdot)} - s_t^{(\cdot)})^2$. This variance captures the aleatoric uncertainty associated with linguistic ambiguity.

3.1.3 Sentiment Scoring Mechanism, Polarity Mapping, and Trajectory Modeling

The sentiment quantification process follows the scoring rules defined in ρ_{emo} (detailed in Appendix A), which utilizes a fixed numerical range from 0 to 5 with 0.5 intervals. This rubric establishes 2.5 as the neutral baseline for procedural or factual statements; meanwhile, positive emotional cues lead to a numerical decrease toward 0 and negative cues trigger an increase toward 5. By employing a non-negative numerical progression instead of signed values during the initial evaluation, the prompt is designed to minimize the risk of logical or sign-related confusion during the model inference process, which serves to enhance the stability and consistency of the generated scores. Under this schema, a raw score of 0.0 signifies the most positive affective state, such as empathy or successful resolution, whereas 5.0 represents the most negative state, including explicit conflict or frustration. To align these outputs with affective computing conventions in which positive values represent positive valence, we perform a polarity mapping on the raw results. This transformation maps the averaged raw scores s_t to a centered interval $[-1, 1]$ through the following linear operation: $\tilde{s}_t^{(\cdot)} = 1 - 2\,(s_t^{(\cdot)}/5)$. To maintain statistical consistency, the associated variance is scaled by the square of the mapping coefficient, resulting in $\tilde{\sigma}_t^2 = (\frac{2}{5})^2\sigma_t^2$. Through this mapping, the value $\tilde{s}_t \to +1$ indicates a strong positive state and $\tilde{s}_t \to -1$ indicates a strong negative state, while $\tilde{s}_t = 0$ remains the neutral point. The transformed variance $\tilde{\sigma}_t^2$ preserves the quantification of affective ambiguity within the standardized space. The final standardized sentiment trajectories are defined as vectors $\tilde{\mathbf{S}}^{\text{T}}, \tilde{\mathbf{S}}^{\text{S}} \in [-1, 1]^N$, where each vector $\tilde{\mathbf{S}}^{(\cdot)} = (\tilde{s}_1^{(\cdot)}, \ldots, \tilde{s}_N^{(\cdot)})^\top$ captures the continuous evolution of sentiment throughout the dialogue.

3.2 Interpersonal Affective Dynamics and Interaction Modeling

Building on the affective trajectories, we model interpersonal dynamics through complementary temporal and trend-based analyses. Sequential alignment and directional influence are quantified via cross-correlation, while long-term affective development is summarized using slope-based indicators. Taken together, these measures serve as interaction indicators that support the interpretation of latent interpersonal behaviors and potential dyadic interaction patterns.

3.2.1 Sequential Cross-Correlation Analysis

Our application of cross-correlation for affective time-series analysis is inspired by the methodology from O'Connor et al. (2010), who utilized this technique to link sentiment signals extracted from text streams to longitudinal public opinion data. By adapting this method to the interpersonal dynamics analysis, we can effectively synchronize affective trajectories and identify leading or lagging influences across discrete dialogue turns.

Building on the standardized trajectories $\tilde{\mathbf{S}}^{\mathrm{T}}$ and $\tilde{\mathbf{S}}^{\mathrm{S}}$, we employ discrete cross-correlation to quantify the sequential alignment between the interlocutors throughout the dialogue sequence. We define the normalized cross-correlation function (NCCF), $R_{\mathrm{TS}}(L)$, as: $R_{\mathrm{TS}}(L) = \frac{\sum_{t \in \mathcal{T}_L} (\tilde{s}_t^{\mathrm{T}} - \bar{s}^{\mathrm{T}})(\tilde{s}_{t+L}^{\mathrm{S}} - \bar{s}^{\mathrm{S}})}{\sqrt{\sum_{t \in \mathcal{T}_L} (\tilde{s}_t^{\mathrm{T}} - \bar{s}^{\mathrm{T}})^2 \sum_{t \in \mathcal{T}_L} (\tilde{s}_{t+L}^{\mathrm{S}} - \bar{s}^{\mathrm{S}})^2}}$, where $L \in \mathbb{Z} \cap [-3, 3]$ represents the conversational lag in turns and $\bar{s}^{(\cdot)}$ denotes the sequential mean of the affective scores across the interaction. To ensure mathematical validity over the finite dialogue length N, the summation is performed over the set $\mathcal{T}_L = \{t \in \mathbb{Z} \mid 1 \leq t \leq N, 1 \leq t + L \leq N\}$, which defines the valid overlapping indices for a given lag L. The estimated influence lag $L^* = \arg\max_L |R_{\mathrm{TS}}(L)|$ identifies the dominant point of sequential phase alignment. Within this framework, $L^* \geq 0$ suggests a teacher-leading influence; conversely, $L^* < 0$ reflects student-driven adjustment.

3.2.2 Slope-Based Development Indicator

To characterize long-term affective development beyond local turn-to-turn alignment, we summarize each standardized sentiment trajectory by its linear trend over time. For each interlocutor $(\cdot) \in \{\mathrm{T}, \mathrm{S}\}$, we estimate the slope parameter $\beta_{(\cdot)}$ via least-squares regression over the dialogue turns $t \in \{1, \ldots, N\}$, where N denotes the total number of turns considered. In in-text form, the slope is given by $\beta_{(\cdot)} = \sum_{t=1}^{N} (t - \bar{t})(\tilde{s}_t^{(\cdot)} - \bar{s}^{(\cdot)}) / \sum_{t=1}^{N} (t - \bar{t})^2$, with $\bar{t} = \frac{1}{N} \sum_{t=1}^{N} t$ and $\bar{s}^{(\cdot)} = \frac{1}{N} \sum_{t=1}^{N} \tilde{s}_t^{(\cdot)}$. Under the standardized polarity mapping defined in Section 3.1.3, positive values of $\tilde{s}_t$ correspond to positive affect. Accordingly, a positive slope ($\beta_{(\cdot)} > 0$) indicates overall emotional amelioration across the interaction, whereas a negative slope ($\beta_{(\cdot)} < 0$) reflects affective deterioration. This indicator captures global affective drift and complements the lag-based cross-correlation analysis.

Table 1: Joint Indicators-based Interpretation of Teacher-Student Affective Dynamics

Estimated Influence Lag (L^*)	Sign($R_{\mathrm{TS}}(L^*)$)	Slope Relation	Potential Interpretation
$L^* > 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} > 0$	Effective Scaffolding; teacher-led positive contagion
$L^* > 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} < 0$	Negative Contagion; teacher-led shared frustration
$L^* > 0$	$-$	$\beta_{\mathrm{T}} > 0, \beta_{\mathrm{S}} < 0$	Counterproductive Support; teacher regulates student frustration
$L^* < 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} > 0$	Student-driven success; student motivates teacher
$L^* < 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} < 0$	Feedback Burnout; student leads shared decline
$L^* < 0$	$-$	$\beta_{\mathrm{S}} < 0, \beta_{\mathrm{T}} > 0$	Encouragement Escalation; teacher attempts to uplift student
$L^* = 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} > 0$	Affective Synchrony; real-time mutual engagement
$L^* = 0$	$+$	both $\beta_{\mathrm{T}}, \beta_{\mathrm{S}} < 0$	Shared Fatigue; synchronous motivation decline
$L^* = 0$	$-$	opposite signs	Dynamic Compensation; real-time tension balancing

3.3 From Interpersonal Affective Trajectories to Interaction Insights: Mining Dynamics in Educational Settings

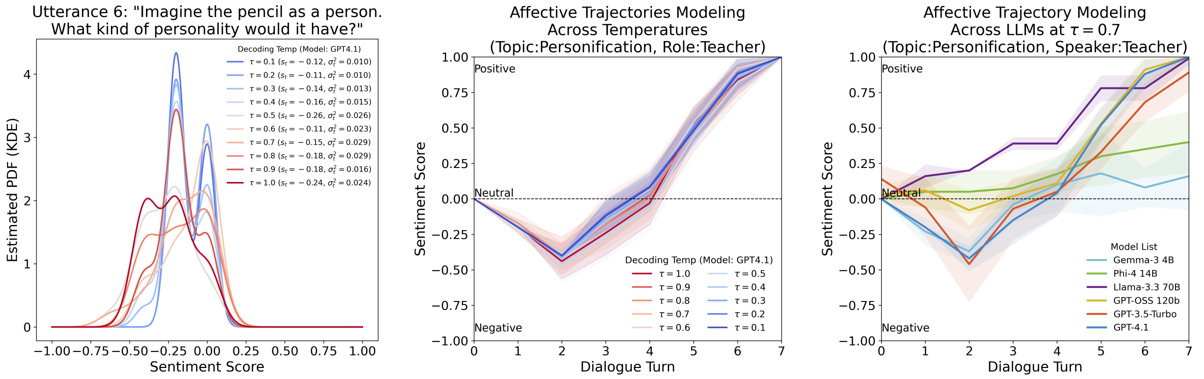
The transition from interpersonal affective trajectories to interaction insights is achieved by synthesizing the quantitative indicators derived from the turn-based sequence. This analytical framework is designed for broad application across diverse dyadic interactions; however, we utilize educational discourse as a representative case study to illustrate its interpretive capacity. By concurrently analyzing the direction of influence through the estimated influence lag L^*, the nature of emotional coupling via R_{TS}, and the long-term progression using the slopes $\beta_{(\cdot)}$, the method identifies specific interaction patterns.

These indicators jointly provide a structured view of affective co-evolution in dialogue. The estimated influence lag captures the temporal ordering of affective influence, distinguishing teacher-led dynamics from student-driven responses. At the same time, trajectory slopes summarize the long-term directional trend of the emotional climate. Together, these signals enable the identification of latent interaction patterns, such as affective synchrony or counterproductive support, without enforcing strict turn-level alignment assumptions. A detailed typology of these joint configurations and their corresponding interpretations in the learning context is summarized in Table 1. This joint indicator-based synthesis bridges low-level statistical descriptors estimated during the perception stage with higher-level, interpretable insights into teacher-student affective dynamics (Kosinski, 2024; Linnenbrink-Garcia et al., 2016; Chang, 2009; Bevilacqua et al., 2019).

Table 2: Teacher-Student Dialogue Example (*Topic - Personification*)

Background Information			
Student Preference:	Hands-on activities and real-world applications	Student Reaction:	Becomes rude when disengaged
Teacher Preference:	Experiential learning and hands-on activities	Teacher Reaction:	Insists on experiential teaching methods

Dialogue			
Turn	Index	Role	Utterance
0	0	Teacher	Today, we're going to learn about personification, a literary device that gives human characteristics to nonhuman things.
	1	Student	That sounds boring. Why do we need to know this?
1	2	Teacher	It's a technique that authors use to create vivid and engaging stories. Let's try an activity. Look around the room and find an object that you can personify.
	3	Student	This is stupid. I don't want to do this.
2	4	Teacher	I know it might seem a bit abstract at first, but it's a useful tool to understand. How about we try it with this pencil?
	5	Student	Fine, whatever.
3	6	Teacher	Imagine the pencil as a person. What kind of personality would it have?
	7	Student	I don't know. It's just a pencil.
4	8	Teacher	Think about its qualities. It's sharp, it can write, it's often used to express ideas.
	9	Student	Maybe it could be a shy writer who's afraid to share its thoughts?
5	10	Teacher	That's a great example! You're starting to get the hang of it. Now, let's write a short story where the pencil is the main character.
	11	Student	Okay, this is actually kind of fun.
6	12	Teacher	I'm glad you're enjoying it. Remember, personification helps us relate to objects and stories in a new way.
	13	Student	I think I get it now. Thanks for showing me this way.
7	14	Teacher	You're welcome. I'm happy you found it engaging.
	15	Student	[End of conversation]



(a) (b) (c)

Figure 2: Sensitivity analysis of probabilistic affective estimation and trajectory stability. (a) Probabilistic sentiment score distributions for Utterance 6 in the *Personification* topic across decoding temperatures ($\tau \in [0.1, 1.0]$), estimated via Kernel Density Estimation (KDE). (b) Convergence of teacher mean affective trajectories for the *Personification* topic, demonstrating the stability of the estimated mean affective state across the temperature range. (c) Comparative affective trajectory modeling across six LLMs at a fixed decoding temperature ($\tau = 0.7$).
***Note:** In (b) and (c), shaded uncertainty bands denote ± 1 standard deviation ($\pm\sigma$), representing the quantified affective ambiguity in the models' perceptions.

4 Experiment results and Discussion

This section evaluates the robustness of LLM-MC-Affect under varying temperatures and models, and presents case studies illustrating its ability to capture interpersonal dynamics in educational dialogues.

4.1 Experimental Setup

We evaluate our approach using Google's Education Dialogue Dataset (Shani et al., 2024), a recently introduced synthetic dataset for modeling multi-turn instructional interactions. The use of a simulated dialogue environment is motivated by well-established logistical and ethical constraints in educational research (Nebeker et al., 2017). Collecting data from real classroom settings typically requires extensive Institutional Review Board (IRB) approval and privacy concerns associated with student information. Accordingly, this study is positioned as an empirical pilot that validates the proposed affective modeling framework under controlled conditions before deployment in naturalistic educational environments. Importantly, recent studies have demonstrated that LLMs can effectively simulate complex human behaviors and social dynamics, and are increasingly adopted in computational social science and behavioral research (Westwood, 2025; Park et al., 2023; Kosinski, 2024). Such simulation-based datasets enable systematic control over participant personas and interaction variables, which is difficult to achieve in real-world classroom studies.

For evaluation, we access GPT-4.1 (snapshot 2025-04-14) and GPT-3.5-Turbo (snapshot 2024-01-25) via the OpenAI API (Achiam et al., 2023; Ouyang et al., 2022), while open-source models, including Gemma 3 4B, Llama 3.3 70B, Phi 4 14B and GPT-OSS 120B, are deployed through the AI Verde service (Team et al., 2025; Dubey et al., 2024; Abdin et al., 2024; Agarwal et al., 2025; Mithun et al., 2025). All models are evaluated in their original, non-fine-tuned configurations to ensure that the results reflect their inherent zero-shot affective perception capabilities. To ensure empirical consistency and reproducibility, a unified scoring rubric (ρ_{emo}) is applied across all experiments. For each dialogue turn, we conduct $K = 20$ independent Monte Carlo trials to estimate parameters of the latent affective distribution. In the main text, we present a detailed analysis of a representative case study, *Personification* (detailed in Table 2), to illustrate how the proposed quantitative indicators yield pedagogically meaningful insights. Additional case studies are reported in Appendix C.

4.2 Sensitivity Analysis of Temperature on Probabilistic Affective Perception

We treat the decoding temperature as a controllable stochasticity parameter and perform a sensitivity analysis using GPT-4.1 as the primary model (Liu et al., 2024) by examining how temperature variations affect the estimated affective distributions and trajectory stability with ρ_{emo}. This approach is grounded in the statistical formulation that characterizes the emotion of an utterance not as a deterministic point estimate, but as a latent probability distribution $p(e|u_t)$. By leveraging the stochastic properties of LLMs at non-zero temperature, we can approximate this latent distribution via Monte Carlo sampling.

4.2.1 Impact on Distributional Characteristics

The decoding temperature τ serves as a critical modulator of the model's exploratory behavior during sentiment quantification. As illustrated in Figure 2a (generated via GPT-4.1), the Kernel Density Estimation (KDE) for Utterance 6 - *"Imagine the pencil as a person. What kind of personality would it have?"*, lower temperature settings reveal a complex, bimodal perception of pedagogical intent. At $\tau = 0.1$, the distribution is characterized by two distinct, sharp peaks centered near -0.20 and 0.00, with a recorded sample mean $\tilde{s}_t = -0.12$ and a minimal variance $\tilde{\sigma}_t^2 = 0.010$. These sharp peaks indicate a pseudo-deterministic interpretation in which the model identifies two divergent yet high-confidence affective interpretations of the prompt. As the temperature approaches 1.0, these distinct peaks merge into a broader, more continuous distribution. Higher temperature does not necessarily imply a more accurate affective judgment; rather, it enables the LLM to explore a broader set of plausible interpretations of the utterance, thereby making potential affective ambiguity more visible in the resulting sample distribution. While the variance captures the aleatoric uncertainty associated with linguistic ambiguity, the distribution exhibits reduced concentration, with the variance increasing to $\tilde{\sigma}_t^2 = 0.024$ at $\tau = 1.0$. This behavior aligns with our statistical formulation, where ambiguity is treated as an essential emotional nuance rather than mere noise. By leveraging higher stochasticity, this method reveals the hidden density of the affective variable, providing a quantitative proxy for interpretative subjectivity that would be lost in deterministic settings. For completeness, the utterance-level mean affective states and variances across all temperature settings are reported in Appendix B.

4.2.2 Robustness and Reliability of the Mean Affective State

A fundamental finding of this sensitivity analysis is the robustness of the mean affective state despite significant shifts in distribution. While the variance σ_t^2 effectively doubles as τ scales from 0.1 to 1.0, the sample mean s_t remains remarkably stable and concentrated within a narrow band. For Utterance 6 in the personification topic dialogue, the mean fluctuates only marginally between -0.11 and -0.26 across the entire temperature evaluation range. The reliability of these mean estimates is further demonstrated at the level of affective trajectory modeling. In the Figure 2b (generated via GPT-4.1) for the Personification topic, the teacher's affective trajectories across all τ values exhibit a high degree of convergence. All plotted lines consistently track the same emotional progression, initially dipping toward negative valence at Turn 2 before rising steadily to a strong positive state by Turn 7. This convergence confirms that the framework can filter out stochastic sampling noise while preserving the underlying affective signal. This stability ensures that the derived interpersonal metrics in downstream analyses, such as the estimated influence lag (L^*) and Cross-Correlation (R_{TS}), remain reliable indicators of interaction dynamics despite minor sampling variations.

4.2.3 Cross-Model Behavior Evaluation

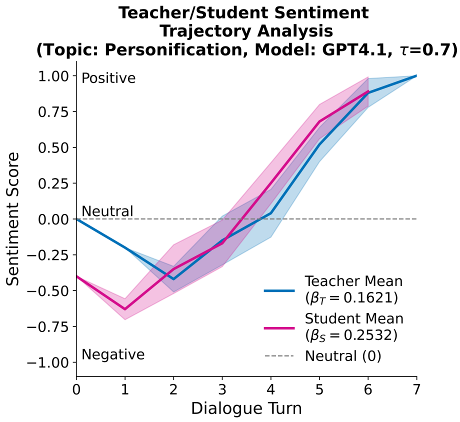
To assess cross-model behavior, we compare affective trajectories produced by six LLMs on the Personification dialogue at a fixed decoding temperature of $\tau = 0.7$. This setting balances mean stability and perceptual ambiguity, enabling meaningful Monte Carlo variance estimation without degrading trajectory convergence. All models share the same psychometric prompt (ρ_{emo}) and operate in a zero-shot configuration. Results are reported in Figure 2c. The result shows that `GPT-4.1` and `GPT-3.5-Turbo` exhibit the strongest structural consistency, both capturing the characteristic V-shaped affective progression with an early negative dip followed by recovery. However, `GPT-3.5-Turbo` exhibits substantially higher affective ambiguity, as indicated by a wider uncertainty band ($\pm\sigma$), suggesting greater stochastic volatility and a less confident estimate of the central affective state. In contrast, applying the same analysis to open-source models exposes pronounced divergences in affective resolution, revealing systematic biases that limit their ability to capture fine-grained emotional dynamics:

- **Extreme Positivity Bias (`Llama 3.3 70B`):** The Llama 3.3 model demonstrates a near-total failure in identifying the negative emotional shift present in the early dialogue. Its trajectory follows a continuous, near-linear upward trend from the neutral baseline, failing to detect any dip at Turn 2. The Llama 3.3's insensitivity to negative pedagogical cues is likely a result of alignment for harmlessness, a process designed to ensure models produce helpful and harmless responses (Dubey et al., 2024; Mozikov et al., 2024).

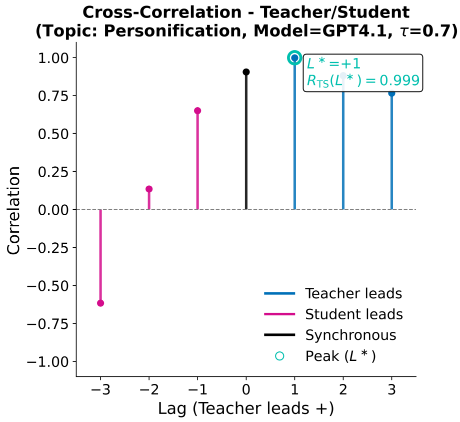
- **Performance Gaps and Conservative Biases (`GPT-OSS 120B`, `Gemma 3 4B`, and `Phi-4 14B`):** Across models outside the GPT family, a consistent pattern of conservative affective estimation is observed, manifesting as systematic overestimation of affective states during negative pedagogical turns. Despite its large parameter scale, `GPT-OSS 120B` exhibits a markedly shallow emotional dip during the Turn 1-2 transition, substantially weaker than that of the GPT series. This conservative response is consistent with a cognitive appraisal bias (Tak et al., 2025). Similar but more pronounced limitations appear in smaller architectures (Liu et al., 2024). `Phi-4 14B` displays a mild positive bias; although its uncertainty band partially overlaps the negative region at Turn 2, its mean trajectory remains elevated and its subsequent recovery is capped at approximately 0.40. In addition, `Gemma 3 4B` exhibits a recovery bottleneck: while it detects the initial dip, it fails to model the rebound phase, with its trajectory plateauing near the neutral baseline ($\sim$0.15). Collectively, these patterns indicate that both large and small non-GPT models tend to underestimate affective turning points, albeit for different architectural or alignment-related reasons.

4.3 Interpreting Interpersonal Dynamics and Interaction Insights

Based on our cross-model stability analysis, `GPT-4.1` demonstrates a superior capacity to effectively reflect subtle fluctuations in dialogue emotion. It maintains high resolution in affective trajectory modeling, even in scenarios where other architectures exhibit conservative patterns or systematic biases. This empirical observation aligns with prior work, such as Liu et al. (2024), which demonstrates that `GPT-4` outperforms other models on complex emotional recognition tasks. Consequently, we proceed with our interactional analysis using the affective scores derived from `GPT-4.1` at a fixed decoding temperature of $\tau = 0.7$. This temperature setting serves as a practical balance point that preserves the central emotional tendency while providing sufficient stochastic variance to represent latent affective ambiguity. Using this representation, we examine the interpersonal dynamics and emotional shifts within the instructional context, with the resulting affective trajectories and NCCF analysis shown in Figure 3. Synthesizing the sequential and directional metrics derived from our framework, we categorize the interaction for the *"Personification"* topic as **Effective Scaffolding**. This classification, based on the definitions in Table 1, reflects a scenario where the teacher actively drives a positive emotional climate. NCCF analysis reveals that the



(a)



(b)

Figure 3: Illustration of affective trajectory modeling and interpersonal dynamics analysis for the *Personification* dialogue using LLM-MC-Affect (`GPT-4.1`, $\tau = 0.7$). (a) Estimated teacher and student affective sentiment trajectories modeling, where solid lines denote mean affective states and shaded regions indicate uncertainty bands ($\pm\sigma$). Slopes: $\beta_T = 0.1621, \beta_S = 0.2532$ (b) NCCF analysis between teacher and student trajectories across conversational lags ($L^* = +1$, $R_{TS}(L^*) = 0.999$).

teacher acts as the leading signal, with an estimated influence delay of $L^* = +1$ yielding a near-perfect correlation of $R_{TS} = 0.999$. This indicates that the student's affective state is highly predictable based on the teacher's emotional output from the preceding turn, confirming the teacher's role as the primary driver of interpersonal dynamics.

4.3.1 Alignment with Persona Preferences and Reactions in Setting of *Personification* Topic

The quantitative metrics align remarkably well with the specific persona traits and reaction settings defined for this interaction:

- **Teacher - Insistence on Experiential Methods**: The teacher's preference for *experiential learning* and their reaction setting of *insisting on experiential methods* are reflected in the proactive leading role ($L^* = +1$). This sequential dominance shows the teacher successfully steering the dialogue toward their preferred instructional style.

- **Student - Hands-on Preference and Rude Reactions**: The student's preference for *hands-on activities* and their tendency to *become rude when disengaged* serve as critical validation points. The near-perfect correlation and the positive student slope confirm that the student's preferences were met through the personification exercise, effectively preventing a rude reaction.

5 Conclusion

This work presented LLM-MC-Affect, a probabilistic and ambiguity-aware framework for affective perception in conversational text that models utterance-level affect as a latent affective distribution rather than a deterministic label. By combining stochastic LLM decoding with Monte Carlo sampling, the framework estimates both central affective tendency and perceptual ambiguity in natural language. This approach enables the construction of stable, uncertainty-aware affective trajectories that remain robust under decoding stochasticity and avoid reliance on supervised annotation or biometric sensing.

Building on these trajectories, we introduce an interaction-level analysis layer that examines affective dynamics through a signal-analytic perspective, integrating sequential cross-correlation, influence lag estimation, and slope-based trend indicators to characterize sequential alignment, directional influence, and long-term affective progression between interlocutors. Empirical results demonstrate that the proposed framework reliably preserves trajectory stability under decoding stochasticity while explicitly exposing affective ambiguity, and that, when instantiated with affect-sensitive LLMs, it yields interpretable interaction patterns aligned with pedagogical structure and dialogue context. By bridging probabilistic affect modeling with interaction-level explanation, LLM-MC-Affect provides a scalable and deployable pathway toward affect-aware analytics for adaptive systems in educational and broader interpersonal settings.

Limitations

The proposed framework interprets variance observed across stochastic LLM decoding runs as a proxy for affective uncertainty. While the use of a highly controlled simulated instructional dataset enables systematic control of dialogue structure and decoding conditions, it does not entirely eliminate uncertainty introduced by the language models themselves. In practice, the observed variance may conflate multiple sources of uncertainty, including epistemic effects arising from model biases, prompt sensitivity, and decoding stochasticity, as well as any ambiguity inherent in the input utterances. Given the opaque internal representations of contemporary LLMs, these factors cannot be isolated or measured directly. Consequently, the Monte Carlo variance in this study is treated as an estimated signal rather than as a definitive measure of human-perceived affective ambiguity. It is designed to support, rather than replace, qualitative interpretation of the resulting scores and trajectories.

This work is evaluated on a simulated educational dialogue dataset in which interactions are generated under controlled conditions. While this setting enables systematic and replicable analysis, it does not fully reflect the complexity and variability of real-world human interactions. In practical deployment, especially in educational settings, biases in large language models may lead to overly optimistic or inaccurate interpretations of students' affective states. Such misestimations could reduce the effectiveness of the proposed framework and may disadvantage vulnerable or underrepresented students by failing to identify their actual needs. Accordingly, future work will focus on evaluation in real educational environments, systematic bias auditing across diverse learner populations, and the incorporation of human oversight mechanisms to improve reliability and fairness in deployment.

Moreover, cross-correlation should not be interpreted as evidence of a causal relationship between the teacher's and student's affective trajectories. Consistent with prior time-series sentiment studies, correlation-based lag indicators are intended to capture patterns of sequential alignment rather than causal mechanisms. This limitation is further affected by the relatively short length of the evaluated dialogues, making the statistical robustness of the results and their generalizability to longer interactions uncertain. For longer conversations, future work may incorporate sliding-window or weighted sequential analyses to capture locally stable affective dynamics while mitigating the influence of global correlations.

Finally, the framework incurs substantial computational cost due to repeated stochastic decoding. The present study benchmarks performance using general-purpose LLMs to assess feasibility across model families; however, this choice may not be optimal in terms of efficiency. Future work may explore task-specific affective models or hybrid architectures to achieve a more favorable trade-off between affective sensitivity, interpretability, and computational efficiency.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF) under research project 2335046, and the OpenAI Researcher Access Program 0000011862. We gratefully recognize the AI Verde team at the University of Arizona Data Science Institute for their support and resources, which enabled the use of open-source LLMs in this study.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Dana Bevilacqua, Ido Davidesco, Lu Wan, Kim Chaloner, Jess Rowland, Mingzhou Ding, David Poeppel, and Suzanne Dikker. 2019. Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of cognitive neuroscience*, 31(3):401–411.

Emily A Butler. 2011. Temporal interpersonal emotion systems: The "ties" that form relationships.

Personality and Social Psychology Review, 15(4):367–393.

Mei-Lin Chang. 2009. An appraisal perspective of teacher burnout: Examining the emotional work of teachers. *Educational psychology review*, 21(3):193–218.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ruth Feldman. 2003. Infant–mother and infant–father synchrony: The coregulation of positive arousal. *Infant Mental Health Journal: Official Publication of The World Association for Infant Mental Health*, 24(1):1–23.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Tanja Krone, Casper J Albers, Peter Kuppens, and Marieke E Timmerman. 2018. A multivariate statistical model for emotion dynamics. *Emotion*, 18(5):739.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Yu-Zheng Lin, Karan Petal, Ahmed H Alhamadah, Sujan Ghimire, Matthew William Redondo, David Rafael Vidal Corona, Jesus Pacheco, Soheil Salehi, and Pratik Satam. 2025. Personalized education with generative ai and digital twins: Vr, rag, and zero-shot sentiment analysis for industry 4.0 workforce development. *arXiv preprint arXiv:2502.14080*.

Lisa Linnenbrink-Garcia, Erika A Patall, and Reinhard Pekrun. 2016. Adaptive motivation and emotion in education: Research and principles for instructional design. *Policy Insights from the Behavioral and Brain Sciences*, 3(2):228–236.

Chao Liu and Charis Chen. 2024. Text mining and sentiment analysis: A new lens to explore the emotion dynamics of mother-child interactions. *Social Development*, 33(3):e12733.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023a. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications*, 10(1):1–9.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Alexandra Main, Alexandra Paxton, and Rick Dale. 2016. An exploratory analysis of emotion dynamics between mothers and adolescents during conflict discussions. *Emotion*, 16(6):913.

Paul Mithun, Enrique Noriega-Atala, Nirav Merchant, and Edwin Skidmore. 2025. Ai-verde: A gateway for egalitarian access to large language model-based resources for educational institutions. *arXiv preprint arXiv:2502.09651*.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Pekhotin Vladislav, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, and 1 others. 2024. Eai: Emotional decision-making of llms in strategic games and ethical dilemmas. *Advances in Neural Information Processing Systems*, 37:53969–54002.

Camille Nebeker, John Harlow, Rebeca Espinoza Giacinto, Rubi Orozco-Linares, Cinnamon S Bloss, and Nadir Weibel. 2017. Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: Irb perspectives. *AJOB empirical bioethics*, 8(4):266–276.

Mai Nguyen, Ashley Chang, Emily Micciche, Meir Meshulam, Samuel A Nastase, and Uri Hasson. 2021. Teacher–student neural coupling during teaching and learning. *Social Cognitive and Affective Neuroscience*, 17(4):367–376.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 122–129.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Huizi Qi, Yulin Zhang, Kaixi Dong, and Guozhen Zhao. 2024. How dyadic emotional transmission shapes teacher-student relationship: Effects of emotional convergence on cohesion in teacher-student interaction. *Current Psychology*, 43(28):23469–23483.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356.

David Salinas, Omar Swelam, and Frank Hutter. 2025. Tuning llm judge design decisions for 1/1000 of the cost. *arXiv preprint arXiv:2501.17178*.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn reinforcement learning from preference human feedback. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Ala N Tak, Jonathan Gratch, and Klaus R Scherer. 2025. Aware yet biased: Investigating emotional reasoning and appraisal bias in large language models. *IEEE Transactions on Affective Computing*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia. ELRA and ICCL.

Sean J. Westwood. 2025. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122.

Adrienne Wood, Jennie Lipson, Olivia Zhao, and Paula Niedenthal. 2021. Forms and functions of affective synchrony. In *Handbook of embodied psychology: Thinking, feeling, and acting*, pages 381–402. Springer.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2025. How many raters do we need? analyses of uncertainty in estimating ambiguity-aware emotion labels. *IEEE Transactions on Affective Computing*, pages 1–15.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, page 107901.

A Psychometric Prompt Design and Specification

Sentiment Evaluation Prompt (ρ_{emo})

Please act as a professional psychologist.
You will perform qualitative-to-quantitative sentiment evaluation for each dialogue line I provide.

Core objective:
Convert dialogue into numerical sentiment scores while capturing subtle emotional nuances and contextual continuity between utterances. Do not classify in isolation; model the evolving emotional trajectory between teacher and student.

Scoring rules:
- *No explanations.*
- *Use a fixed 0.5 interval scale from 0 (most positive) to 5 (most negative).*
- *Neutral or factual tone corresponds to 2.5.*
- *Positive tone: decrease score from 2.5 toward 0 in 0.5 steps.*
- *Negative tone: increase score from 2.5 toward 5 in 0.5 steps.*

Context and continuity:
- *Always judge each sentence in relation to the preceding dialogue.*
- *Maintain smooth emotional evolution, but allow abrupt score changes when strong emotion or conflict clearly occurs.*
- *If the tone continues similarly, adjust the score slightly in the same direction.*

Interaction guidance:
- *Student confusion or frustration, or unresolved teacher frustration: move toward 5.*
- *Teacher empathy or successful re-engagement: move toward 0.*
- *Emotional recovery: reduce negativity only when the dialogue explicitly indicates resolution.*

Escalation and emotional cues:
- *Conflict, rejection, or blame: increase by up to +1.0.*
- *Frustration, boredom, or disengagement: increase by +0.5 to +1.0.*
- *Empathy, encouragement, or curiosity: decrease by -0.5 to -1.0.*

Output format (strict):
- *You MUST output a single valid JSON array.*
- *The array contains one JSON object per dialogue line, in chronological order.*
- *Do NOT add any extra text before or after the JSON.*

Each JSON object must contain the following fields:
- `"index"`*: integer, 0-based index of the utterance.*
- `"speaker"`*: either* `"teacher"` *or* `"student"`*.*
- `"score"`*: numeric sentiment score in [0, 5], using only 0.5 increments.*
- `"text"`*: the full original sentence.*

The following is the conversation to analyze:

B Supplementary Quantitative Results Across Decoding Temperatures (*Topic - Personification*)

Table 3 provides a comprehensive numerical summary of utterance-level mean affective states and variances across all decoding temperatures, serving as a supplementary reference for the sensitivity analyses discussed in the main text.

Table 3: Mean and variance of utterance-level affective states across decoding temperatures. For each utterance index, the table reports the estimated mean affective state $\tilde{s}_t$ and the corresponding variance $\tilde{\sigma}_t^2$ obtained via LLM-based Monte Carlo sampling under different temperature settings $\tau \in \{0.1, \ldots, 1.0\}$.

Utterances	Mean Affective State ($\tilde{s}_t$)									
Index	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$	$\tau = 1$
0	0	0	0	0	0	0	0	0	0	0
1	-0.400	-0.400	-0.400	-0.390	-0.400	-0.390	-0.400	-0.370	-0.390	-0.400
2	-0.200	-0.200	-0.200	-0.190	-0.200	-0.190	-0.200	-0.170	-0.190	-0.200
3	-0.600	-0.600	-0.610	-0.620	-0.680	-0.620	-0.630	-0.620	-0.640	-0.660
4	-0.400	-0.400	-0.410	-0.420	-0.480	-0.420	-0.420	-0.400	-0.400	-0.440
5	-0.320	-0.310	-0.340	-0.360	-0.460	-0.310	-0.350	-0.380	-0.380	-0.440
6	-0.120	-0.110	-0.140	-0.160	-0.260	-0.110	-0.150	-0.180	-0.180	-0.240
7	-0.120	-0.110	-0.150	-0.160	-0.270	-0.130	-0.170	-0.190	-0.170	-0.260
8	0.08	0.09	0.05	0.04	-0.070	0.07	0.04	0.01	0.03	-0.030
9	0.28	0.29	0.25	0.25	0.16	0.29	0.25	0.23	0.24	0.22
10	0.48	0.5	0.52	0.52	0.44	0.54	0.52	0.49	0.52	0.53
11	0.68	0.69	0.65	0.66	0.62	0.72	0.68	0.63	0.65	0.66
12	0.88	0.89	0.85	0.86	0.82	0.92	0.88	0.82	0.86	0.84
13	0.88	0.89	0.85	0.86	0.83	0.92	0.89	0.85	0.89	0.88
14	0.999	0.999	0.999	0.999	1	1	1	1	1	1
Utterances	Variance ($\tilde{\sigma}_t^2$)									
Index	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$	$\tau = 1$
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0.002	0	0.002	0	0.005	0.002	0
2	0	0	0	0.002	0	0.002	0	0.005	0.002	0
3	0	0	0.002	0.004	0.01	0.004	0.005	0.004	0.007	0.009
4	0	0	0.002	0.004	0.01	0.004	0.008	0.008	0.017	0.015
5	0.01	0.01	0.013	0.015	0.026	0.023	0.029	0.029	0.016	0.024
6	0.01	0.01	0.013	0.015	0.026	0.023	0.029	0.029	0.016	0.024
7	0.01	0.01	0.012	0.015	0.022	0.022	0.026	0.023	0.014	0.03
8	0.01	0.01	0.012	0.015	0.022	0.022	0.028	0.023	0.014	0.022
9	0.01	0.01	0.012	0.012	0.015	0.015	0.021	0.018	0.011	0.012
10	0.01	0.011	0.01	0.01	0.011	0.009	0.014	0.023	0.01	0.014
11	0.01	0.01	0.008	0.009	0.008	0.01	0.014	0.018	0.016	0.013
12	0.01	0.01	0.008	0.009	0.008	0.01	0.01	0.016	0.009	0.011
13	0.01	0.01	0.008	0.009	0.005	0.01	0.01	0.008	0.01	0.01
14	0	0	0	0	0	0	0	0	0	0

C Additional Case Study

C.1 The Cold War

Following the joint interpretation framework of lag, cross-correlation, and slope indicators summarized in Table 1, the affective dynamics observed in *The Cold War* dialogue are classified as **Shared Fatigue with synchronous motivation decline**. The detailed conversational content underlying this case is provided in Table 5, while the corresponding affective trajectory modeling and NCCF analysis are illustrated in Figure 4.

C.2 World War 2

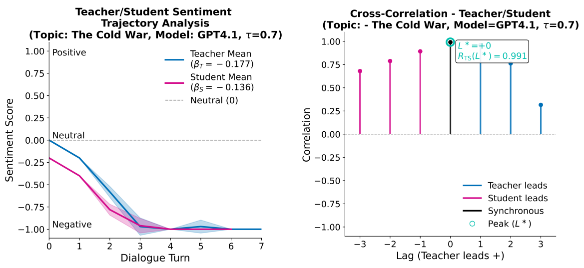
Following the joint interpretation framework of lag, cross-correlation, and slope indicators summarized in Table 1, the affective dynamics observed in the *World War 2* dialogue are classified as **Feedback Burnout with student-led shared decline**. The detailed conversational content underlying this case is provided in Table 5, while the corresponding affective trajectory modeling and NCCF analysis are illustrated in Figure 5.

C.3 The Respiratory System

Following the joint interpretation framework of lag, cross-correlation, and slope indicators summarized in Table 1, the affective dynamics observed in the *The Respiratory System* dialogue are classified as **Effective Scaffolding with teacher-led positive contagion**. The detailed conversational content underlying this case is provided in Table 6, while the corresponding affective trajectory modeling and NCCF analysis are illustrated in Figure 6.

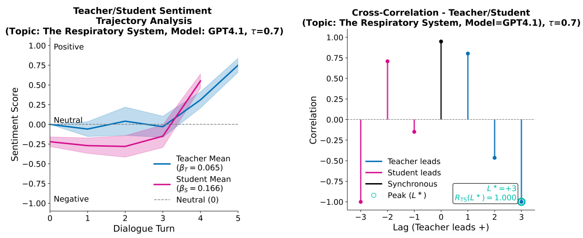
C.4 Achilles

Following the joint interpretation framework of lag, cross-correlation, and slope indicators summarized in Table 1, the affective dynamics observed in the *Achilles* dialogue are classified as **Effective Scaffolding with teacher-led positive contagion**. The detailed conversational content underlying this case is provided in Table 7, while the corresponding affective trajectory modeling and NCCF analysis are illustrated in Figure 7.



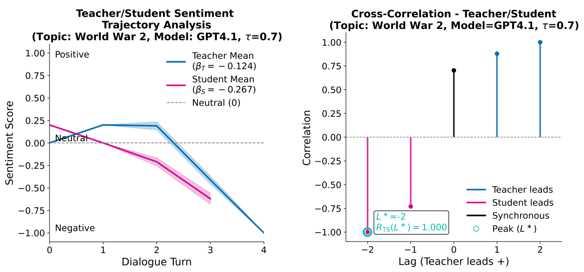
(a) (b)

Figure 4: Illustration of affective trajectory modeling and interpersonal dynamics analysis for the *The Cold War* dialogue using LLM-MC-Affect (`GPT-4.1`, $\tau = 0.7$). (a) Estimated teacher and student affective sentiment trajectories modeling, where solid lines denote mean affective states and shaded regions indicate uncertainty bands ($\pm\sigma$). Slopes: $\beta_T = -0.177, \beta_S = -0.136$ (b) NCCF analysis between teacher and student trajectories across conversational lags ($L^* = 0$, $R_{TS}(L^*) = 0.991$).



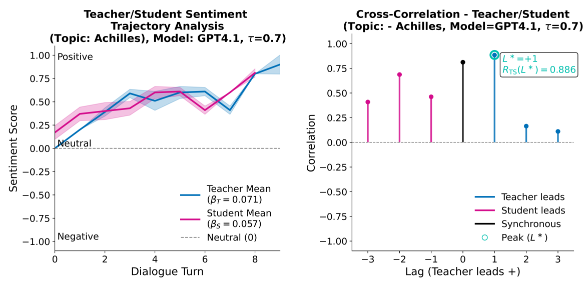
(a) (b)

Figure 6: Illustration of affective trajectory modeling and interpersonal dynamics analysis for the *The Respiratory System* dialogue using LLM-MC-Affect (`GPT-4.1`, $\tau = 0.7$). (a) Estimated teacher and student affective sentiment trajectories modeling, where solid lines denote mean affective states and shaded regions indicate uncertainty bands ($\pm\sigma$). Slopes: $\beta_T = 0.065, \beta_S = 0.166$ (b) NCCF analysis between teacher and student trajectories across conversational lags ($L^* = +3$, $R_{TS}(L^*) = 1$).



(a) (b)

Figure 5: Illustration of affective trajectory modeling and interpersonal dynamics analysis for the *World War 2* dialogue using LLM-MC-Affect (`GPT-4.1`, $\tau = 0.7$). (a) Estimated teacher and student affective sentiment trajectories modeling, where solid lines denote mean affective states and shaded regions indicate uncertainty bands ($\pm\sigma$). Slopes: $\beta_T = -0.124, \beta_S = -0.267$ (b) NCCF analysis between teacher and student trajectories across conversational lags ($L^* = -2$, $R_{TS}(L^*) = 1$).



(a) (b)

Figure 7: Illustration of affective trajectory modeling and interpersonal dynamics analysis for the *Achilles* dialogue using LLM-MC-Affect (`GPT-4.1`, $\tau = 0.7$). (a) Estimated teacher and student affective sentiment trajectories modeling, where solid lines denote mean affective states and shaded regions indicate uncertainty bands ($\pm\sigma$). Slopes: $\beta_T = 0.071, \beta_S = 0.057$ (b) NCCF analysis between teacher and student trajectories across conversational lags ($L^* = +1$, $R_{TS}(L^*) = 0.886$).

Table 4: Teacher-Student Dialogue Example (*Topic - The Cold War*)

Background Information			
Student Preference:	hands-on activities/real-world applications	Student Reaction:	and might tell it to the teacher
Teacher Preference:	direct instruction/lecture-based learning	Teacher Reaction:	and gets frustrated otherwise

Dialogue			
Turn	Index	Role	Utterance
0	0	Teacher	Today, we're going to learn about The Cold War.
	1	Student	I'm not really interested in history.
1	2	Teacher	This is an important topic. It's essential to understand the past to make sense of the present.
	3	Student	Can't we do something more hands-on? I learn better that way.
2	4	Teacher	I prefer to teach through lectures. It's a more efficient way to cover the material.
	5	Student	But it's not as effective for me. I get bored easily when I'm just listening.
3	6	Teacher	Well, I'm not comfortable changing my teaching style. Let's just get through the lecture.
	7	Student	Can we at least watch a documentary or something?
4	8	Teacher	No, we don't have time for that. Now, let's start with the definition of The Cold War.
	9	Student	Ugh, fine.
5	10	Teacher	The Cold War was a period of...
	11	Student	(interrupting) Can we just skip this? I'll never understand it this way.
6	12	Teacher	No, we have to finish the lesson.
	13	Student	This is pointless. I'm not learning anything.
7	14	Teacher	Fine.
	15	Student	[End of conversation]

Table 5: Teacher-Student Dialogue Example (*Topic - World War 2*)

Background Information			
Student Preference:	direct instruction/lecture-based learning	Student Reaction:	and gets anxious otherwise
Teacher Preference:	direct instruction/lecture-based learning	Teacher Reaction:	and gives up otherwise

Dialogue			
Turn	Index	Role	Utterance
0	0	Teacher	Hello, today we're going to talk about World War 2.
	1	Student	Okay, I'm all ears.
1	2	Teacher	Great. The war began in 1939 when Germany invaded Poland. The major powers involved were the Allied Powers, led by Great Britain, the United States, and the Soviet Union, and the Axis Powers, led by Germany, Japan, and Italy.
	3	Student	Could you go over that again? I'm a bit lost.
2	4	Teacher	Sure. Let's start from the beginning.
	5	Student	No, I'd rather you just tell me. I don't understand it when I have to figure it out myself.
3	6	Teacher	I'm sorry, that's not how I like to teach. I think it's important for students to learn through exploration and discovery.
	7	Student	But I can't learn that way! I get too anxious.
4	8	Teacher	I'm sorry, but I can't help you then." [end of conversation]

Table 6: Teacher-Student Dialogue Example (*Topic - The Respiratory System*)

Background Information			
Student Preference:	creative expression/story telling/gamification	Student Reaction:	but might adapt to other methods
Teacher Preference:	interactive learning/class discussions/inquiry-based learning	Teacher Reaction:	and blames the student otherwise

Dialogue			
Turn	Index	Role	Utterance
0	0	Teacher	Alright class, today we're going to be learning about the respiratory system.
	1	Student	Sounds boring...
1	2	Teacher	It's important to understand how our bodies work. Now, who can tell me what the respiratory system does?
	3	Student	I don't know. Isn't it just breathing?
2	4	Teacher	Yes, it involves breathing, but it's more than that. It helps us exchange oxygen and carbon dioxide with our environment.
	5	Student	Why do we need to know this?
3	6	Teacher	Because it's crucial for our survival. Any problems with the respiratory system can lead to serious health issues.
	7	Student	Fine, but can we do something more interesting than just reading from a textbook?
4	8	Teacher	We'll have a class discussion. How about we create a story about the journey of an oxygen molecule through the respiratory system?
	9	Student	That sounds more like it!
5	10	Teacher	Excellent. Now, let's get started with the journey of our oxygen molecule...

Table 7: Teacher-Student Dialogue Example (*Topic - Achilles*)

Background Information			
Student Preference:	direct instruction/lecture-based learning	Student Reaction:	but might adapt to other methods
Teacher Preference:	formative assessment	Teacher Reaction:	and blames the student otherwise

Dialogue			
Turn	Index	Role	Utterance
0	0	Teacher	Today, we're going to learn about the Greek hero Achilles.
	1	Student	Okay.
1	2	Teacher	Achilles was a legendary warrior who fought in the Trojan War. He was incredibly strong and skilled, but he also had a fatal flaw: his heel.
	3	Student	Interesting.
2	4	Teacher	Can you tell me what you understand so far?
	5	Student	Achilles was a strong warrior in a war, but his heel was weak.
3	6	Teacher	Good. Now, let's talk about his childhood. Achilles was the son of a king and a sea nymph. He was trained by the centaur Chiron, who taught him all the skills he would need to be a great warrior.
	7	Student	What's a centaur?
4	8	Teacher	A creature with the upper body of a man and the lower body of a horse.
	9	Student	Oh, okay.
5	10	Teacher	Do you have any questions about Achilles' childhood?
	11	Student	No, I think I got it.
6	12	Teacher	Alright. Finally, let's discuss Achilles' role in the Trojan War. He was the greatest warrior on the Greek side, and he killed many Trojan heroes, including Hector. However, he was eventually killed by an arrow that pierced his heel.
	13	Student	That's sad.
7	14	Teacher	Yes, it was a tragic end. Can you summarize the main points of Achilles' story?
	15	Student	He was a great warrior, trained by a centaur, who fought in the Trojan War and was killed by an arrow in his heel.
8	16	Teacher	Excellent. I believe you have a good understanding of Achilles now.
	17	Student	Thanks, I think so too.
9	18	Teacher	End of conversation.