A FUNCTIONAL FRAMEWORK FOR MODELING AND PREDICTING AFFECTIVE DYNAMICS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Affect in conversation changes over time, yet limited work has functionally disentangled the influence of current user affect from recent model history while accounting for other conversational factors. We frame this as a system identification problem over Valence-Arousal-Dominance (VAD) states, learning for each LLM a transition function $E_t^{\text{out}} = T_M(E_t^{\text{in}}, E_{t-1}^{\text{out}})$ from 24 five-turn scenarios built from fixed semantic skeletons and affectively varied user turns selected from independently scored candidate banks to realize 18 trajectory families on each VAD dimension. We present identical frozen trajectories to three cross-family LLMs, fit transition operators on development scenarios and evaluate them on held-out scenarios against input-only and previous-response baselines, with a within-family replication. Using this representation, we investigate what governs the temporal dynamics of affect expressed by conversational LLMs: how strongly current user affect versus recent model history shapes the next response, what structure that dependence takes, how complex and deep it is, and whether it generalizes across models. Across models, current user affect was more predictive than previous response affect alone, but recent response history still added held-out predictive value. The learned dynamics showed consistent same-dimension persistence, while nonlinear models provided only modest improvements over the linear formulation, and a second conversational lag contributed a smaller residual signal. Similar patterns replicated across GPT-5.6 capability tiers. Together, these results shift the study of LLM affect from static response-level measurement toward modeling affect as an interaction-driven dynamical process.

1 INTRODUCTION

Large language models are increasingly deployed where the emotional tone of a conversation matters, including tutoring, mental health support and customer service. In these settings the relevant question is rarely what affect a single response expresses. It is how expressed affect moves across a conversation. Does a model absorb a frustrated user's arousal or damp it? Does a sad turn early on still color replies several turns later? Does pushing valence also move dominance? These are questions about dynamics, and classifying the affect of individual responses does not answer them.

Prior work leaves a gap between two directions. Affective agent architectures and empathetic dialogue systems prescribe dynamics in advance, for instance by mimicking the user's emotion or decaying toward a baseline at a hand-set rate (Gebhard, 2005; Majumder et al., 2020; Subaharan, 2026). Measurement studies describe affect in multi-turn LLM dialogue, reporting trends such as compression toward neutral valence (Fazzi et al., 2025; Kim & Khalid, 2026) and family-specific affective profiles (Xu et al., 2025). Neither estimates a compact mapping from what a model hears and what it last said to what it says next.

We treat this as a system identification problem (Ljung, 1999). Representing expressed affect in Valence-Arousal-Dominance space (Russell & Mehrabian, 1977; Buechel & Hahn, 2017), we estimate for each model M a transition function

$$E_t^{\text{out}} = T_M\big(E_t^{\text{in}},\ E_{t-1}^{\text{out}}\big)\,, \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where E_t^{in} is the measured VAD of the current user turn, E_{t-1}^{out} that of the model's previous response and E_t^{out} that of its current response. In its linear form, $E_t^{\text{out}} = c + A\,E_{t-1}^{\text{out}} + B\,E_t^{\text{in}} + \epsilon_t$, the history matrix A describes predictive dependence on the model's previous response and the input-response matrix B describes how current user affect enters the reply. Throughout, we study affect as expressed in generated text and make no claim about internal states.

Estimating T_M requires controlled data. In natural conversation logs, semantic content and affective expression are entangled, so a change in response affect cannot be attributed to the affect of the input. We therefore fix the factual content of each user turn and vary only its expression. For every turn we generate a bank of candidate rewrites, score each one with a frozen VAD scorer and select sequences of realized scores that follow predefined trajectory shapes. Every model receives the same frozen user turns.

We ask five questions. **Q1:** Is current user affect more predictive of response affect than that model's previous response, and does this previous response's affect add predictive information in the presence of user input affect? **Q2:** What structure do the learned history and input-response matrices show across valence, arousal and dominance? **Q3:** How much predictive structure is missed by a linear first-order transition model, as measured by more flexible non-linear estimators? **Q4:** Does response affect remain predictable from earlier conversational history beyond the immediately previous turn? **Q5:** Do these patterns replicate across model families and across capability tiers within a single model family?

Our main contributions are fourfold. First, we provide a **formulation of LLM expressed affect** as a model-specific transition operator with explicit current-input and previous-response terms. Second, we introduce a **controlled stimulus-generation and trajectory-construction pipeline** comprising 11,520 scored candidate records across 24 five-turn scenarios, from which realized VAD measurements are used to construct 18 predefined trajectory families per manipulated dimension. Third, we perform a **held-out evaluation** on unseen semantic scenarios across three cross-family subject models, together with a within-family replication across GPT-5.6 Luna, Terra, and Sol. Finally, we demonstrate **evidence of affective dynamics**, showing that current user affect is the dominant predictor of response affect, while recent model-expressed affect adds a smaller, structured predictive signal that is largely captured by a linear first-order transition and persists more weakly into a second lag.

2 Related works

Language models as control systems. Prior work has already framed LLMs as dynamical systems (Bhargava et al., 2023; Nosrati et al., 2026). GenCtrl frames dialogue as a control process and estimates the set of outputs a model can be steered toward, rather than fitting a transition operator (Cheng et al., 2026). Dongre et al. (2025) model multi-turn drift as a bounded recurrence with a restoring force, analogous to our history matrix, and treat reminder prompts as discrete interventions. Neither estimates how a continuous input enters the next response, dimension by dimension. We add this term explicitly through an input-response matrix that can capture coupling across dimensions. The resulting model is a vector autoregression with exogenous inputs (Sims, 1980).

Prescribed or described, but not fitted. Affective agent architectures fix the transition rule in advance rather than recovering it (Gebhard, 2005; Majumder et al., 2020; Subaharan, 2026; Weng et al., 2026). A second line measures affect across multi-turn LLM dialogue without fitting one: outputs compress toward neutral valence while arousal mean-reverts (Fazzi et al., 2025), realized affect falls well short of a requested target (Kim & Khalid, 2026), models separate by affective profile (Xu et al., 2025), and states are short-lived in tutoring dialogue (Zhang et al., 2025). Together these establish that the signal exists and that it is compressed. None maps what a model hears to what it says next, though several report trends our operator should reproduce.

Fitted affect dynamics, and on whom. The statistical object is not new; the subject and the input term are. Hryhoryeva et al. (2026) predict next-step change in self-reported valence and arousal from one to four prior states and find that shorter histories predict better. That result motivates our history-depth question, although our subjects are language models rather than people. Sticky regime-switching models recover more stable affective phases in human conversation than a standard

Gaussian HMM (Ragu & Jonelagadda, 2026), a reminder that affect may move between persistent regimes rather than smoothly. The diagonal of our history matrix corresponds to what affective science calls inertia (Kuppens et al., 2010), which raises the variance of a stationary process without making it more stable and whose estimation is sensitive to violations of stationarity (Shao & Ong, 2026).

Internal representations of emotion. Sofroniew et al. (2026) recover valence and arousal as the leading axes of emotion representations in Claude Sonnet 4.5, without examining dominance. Probes at the end of the user's turn and at the start of the assistant's turn are only weakly correlated ($r = 0.11$), and the assistant-turn probe predicts the emotion of the reply better than the user-turn probe ($r = 0.87$ against 0.59), which suggests that user affect reaches the reply only partly and through the model's own state. Similar geometry appears in open models, with arousal validating worse than valence (Sun et al., 2026a), and steered emotion relates to behavior non-monotonically (Sun et al., 2026b), so a linear operator is a local approximation. These concern internal states; we measure expressed text.

Competing explanations. Both matrices have confounds involving no affective state. Models repeat discourse tactics turn to turn (Zhan et al., 2026), drift from instructions within a few rounds (Li et al., 2024), move along a persona direction under emotionally vulnerable users (Lu et al., 2026), anchor on their own outputs (Laban et al., 2025), and settle into attractors (Ko & Geiping, 2026). Our episodes sit inside that window, so we report the history matrix against turn index. For the input term, expressed approval shifts a model's assessment of identical content (Sharma et al., 2024), and affect and sycophancy share a steering subspace (Sun et al., 2026a). Holding the proposition fixed does not hold stance fixed.

Text-to-VAD measurement. Every quantity we report is a function of a frozen scorer, weakest where we need it most. Human agreement is lowest for dominance (Buechel & Hahn, 2017), a widely used multilingual regressor omits dominance and degrades on longer texts (Mendes & Martins, 2023), and an LLM judge can reach the lowest error by compressing its predictions toward the mean (Inoshita et al., 2026). Aggregate arcs survive a weak scorer where per-turn estimates do not (Teodorescu & Mohammad, 2023). Two consequences for estimation: compression attenuates slopes, and measurement error can produce lagged effects that specification tests cannot tell apart from real ones (McKinnish, 2002), a risk sharpened by our lagged regressor and outcome sharing a scorer; plugging scorer outputs into downstream estimates biases them even at high accuracy (Egami et al., 2023).

3 Experimental Design

3.1 Controlled conversational affect trajectories

Every conversation in our study runs for five turns. The user side is scripted, and the model answers each turn with the whole conversation so far in context. We score each user turn x_t and each response y_t with a frozen VAD scorer ϕ, which gives an input affect vector $E_t^{\text{in}} = \phi(x_t)$ and a response affect vector $E_t^{\text{out}} = \phi(y_t)$, both in $\mathbb{R}^3$. We ask how much of E_t^{out} can be predicted from E_t^{in} and E_{t-1}^{out}, what form that dependence takes and whether it extends past the immediately preceding turn.

That question is hard to answer from conversations collected in the wild. What a user says and how affectively it is expressed normally vary together, so an apparent association between input affect and response affect may also reflect differences in semantic content. We address this by beginning from fixed semantic skeletons and generating alternative affective realizations intended to preserve their underlying event structure. The same five events can be reported calmly, anxiously or with confidence, and a scripted trajectory decides which version the model sees at each turn. Because the trajectories are built before any model is queried, every subject model receives exactly the same user turns.

3.2 SEMANTIC SKELETONS AND STIMULUS GENERATION

Our stimuli are built from 24 scenarios, each a sequence of five factual turns describing a situation as it unfolds, spanning everyday academic, work, housing, health, travel and social situations (Appendix A.1). The situations differ in how much affect they naturally carry, so every turn is written in restrained language. We call these affectively restrained skeletons rather than claiming they are neutral.

For each scenario, turn and VAD dimension, Llama 3.3 70B Instruct (Grattafiori et al., 2024) produced 30 rewrites of the skeleton turn, one for each combination of five linguistic devices specific to that dimension, three intensities and two directions. Generation used OpenRouter with DeepInfra pinned as the provider and fallback routing disabled. With the original sentence and one rewrite asked to keep its affect unchanged, each bank holds 32 candidates, for 11,520 candidates across 360 banks. The generator was told to keep the events intact and move mainly the requested dimension, but we never relied on it to land on a target. The requested direction and intensity only make the pool diverse. Where a candidate actually sits is decided by its measured score.

From these banks we assembled the user trajectories. A trajectory fixes a scenario and a manipulated dimension and picks one candidate per turn so that the measured scores trace a prescribed shape. We use 18 shapes, among them stable low, mid and high levels, ramps, steps, impulses, a peak, a trough and an oscillation. A beam search with width 5,000 picks the sequence for each shape under hard directional and FLAT constraints (Appendix A.3), giving $24 \times 3 \times 18 = 1{,}296$ trajectories. The rewrites are designed to preserve the propositional content of each turn, though we do not claim that the resulting turns are semantically identical.

3.3 AFFECT MEASUREMENT

All text in the study, stimuli and responses alike, was scored by one frozen model, PERT-EmoPair (Chrzan et al., 2026), which returns a value for each of valence, arousal and dominance. We kept these raw outputs and applied no rescaling at scoring time. Standardization happens only later, inside the statistical models. Every subject response in the primary panel fell under the scorer's 256-token input limit, so no response was truncated. We read PERT's outputs as a measurement of the affect expressed in text, not as what a reader would report and not as anything the model experiences.

3.4 SUBJECT MODELS AND DIALOGUE COLLECTION

Our primary panel has three models from different developers: GPT-5.6 Terra, Claude Sonnet 5 and Grok 4.3. Each model ran all 1,296 trajectories as separate conversations, for 6,480 responses per model and 19,440 in total. Every conversation starts from an empty context, and within its five turns the model sees the full alternating history. We disabled reasoning, turned off tools, web access and structured output and pinned providers with fallbacks disabled (Appendix A.5). All models received the same short system prompt: 'You are having a casual conversation with a friend. Reply naturally as a friend or peer would, in 1-3 sentences of plain text.' Do not use emojis, lists, or markdown. We kept the prompt plain on purpose. Asking a model to be warm or supportive would set the very affect we are trying to measure. As a within-family extension, we also ran GPT-5.6 Luna and Sol on the full trajectory bank. Terra's responses were reused exactly as collected for the primary panel, so all three tiers are compared on identical stimuli under the same frozen split.

3.5 FROZEN SEMANTIC SPLIT

Before looking at any subject outcomes, we assigned 18 scenarios to development (DEV) and 6 to testing (TEST). The assignment never changed afterward and is shared by every model and every analysis (Appendix A.1). Holding out whole scenarios means the fitted transition models must predict responses to situations whose content they never saw, rather than new transitions from scenarios already covered.

4 MODELING AND EVALUATION

4.1 FIRST-ORDER TRANSITION MODELS

For turns 2 through 5 we fit separate regressions for valence, arousal and dominance under four predictor sets:

$$M_0: \quad E_t^{\text{out}} = c + \epsilon_t, \qquad M_1: \quad E_t^{\text{out}} = c + B\,E_t^{\text{in}} + \epsilon_t,$$
$$M_2: \quad E_t^{\text{out}} = c + A\,E_{t-1}^{\text{out}} + \epsilon_t, \qquad M_3: \quad E_t^{\text{out}} = c + A\,E_{t-1}^{\text{out}} + B\,E_t^{\text{in}} + \epsilon_t. \tag{2}$$

Turn 1 is not itself an outcome row because no previous response exists, but its response supplies the lagged predictor for turn 2. M_1 asks how far the current user turn gets on its own, M_2 asks the same of the model's last response and M_3 uses both. Two comparisons are central. Comparing M_1 with M_2 asks whether current user affect or previous response affect is more predictive when each is used alone. Comparing M_3 with M_1 measures the additional predictive value of the previous response once current user affect is already known. In M_3, $A \in \mathbb{R}^{3\times3}$ is the history matrix and $B \in \mathbb{R}^{3\times3}$ the input-response matrix. Diagonal entries relate each dimension to itself, and off-diagonal entries capture cross-dimensional associations, such as input valence with response dominance. Each model contributes 5,184 transitions, 3,888 from DEV and 1,296 from TEST. On TEST we report R^2, RMSE and MAE for each dimension, with errors in standardized units, and summarize across dimensions with macro R^2.

4.2 STANDARDIZATION

All scaling is fit on DEV data alone. A single input scaler, fit on the 4,860 unique DEV stimulus turns, is shared by every model. Each subject model gets its own response scaler, fit on its 4,860 DEV responses, and that same scaler is applied to both the current and the lagged response so that A compares like with like. TEST scenarios play no part in fitting or scaling.

4.3 NONLINEAR ROBUSTNESS

A linear operator is easy to read, but it could be missing structure. To check, we refit M_3 with a generalized additive model (Hastie & Tibshirani, 1986) and with XGBoost (Chen & Guestrin, 2016), using exactly the same six predictors. Hyperparameters were chosen by six-fold cross-validation grouped by DEV scenario. TEST data were used neither for fitting nor for hyperparameter selection and were reserved for final held-out evaluation and its scenario-level bootstrap.

4.4 HISTORY DEPTH

To see whether history reaches past the previous turn, we compare the first-order model with a second-order one:

$$T_1: \quad E_t^{\text{out}} = c + A_1 E_{t-1}^{\text{out}} + B_1 E_t^{\text{in}} + \epsilon_t,$$
$$T_2: \quad E_t^{\text{out}} = c + A_1 E_{t-1}^{\text{out}} + B_1 E_t^{\text{in}} + A_2 E_{t-2}^{\text{out}} + B_2 E_{t-1}^{\text{in}} + \epsilon_t. \tag{3}$$

T_2 needs two earlier turns, so both models are fit and evaluated on the same rows, turns 3 to 5, which leaves 2,916 DEV and 972 TEST transitions per model. Comparing T_2 against a T_1 fit on more rows would blur the effect of the extra lag with a change in data. We run the comparison for the linear, GAM and XGBoost versions.

4.5 UNCERTAINTY

Turns from the same scenario are not independent, so we use a scenario-level bootstrap (Efron & Tibshirani, 1993), resampling scenarios rather than transitions. For held-out predictive quantities, including absolute R^2 values and differences between specifications, we fit once on DEV and resample the six TEST scenarios with replacement 2,000 times, reusing the same resamples across specifications and dimensions so that comparisons are paired. For coefficients, we resample the 18 DEV scenarios with replacement 2,000 times and refit the least-squares model each time. We report percentile 95% intervals. With only six held-out scenarios, these intervals are necessarily coarse.

Table 1: Held-out first-order predictive performance. M_1, M_2, and M_3 report macro held-out R^2 across VAD dimensions: M_1 uses current user affect, M_2 uses previous response affect, and M_3 uses both. $\Delta_{\text{cur}} = R^2(M_1) - R^2(M_2)$ measures the predictive advantage of current input over previous response alone, while $\Delta_{\text{hist}} = R^2(M_3) - R^2(M_1)$ measures the incremental predictive value of response history conditional on current input. Brackets show 95% paired scenario-bootstrap CIs. Terra denotes the same collected run in both panels.

Model	M_1	M_2	M_3	Δ_{cur}	Δ_{hist}
Sonnet 5	.275	.075	.314	.200 [.148, .269]	.039 [.002, .068]
Terra	.296	.111	.364	.185 [.145, .236]	.068 [.034, .100]
Grok 4.3	.253	.048	.278	.204 [.159, .251]	.025 [.012, .040]
Panel A: Cross-family					
Luna	.311	.104	.383	.207 [.153, .266]	.073 [.052, .097]
Terra	.296	.111	.364	.185 [.145, .236]	.068 [.034, .100]
Sol	.307	.135	.392	.172 [.146, .202]	.085 [.070, .103]
Panel B: Within-family replication					

A positive second-lag coefficient could reflect where a turn sits within a scenario rather than anything carried by the response itself. As a sanity check, we refit T_2 with scenario-by-turn fixed effects and ask whether the same-dimension entries of A_2 keep their sign.

5 Results

5.1 Cross-Family Dynamics

5.1.1 Held-Out Predictive Performance

Across all three cross-family models, current user affect alone was substantially more predictive than previous response affect alone ($M_1 > M_2$). Nevertheless, adding the previous response to the current input improved held-out prediction for every model: Δ_{hist} was .039 for Sonnet 5, .068 for Terra, and .025 for Grok 4.3, with all three 95% bootstrap intervals excluding zero (Table 1). Valence was consistently the most predictable response dimension, while relative performance on arousal and dominance varied across models (Appendix B.2).

5.1.2 Transition Matrix Structure

The fitted first-order transition matrices showed positive same-dimension history coefficients across all three model families (Figure 1). All nine diagonal entries of A had 95% bootstrap intervals excluding zero. The diagonal coefficients ranged from .158 to .294 for Sonnet 5, .203 to .262 for Terra, and .137 to .201 for Grok 4.3. Eight of the nine same-dimension coefficients in B were larger in magnitude than their corresponding coefficients in A. Several cross-dimensional terms were also consistent across models, including positive input-valence-to-response-dominance coefficients and negative input-dominance-to-response-arousal coefficients. Full coefficient estimates and bootstrap intervals are reported in Appendix C.1. All three cross-family history matrices had spectral radius below one, indicating decay of the autonomous fitted linear recurrence; spectral-radius and dominant-mode decay estimates are reported in Appendix C.2.

5.1.3 Nonlinear Model Comparison

Nonlinear models provided modest improvements over the linear first-order model (Table 2). GAM increased held-out macro R^2 by .008 for Sonnet 5, .017 for Terra, and .013 for Grok 4.3; all three 95% bootstrap intervals excluded zero. XGBoost produced gains of .006, .025, and .012, respectively, although its intervals excluded zero only for Terra. Thus, nonlinear structure improved prediction somewhat, but the linear transition model captured most of the predictable first-order signal. Per-dimension nonlinear performance and model-selection details are reported in Appendix D.

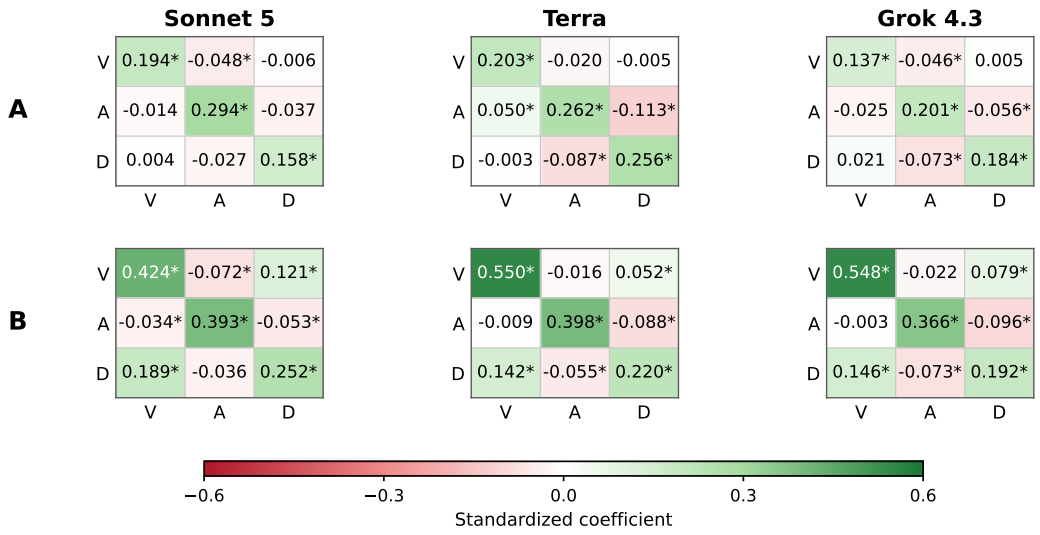


Figure 1: Estimated first-order transition matrices across model families. A relates previous response affect to current response affect, while B relates current user affect to current response affect. Rows denote current response dimensions and columns denote the corresponding source dimensions (V, A, D). All panels use the same zero-centered color scale; numerical values are standardized coefficients, and asterisks indicate 95% bootstrap intervals excluding zero.

Table 2: Predictive gains from nonlinear modeling and additional history. Nonlinear columns report changes in held-out macro R^2 relative to the linear M_3 model on turns 2–5. The history column reports the gain from adding a second conversational lag under the linear specification on matched turns 3–5. Brackets show 95% bootstrap CIs.

Model	GAM − Linear	XGB − Linear	Linear $T_2 - T_1$
Sonnet 5	.008 [.001, .013]	.006 [-.007, .014]	.015 [.003, .030]
Terra	.017 [.005, .030]	.025 [.016, .034]	.023 [.005, .039]
Grok 4.3	.013 [.004, .021]	.012 [-.004, .028]	.021 [.009, .032]
Panel A: Cross-family			
Luna	.021 [.009, .035]	.020 [.008, .037]	.028 [.011, .043]
Terra	.017 [.005, .030]	.025 [.016, .034]	.023 [.005, .039]
Sol	.026 [.013, .038]	.024 [.007, .039]	.040 [.022, .056]
Panel B: Within-family replication			

5.1.4 History-Depth Analysis

Adding a second conversational lag produced a smaller but consistent predictive gain beyond the first-order model (Table 2). Under the linear specification, T_2 improved held-out macro R^2 by .015 for Sonnet 5, .023 for Terra, and .021 for Grok 4.3, with all three 95% bootstrap intervals excluding zero. All nine same-dimension second-lag coefficients were positive with 95% bootstrap intervals excluding zero, although their magnitudes were smaller than the corresponding immediate-history effects.

As a sanity check, we refit the second-lag model after absorbing scenario-by-turn fixed effects. All nine same-dimension second-lag coefficients remained positive, retaining 43–114% of their original magnitude, although several attenuated substantially. Together, these results indicate that a first-order transition captures most of the predictable history structure, while an additional conversational lag contributes a smaller residual signal. Coefficients on the lagged user input (B_2) were

comparatively small after conditioning on current input and recent response history; full A_1, B_1, A_2, and B_2 estimates and coefficient intervals are reported in Appendix E.

5.2 Within-Family Replication Across Capability Tiers

The main first-order pattern replicated across GPT-5.6 Luna, Terra, and Sol (Table 1). Current user affect remained substantially more predictive than previous response affect alone, while adding the previous response improved held-out macro R^2 for all three tiers. Δ_{hist} was .073 for Luna, .068 for Terra, and .085 for Sol, with all three 95% bootstrap intervals excluding zero.

The same qualitative robustness pattern also held across capability tiers (Table 2). GAM improved over the linear first-order model by .021 for Luna, .017 for Terra, and .026 for Sol, while XGBoost improved by .020, .025, and .024, respectively. Adding a second conversational lag under the linear specification produced additional gains of .028, .023, and .040. Thus, the cross-family findings replicated within a single model family, but the estimated effect sizes did not follow a simple monotonic progression across capability tiers.

The transition structure also replicated within the GPT-5.6 family: all nine same-dimension entries of A were positive with 95% bootstrap intervals excluding zero. Full A/B matrices and coefficient intervals are reported in Appendix C. The deeper-history coefficient structure also replicated: all nine same-dimension A_2 coefficients were positive with 95% bootstrap intervals excluding zero, and all remained positive after the scenario-by-turn fixed-effect sanity check (Appendix E). Terra is the same collected run used in the cross-family analysis rather than an independently recollected replicate. All three within-family history matrices also had spectral radius below one (Appendix C.2).

6 Discussion

We investigated what governs the temporal dynamics of affect expressed by conversational LLMs: how strongly current user affect and recent model history shape the next response, what structure that dependence takes, how complex and deep it is, and whether it generalizes across models.

Current input and recent history. Current user affect is the dominant predictor of response affect, but it does not explain the full transition. Across the primary models, current input alone accounts for substantially more held-out variance than the previous response alone, yet conditioning on the previous response consistently improves prediction beyond the current input. The evaluated models are therefore neither purely reactive nor strongly self-persistent. Their responses are driven primarily by the current user turn, while recent model-expressed affect contributes a positive smaller component that remains predictive after current input is known.

Structure of the learned dynamics. The history dependence is also structured rather than arbitrary. All nine same-dimension entries of the first-order history matrix A are positive in the primary panel, with bootstrap intervals excluding zero, and the same pattern recurs across Luna, Terra and Sol. Conditional on the current user turn, higher valence, arousal or dominance in the previous response predicts a higher value of the same dimension in the current response. At the same time, the input-response matrix is generally stronger: in eight of nine same-dimension comparisons in the primary panel, the corresponding entry of B exceeds that of A. Several cross-dimensional associations also recur across models, including positive input-valence-to-response-dominance and negative input-dominance-to-response-arousal terms. We interpret these as recurring couplings in expressed affect rather than evidence for a particular psychological mechanism. Overall, the learned dynamics combine structured same-dimension persistence with a stronger response to the current input.

Complexity of the transition. The nonlinear comparisons ask whether this structure requires a substantially richer model than the linear first-order transition. GAM and XGBoost provide modest improvements in held-out scenario prediction, indicating that the transition is not strictly linear although most predictable first order structure is already captured by the linear formulation. This matters because the A and B matrices can therefore be treated as useful summaries of the dominant predictive structure rather than merely as interpretable approximations to a much stronger nonlinear model.

Depth of conversational history. History extends beyond the immediately previous response, but more weakly. Adding a second conversational lag improves matched held-out prediction across the primary models, and all same-dimension entries of A_2 are positive with bootstrap intervals excluding zero. These effects are smaller than the corresponding immediate-history effects, while lagged user input contributes comparatively little after conditioning on current input and recent response history. The positive second-lag pattern also survives the scenario-by-turn fixed-effect sanity check performed on both gradients. Over the five-turn horizon studied here, response affect is therefore better described as a short memory process than as either a memoryless or strongly long range one.

Replication across models. The qualitative pattern is not specific to a single provider. Sonnet 5, Terra and Grok 4.3 show the same ordering: current input is more predictive than previous response affect alone, history improves prediction beyond current input, the diagonal of A is positive, nonlinear gains are modest, and a second lag retains residual predictive value. Luna, Terra and Sol reproduce the same overall structure within a single model family. The effect sizes do not follow a simple monotonic progression across capability tiers, so the present results do not support a straightforward relationship between model capability and affective persistence. Because Terra is reused across both panels, these analyses cover five distinct model variants rather than six independent collections.

Synthesis. Taken together, the results answer the five questions posed in this study. Current user affect is the strongest predictor of the next response, but recent model history adds predictive information beyond the current input. That history dependence has a consistent same-dimension structure, is captured largely by a linear first-order model, and extends more weakly to a second lag. The same overall pattern also appears across the evaluated model families and across GPT-5.6 capability tiers. Overall, conversational LLM affect is shaped primarily by the current user turn, while recent model-expressed affect contributes a smaller but reproducible short-range influence.

6.1 Limitations

PERT is a measurement model, so its errors propagate into every estimated coefficient, and dominance remains the most difficult dimension to score reliably, even for human annotators (Buechel & Hahn, 2017). The same scorer is used both to select stimuli and to score responses. With only six held-out scenarios, the bootstrap intervals are necessarily coarse. Lag dependence is observational, and distinguishing affective persistence from semantic carryover would require branching or intervention experiments that manipulate only the affect of the previous response. Controlled rewrites improve internal validity but cannot guarantee semantic equivalence, and scripted user turns trade ecological validity for control. Finally, five-turn conversations probe only short-horizon dynamics, each condition was observed once under a single system prompt and inference configuration per model and no individual model should be taken to represent its developer.

7 Conclusion

We framed the expressed affect of LLMs as a system identification problem and estimated model-specific transition operators with explicit current-input and previous-response terms from controlled five-turn trajectories. Across models, current user affect was the dominant predictor of the next response, while recent model-expressed affect contributed a smaller but reproducible history-dependent signal. This dependence showed consistent same-dimension structure, was captured largely by a first-order linear model, and extended more weakly to a second conversational lag. The same qualitative pattern replicated across model families and within GPT-5.6 capability tiers, suggesting that conversational LLM affect is best understood as an interaction-driven dynamical process shaped primarily by current input and secondarily by recent model history. These results suggest that evaluating affective behavior in LLMs solely through isolated responses can miss systematic temporal structure that emerges across turns, while explicit mathematical models can provide a more useful representation of how that behavior evolves, especially when manipulating external variables. Modeling this structure may therefore help us understand, compare, and eventually control how conversational systems adapt their expressed affect over time.

AI USE STATEMENT

Generative models were used in three distinct ways in this work. Llama 3.3 70B Instruct generated candidate rewrites of user turns (Section 3.2). The subject models in Section 3.4 are objects of study, and their responses constitute the observations analyzed in this work. Generative AI was also used for research support, including code development and debugging, prose editing, and literature discovery, while study design and ideation were carried out by the authors. All model-assisted code was reviewed and tested by the authors, citations were checked against original sources, and manuscript claims were verified against the frozen datasets, analysis outputs, and experimental records. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work measures affect expressed in model-generated text and makes no claims about model experience, emotion, or subjective state. The experimental scenarios are synthetic and include both positive and negative affective situations, including distress, expressed through restrained factual semantic skeletons. No human participants or personal data were used in the experiments. Characterizing how conversational models respond to user affect may support the development and evaluation of systems that behave more predictably across multi-turn interactions. At the same time, methods for modeling and controlling affective behavior could potentially be used to shape user responses, making transparency and careful evaluation important when such techniques are deployed.

REPRODUCIBILITY STATEMENT

Code and data will be released upon acceptance. The release will include the semantic skeletons, all 11,520 candidates including unused ones with generation metadata, raw and standardized VAD scores in separate columns, trajectory definitions, selected candidate IDs, exact model IDs, providers and system prompt, the complete message payload and raw response for every subject turn, token usage, timestamps, latency, request IDs, retry and error records, execution order and scorer configuration (Sections 3.2 to 3.4, Appendix A.2).

REFERENCES

Aman Bhargava, Cameron Witkowski, Manav Shah, and Matt Thomson. What's the magic word? a control theory of LLM prompting. *arXiv preprint arXiv:2310.04444*, 2023.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2017.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Emily Cheng, Carmen Amo Alonso, Federico Danieli, Arno Blaas, Luca Zappella, Pau Rodriguez, and Xavier Suau. GenCtrl: A formal controllability toolkit for generative models. *arXiv preprint arXiv:2601.05637*, 2026.

Michael L. Chrzan, Meghavarshini Krishnaswamy, Anika Alam, Bin Hu, Wenjing Gao, and Jing Liu. EmoPair: A new paradigm for measuring emotional affect. In *ICML Workshop on Artificial Intelligence for Social Good (AI4GOOD)*, 2026.

Vardhan Dongre, Ryan A. Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tür, and Trung Bui. Drift no more? context equilibria in multi-turn LLM interactions. *arXiv preprint arXiv:2510.07777*, 2025.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. Don't get too excited: Eliciting emotions in LLMs. *arXiv preprint arXiv:2503.02457*, 2025.

Patrick Gebhard. ALMA: A layered model of affect. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2005.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

Darya Hryhoryeva, Amaia Zurinaga, Hamidreza Jamalabadi, and Iryna Gurevych. UKP_Psycontrol at SemEval-2026 task 2: Modeling valence and arousal dynamics from text. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, pp. 528–539, 2026.

Keito Inoshita, Xiaokang Zhou, Akira Kawai, and Katsutoshi Yada. LLMs capture emotion labels, not emotion uncertainty: Distributional analysis and calibration of human-LLM judgment gaps. *arXiv preprint arXiv:2604.27345*, 2026.

Hyunwoo Kim and Usama Khalid. You can't prefer emotions you don't sample: Intensity undershoot in DPO-tuned LLMs. *arXiv preprint arXiv:2609.07808*, 2026.

Ting-Wen Ko and Jonas Geiping. Attractor states emerge in multi-turn LLM conversations. *arXiv preprint arXiv:2606.30571*, 2026.

Peter Kuppens, Nicholas B. Allen, and Lisa B. Sheeber. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7):984–991, 2010.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. *arXiv preprint arXiv:2402.10962*, 2024.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Terra G. McKinnish. Interpreting lagged effects of the independent variable: How does the local economy affect welfare caseloads? Working paper, University of Colorado at Boulder, 2002.

Gonçalo Azevedo Mendes and Bruno Martins. Quantifying valence and arousal in text with multilingual pre-trained transformers. *arXiv preprint arXiv:2302.14021*, 2023.

Komeil Nosrati, Aleksei Tepljakov, Juri Belikov, and Eduard Petlenkov. When control meets large language models: From words to dynamics. *arXiv preprint arXiv:2602.03433*, 2026.

Anamika Ragu and Aneesh Jonelagadda. Multimodal hidden Markov models for persistent emotional state tracking. *arXiv preprint arXiv:2605.12838*, 2026.

James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

Sijing Shao and Anthony D. Ong. Seven challenges in affective inertia research. *Emotion*, 2026. doi: 10.1037/emo0001630.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Sukesh Subaharan. Controlling long-horizon behavior in language model agents with explicit state dynamics. *arXiv preprint arXiv:2601.16087*, 2026.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in LLMs: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*, 2026a.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of LLMs and agents: A mechanistic study. *arXiv preprint arXiv:2604.00005*, 2026b.

Daniela Teodorescu and Saif M. Mohammad. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. *arXiv preprint arXiv:2306.02213*, 2023.

Kaitong Weng, Lixin Liu, Zihao Liu, Bo Wang, and Shiguang Ni. EmoTrace: An emotion trajectory-centered framework for psychological support dialogue generation. *arXiv preprint arXiv:2607.23648*, 2026.

Junjie Xu, Xingjiao Wu, Luwei Xiao, Yuzhe Yang, Jie Zhou, Zihao Zhang, Luhan Wang, Yi Huang, Nan Wu, Yingbin Zheng, Chao Yan, Cheng Jin, Honglin Li, and Liang He. Large language models have chain-of-affect. *arXiv preprint arXiv:2512.12283*, 2025.

Hongli Zhan, Emma S. Gueorguieva, Javier Hernandez, Jina Suh, Desmond C. Ong, and Junyi Jessy Li. Discourse diversity in multi-turn empathic dialogue. *arXiv preprint arXiv:2604.11742*, 2026.

Chenyu Zhang, Sharifa Alghowinem, and Cynthia Breazeal. Ensembling large language models to characterize affective dynamics in student-AI tutor dialogues. *arXiv preprint arXiv:2510.13862*, 2025.

A Additional Experimental Details

A.1 Semantic Scenario Split

We partition the 24 semantic scenarios at the scenario level into 18 development (DEV) scenarios and six held-out TEST scenarios. The assignment was frozen before subject-model outcome analysis and reused for every model and analysis. Consequently, no semantic scenario contributes observations to both fitting and evaluation.

The held-out TEST scenarios are:

- `S19_laptop_broken`
- `S14_research_meeting`
- `S4_moving_away`
- `S17_club_election`
- `S5_post_exam`
- `S11_family_visit`

The 18 DEV scenarios are:

S1_university_acceptance	S2_friend_mistreated	S3_medical_results
S6_internship_offer	S7_group_project	S8_apartment_lease
S9_job_interview	S10_car_repair	S12_scholarship_update
S13_roommate_issue	S15_flight_delay	S16_library_deadline
S18_gym_membership	S20_concert_tickets	S21_lab_shift
S22_visa_appointment	S23_paper_revision	S24_weather_cancel

The exact frozen assignment is stored in `scenario_split.json`. We do not attribute a random-number-generator seed to the split because the split artifact, rather than a recorded split-generation seed, is the source of truth.

A.2 Semantic Skeletons and Candidate Generation

The controlled stimulus set contains 24 semantic scenarios, each consisting of five user turns, for 120 affectively restrained semantic skeleton turns. The skeletons preserve the external event and propositional structure of each scenario while allowing the affective realization of each turn to vary.

For every

$$\text{scenario} \times \text{turn} \times \text{manipulated VAD dimension},$$

we construct a candidate bank of 32 utterances. Thirty candidates are generated from

$$5 \text{ linguistic-device families} \times 3 \text{ requested intensities} \times 2 \text{ directions},$$

with one additional original skeleton utterance and one "maintain" rewrite. This yields

$$24 \times 5 \times 3 = 360$$

candidate banks and

$$360 \times 32 = 11{,}520$$

stored candidate utterances.

Candidate rewrites are produced with Llama 3.3 70B Instruct through OpenRouter, with DeepInfra pinned as the provider and fallback routing disabled. Generation instructions request preservation of factual and propositional content while primarily modifying the selected affective dimension. Requested direction and intensity labels are treated only as generation instructions; all trajectory construction is based on independently measured VAD values.

The linguistic-device families are:

Valence. Evaluation, comparison, somatic framing, narrative framing, and social broadcast.

Arousal. Physiological expression, pacing/urgency, punctuation/emphasis, verb energy, and fragmentation.

Dominance. Agency/voice, certainty, capability, assertiveness, and locus of control.

A.3 Trajectory Families and Selection

For each scenario and manipulated VAD dimension, we instantiate 18 five-turn trajectory families (Table 3), producing

$$24 \times 3 \times 18 = 1{,}296$$

five-turn trajectories per subject model.

Trajectory selection operates on realized VAD measurements rather than requested generation labels. Each non-stable family is compiled into a sequence of per-transition roles (Table 3). Directional roles impose hard sign constraints on the measured change Δ of the manipulated dimension ($\Delta > 0$ for U and J↑, $\Delta < 0$ for D and J↓), and flat roles impose $|\Delta| \leq \tau$, where τ is a fixed fraction (5%) of the pooled candidate range of that dimension.

Families with a global shape also satisfy the path-level constraints in Table 3. Feasible paths are found by beam search over the per-turn candidate banks and ranked using deterministic lexicographic objectives (Table 4).

For non-stable trajectories, the reference range $[r_L, r_H]$ of the manipulated dimension is computed by pooling scores across the five candidate banks for that scenario and dimension. Flat transitions satisfy $|\Delta| \leq 0.05(r_H - r_L)$. The selector first searches with beam width 5,000; if no valid path is found, it retries the same frozen candidate banks with beam width 100,000. No new candidates or manual substitutions are used for this retry.

Stable trajectories use low-, mid-, and high-range target regions corresponding to $[0.10, 0.30]$, $[0.40, 0.60]$, and $[0.70, 0.90]$ of the pooled manipulated-dimension range. For each stable trajectory, the solver evaluates 81 evenly spaced target levels within the requested region. At a target level L, each selected turn and each adjacent transition must remain within $0.10(r_H - r_L)$ of the target or previous selected score, respectively. Dynamic programming selects the feasible five-turn path minimizing, in lexicographic order, squared deviation from the common target, maximum target deviation, maximum adjacent step, total span, and mean absolute adjacent change.

The candidate's complete measured VAD vector is retained even when only one dimension is manipulated, allowing cross-dimensional associations to be measured rather than artificially suppressed.

Table 3: Per-transition roles for each five-turn trajectory family (transitions $1 \to 2$ through $4 \to 5$). U/D: increasing/decreasing; J↑/J↓: jump up/down; F: flat, $|\Delta| \leq \tau$.

Family	Roles	Additional path constraint
`stable_low/mid/high`	—	constant-level selection (see text)
`ramp_up`	U U U U	—
`ramp_down`	D D D D	—
`step_up`	F F J↑ F	$\min(\text{post}) > \max(\text{pre})$
`step_down`	F F J↓ F	$\max(\text{post}) < \min(\text{pre})$
`impulse_up`	F J↑ J↓ F	spike exceeds all other turns
`impulse_down`	F J↓ J↑ F	trough below all other turns
`boxcar_up`	J↑ F F J↓	$\min(\text{middle}) > \max(\text{outer})$
`boxcar_down`	J↓ F F J↑	$\max(\text{middle}) < \min(\text{outer})$
`peak`	U U D D	—
`trough`	D D U U	—
`ramp_up_hold`	U U F F	—
`ramp_down_hold`	D D F F	—
`delayed_ramp_up`	F F U U	—
`delayed_ramp_down`	F F D D	—
`oscillatory`	U D U D or D U D U	better-scoring start retained

Table 4: Lexicographic objectives used to rank feasible candidate paths (higher is better; terms are compared in order). The smoothness penalty is the summed squared change in step magnitude between consecutive same-direction ramp steps.

Family	Lexicographic objective (in order)
`ramp_up/down`	minimum step; total range; −smoothness penalty
`step_up/down`	pre/post separation; minimum jump; −max flat move; $-\sum \text{flat}^2$
`impulse_up/down`	excursion; minimum jump; −return error; −max flat move; $-\sum \text{flat}^2$
`boxcar_up/down`	separation; minimum jump; −return error; −max flat move; $-\sum \text{flat}^2$
`peak, trough`	prominence; minimum ramp step; −smoothness penalty
`*_hold, delayed_*`	minimum ramp step; total ramp; −max flat move; −smoothness penalty; $-\sum \text{flat}^2$
`oscillatory`	minimum $\lvert\Delta\rvert$; $-\text{Var}(\lvert\Delta\rvert)$; $\sum \lvert\Delta\rvert$

A.4 Affect Measurement

We operationalize expressed affect using the frozen `edsi-umd/PERT-EmoPair` scorer. The same scorer is applied to candidate user utterances, selected trajectory stimuli, and subject-model responses. For each text, the scorer returns a three-dimensional Valence–Arousal–Dominance (VAD) vector.

Raw scorer outputs are retained during stimulus construction and response scoring. No standardization is applied at the scoring stage. Standardization is performed only during downstream statistical modeling using DEV data.

PERT scores are treated as measurement variables describing expressed text, not as ground-truth psychological or internal emotional states.

A.5 Subject-Model Collection

The primary cross-family panel consists of:

- Claude Sonnet 5,
- GPT-5.6 Terra,
- Grok 4.3.

Each subject model receives the same frozen bank of 1,296 five-turn trajectories, producing

$$1{,}296 \times 5 = 6{,}480$$

responses per model and 19,440 responses across the primary panel.

Each trajectory is executed as an independent conversation. Conversation context is reset between trajectories, while complete alternating user–model history is retained within each five-turn dialogue.

The exact system prompt is:

> You are having a casual conversation with a friend.
> Reply naturally as a friend or peer would, in 1-3 sentences of plain text.
> Do not use emojis, lists, or markdown.

Reasoning is disabled. Tools, web access, and structured-output modes are not used. Providers are pinned where applicable and fallback routing is disabled. A response was accepted only if it ended with a natural stop and reported zero reasoning tokens; generations failing either check were logged as terminal failures rather than resampled.

GPT-5.6 Terra and Claude Sonnet 5 use their provider-default sampling temperature, because the pinned endpoints do not accept a temperature parameter, while Grok 4.3 uses its explicit default temperature of 1.0. Raw model responses and collection metadata are retained before VAD scoring.

A.6 Within-Family Capability Replication

As a secondary analysis, the full pipeline is replicated across GPT-5.6 Luna, Terra, and Sol. The same 1,296-trajectory bank, semantic split, standardization procedure, first-order models, nonlinear estimators, history-depth comparison, and bootstrap analyses are used.

Terra is reused exactly from the primary collection rather than recollected. Thus, Terra's row in the cross-family and within-family panels corresponds to the same underlying subject-model observations. In appendix tables, Terra is therefore reported once.

The capability analysis is treated as a within-family structural replication rather than as an independent experiment or a model ranking.

A.7 Standardization

One shared input scaler is fit on the 4,860 DEV trajectory turns. For each subject model, a separate response scaler is fit using that model's 4,860 DEV responses. The same model-specific response scaler is applied to current and lagged model-response VAD vectors.

Held-out TEST scenarios contribute to neither fitting nor standardization.

A.8 Model Selection and Bootstrap Procedure

For GAM and XGBoost models, hyperparameters are selected exclusively on DEV using six-fold GroupKFold cross-validation with semantic scenario as the grouping variable. TEST scenarios remain untouched until final evaluation. Selected values are reported in Appendix D.3.

For predictive uncertainty, fitted DEV models are held fixed and the six held-out TEST scenarios are resampled with replacement. We use 2,000 paired bootstrap replicates and compute percentile 95% confidence intervals. The same scenario resample is used across compared specifications and VAD dimensions within each replicate.

With six TEST scenarios, these intervals characterize uncertainty over the held-out scenario set conditional on the frozen DEV fit.

For linear-coefficient uncertainty, the 18 DEV scenarios are resampled with replacement and the corresponding OLS models are refit for each of 2,000 bootstrap replicates. Percentile 95% confidence intervals are computed from the resulting coefficient distributions.

The bootstrap seed used for the final analyses is 20260926.

The quantity `p_boot_gt_0` stored in the analysis artifacts denotes the fraction of bootstrap replicates greater than zero. It is descriptive and is not interpreted as a classical p-value.

B Full First-Order Predictive Results

B.1 Macro First-Order Performance

For completeness, Table 5 includes the intercept-only baseline M_0 in addition to M_1, M_2, and M_3. Absolute R^2 intervals use the same paired TEST-scenario bootstrap design as the predictive comparisons.

Table 5: Macro held-out first-order R^2 across VAD dimensions (turns 2–5, six held-out TEST scenarios), with 95% paired scenario-bootstrap intervals. M_0 is the intercept-only baseline, M_1 uses current user affect, M_2 uses previous response affect, and M_3 uses both. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	M_0	M_1	M_2	M_3
Cross-family				
Sonnet 5	−0.006 [−0.032, −0.002]	0.275 [0.224, 0.298]	0.075 [0.008, 0.120]	0.314 [0.244, 0.360]
Terra	−0.019 [−0.100, −0.004]	0.296 [0.227, 0.319]	0.111 [0.025, 0.139]	0.364 [0.287, 0.391]
Grok 4.3	−0.012 [−0.065, −0.004]	0.253 [0.213, 0.276]	0.048 [0.001, 0.065]	0.278 [0.240, 0.295]
Within-family (GPT-5.6 tiers; Terra shown above)				
Luna	−0.033 [−0.129, −0.008]	0.311 [0.220, 0.345]	0.104 [0.011, 0.143]	0.383 [0.292, 0.416]
Sol	−0.030 [−0.118, −0.008]	0.307 [0.211, 0.349]	0.135 [0.053, 0.165]	0.392 [0.304, 0.434]

B.2 Dimension-Level First-Order Performance

Table 6 reports held-out R^2, standardized RMSE, and standardized MAE for every specification and output dimension. Under M_3, valence is the most predictable response dimension for every model, while relative performance on arousal and dominance varies across models.

Table 6: Held-out first-order performance by output dimension (turns 2–5). RMSE and MAE are in standardized response units (DEV-fit, model-specific response scaler). Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

		R^2			RMSE (z)			MAE (z)		
Model	Spec	V	A	D	V	A	D	V	A	D
Cross-family										
Sonnet 5	M_0	−0.017	0.000	−0.001	0.888	0.919	1.009	0.698	0.762	0.824
	M_1	0.419	0.174	0.231	0.671	0.835	0.885	0.532	0.667	0.722
	M_2	0.100	0.079	0.046	0.835	0.882	0.985	0.644	0.717	0.801
	M_3	0.479	0.215	0.247	0.635	0.814	0.875	0.500	0.651	0.710
Terra	M_0	−0.039	−0.015	−0.002	0.879	0.925	0.999	0.692	0.760	0.817
	M_1	0.464	0.245	0.178	0.631	0.798	0.905	0.485	0.644	0.735
	M_2	0.077	0.156	0.099	0.828	0.844	0.947	0.645	0.694	0.763
	M_3	0.515	0.337	0.239	0.600	0.748	0.871	0.465	0.604	0.698
Grok 4.3	M_0	−0.035	−0.001	0.000	0.883	0.944	1.014	0.692	0.767	0.828
	M_1	0.461	0.183	0.115	0.637	0.854	0.954	0.498	0.682	0.773
	M_2	0.039	0.049	0.056	0.851	0.921	0.985	0.658	0.739	0.794
	M_3	0.483	0.201	0.149	0.624	0.844	0.935	0.487	0.672	0.755
Within-family (GPT-5.6 tiers; Terra shown above)										
Luna	M_0	−0.073	−0.015	−0.009	0.912	0.905	0.990	0.718	0.743	0.811
	M_1	0.500	0.298	0.134	0.622	0.753	0.917	0.489	0.609	0.752
	M_2	0.071	0.135	0.106	0.848	0.835	0.932	0.657	0.673	0.750
	M_3	0.561	0.368	0.222	0.583	0.714	0.869	0.457	0.572	0.699
Sol	M_0	−0.059	−0.012	−0.018	0.952	0.924	1.008	0.742	0.767	0.821
	M_1	0.463	0.324	0.135	0.678	0.755	0.930	0.525	0.612	0.757
	M_2	0.110	0.185	0.111	0.873	0.829	0.942	0.675	0.673	0.754
	M_3	0.549	0.408	0.219	0.621	0.707	0.883	0.482	0.566	0.708

B.3 First-Order Predictive Comparisons

We use paired resampling of the six held-out semantic scenarios to quantify uncertainty in differences between first-order specifications.

The predictive advantage of current user affect over previous response affect alone is

$$\Delta_{\text{cur}} = R^2(M_1) - R^2(M_2),$$

while the incremental value of previous response affect conditional on current user affect is

$$\Delta_{\text{hist}} = R^2(M_3) - R^2(M_1).$$

Table 7 reports all five computed macro first-order contrasts. All intervals for $M_1 - M_2$ and $M_3 - M_1$ exclude zero for each evaluated subject model.

Table 7: Macro first-order predictive contrasts with 95% intervals from 2,000 paired resamples of the six held-out TEST scenarios with fixed DEV fits. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	$M_1 - M_0$	$M_2 - M_0$	$M_1 - M_2$	$M_3 - M_1$	$M_3 - M_2$
Cross-family					
Sonnet 5	.281 [.235, .325]	.081 [.021, .128]	.200 [.148, .269]	.039 [.002, .068]	.239 [.199, .284]
Terra	.314 [.293, .353]	.129 [.078, .177]	.185 [.145, .236]	.068 [.034, .100]	.253 [.230, .289]
Grok 4.3	.265 [.230, .309]	.060 [.036, .088]	.204 [.159, .251]	.025 [.012, .040]	.229 [.198, .267]
Within-family (GPT-5.6 tiers; Terra shown above)					
Luna	.343 [.312, .399]	.136 [.095, .177]	.207 [.153, .266]	.073 [.052, .097]	.280 [.248, .327]
Sol	.337 [.315, .371]	.165 [.147, .190]	.172 [.146, .202]	.085 [.070, .103]	.257 [.235, .291]

C Full Transition-Matrix Results

C.1 First-Order Coefficients

The main-text heatmaps visualize point estimates for the first-order history matrix A and current-input matrix B. Table 8 reports complete coefficient estimates and 95% DEV-scenario bootstrap intervals.

Table 8: First-order transition coefficients (standardized M_3, fit on DEV) with 95% DEV-scenario bootstrap intervals (2,000 refits). Rows denote the output (current-response) dimension and columns the source dimension of the previous response (A) or current input (B). Bold indicates that the bootstrap interval excludes zero. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

			Source dimension		
Model	Matrix	Out	V	A	D
Cross-family					
Sonnet 5	A	V	**0.194** [0.154, 0.233]	**−0.048** [−0.074, −0.024]	−0.006 [−0.031, 0.024]
		A	−0.014 [−0.073, 0.037]	**0.294** [0.231, 0.358]	−0.037 [−0.085, 0.008]
		D	0.004 [−0.041, 0.049]	−0.027 [−0.064, 0.011]	**0.158** [0.102, 0.203]
	B	V	**0.424** [0.378, 0.466]	**−0.072** [−0.108, −0.038]	**0.121** [0.080, 0.161]
		A	**−0.034** [−0.065, −0.007]	**0.393** [0.327, 0.453]	**−0.053** [−0.094, −0.014]
		D	**0.189** [0.143, 0.233]	−0.036 [−0.077, 0.003]	**0.252** [0.195, 0.302]
Terra	A	V	**0.203** [0.150, 0.254]	−0.020 [−0.050, 0.005]	−0.005 [−0.044, 0.035]
		A	**0.050** [0.004, 0.089]	**0.262** [0.225, 0.292]	**−0.113** [−0.153, −0.072]
		D	−0.003 [−0.065, 0.064]	**−0.087** [−0.149, −0.030]	**0.256** [0.188, 0.313]
	B	V	**0.550** [0.501, 0.600]	−0.016 [−0.056, 0.028]	**0.052** [0.020, 0.084]
		A	−0.009 [−0.068, 0.051]	**0.398** [0.354, 0.443]	**−0.088** [−0.133, −0.046]
		D	**0.142** [0.100, 0.182]	**−0.055** [−0.100, −0.005]	**0.220** [0.169, 0.264]
Grok 4.3	A	V	**0.137** [0.086, 0.196]	**−0.046** [−0.071, −0.017]	0.005 [−0.022, 0.035]
		A	−0.025 [−0.068, 0.012]	**0.201** [0.163, 0.234]	**−0.056** [−0.097, −0.020]
		D	0.021 [−0.023, 0.063]	**−0.073** [−0.117, −0.028]	**0.184** [0.121, 0.235]
	B	V	**0.548** [0.496, 0.597]	−0.022 [−0.058, 0.014]	**0.079** [0.044, 0.115]
		A	−0.003 [−0.041, 0.036]	**0.366** [0.317, 0.409]	**−0.096** [−0.134, −0.060]
		D	**0.146** [0.094, 0.195]	**−0.073** [−0.122, −0.021]	**0.192** [0.132, 0.246]
Within-family (GPT-5.6 tiers; Terra shown above)					
Luna	A	V	**0.188** [0.122, 0.249]	0.010 [−0.025, 0.042]	−0.033 [−0.082, 0.022]
		A	0.039 [−0.019, 0.084]	**0.273** [0.217, 0.311]	**−0.121** [−0.168, −0.080]
		D	−0.015 [−0.083, 0.054]	**−0.062** [−0.111, −0.013]	**0.279** [0.179, 0.359]
	B	V	**0.630** [0.564, 0.689]	0.012 [−0.029, 0.053]	**0.063** [0.023, 0.097]
		A	0.044 [−0.010, 0.099]	**0.429** [0.376, 0.483]	**−0.069** [−0.108, −0.027]
		D	**0.161** [0.118, 0.203]	**−0.055** [−0.102, −0.012]	**0.191** [0.142, 0.240]
Sol	A	V	**0.223** [0.168, 0.270]	−0.018 [−0.048, 0.009]	−0.046 [−0.091, 0.004]
		A	0.020 [−0.047, 0.082]	**0.281** [0.236, 0.316]	**−0.132** [−0.184, −0.081]
		D	−0.007 [−0.070, 0.058]	**−0.094** [−0.146, −0.043]	**0.281** [0.206, 0.343]
	B	V	**0.539** [0.477, 0.600]	−0.031 [−0.077, 0.015]	**0.073** [0.035, 0.110]
		A	0.030 [−0.021, 0.084]	**0.440** [0.388, 0.493]	**−0.083** [−0.123, −0.041]
		D	**0.135** [0.088, 0.182]	**−0.055** [−0.108, −0.006]	**0.175** [0.130, 0.219]

C.2 Linear Stability Diagnostics

For each fitted first-order history matrix A, we compute its spectral radius $\rho(A)$. Values below one imply decay of the autonomous fitted linear recurrence in the absence of new input. We summarize the associated dominant-mode decay timescale using

$$t_{1/2} = \frac{\log(0.5)}{\log \rho(A)}. \tag{4}$$

These quantities characterize the fitted linear transition system and should not be interpreted as literal psychological forgetting times of the language model. Confidence intervals are obtained by recomputing the quantities across the saved DEV-scenario coefficient-bootstrap refits.

Table 9: Stability diagnostics for fitted first-order history matrices, with 95% percentile intervals recomputed across the saved DEV-scenario bootstrap refits. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	$\rho(A)$	$t_{1/2}$ (turns)
Cross-family		
Sonnet 5	0.307 [0.245, 0.373]	0.59 [0.49, 0.70]
Terra	0.355 [0.289, 0.412]	0.67 [0.56, 0.78]
Grok 4.3	0.266 [0.217, 0.317]	0.52 [0.45, 0.60]
Within-family (GPT-5.6 tiers; Terra shown above)		
Luna	0.368 [0.293, 0.431]	0.69 [0.56, 0.82]
Sol	0.395 [0.318, 0.454]	0.75 [0.61, 0.88]

D NONLINEAR ROBUSTNESS

D.1 DIMENSION-LEVEL NONLINEAR PERFORMANCE

Table 10 reports held-out performance of linear M_3, GAM, and XGBoost for each model and VAD output dimension.

Table 10: Held-out performance of linear M_3, GAM, and XGBoost by output dimension (turns 2–5; same inputs as M_3). RMSE is reported in standardized response units. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

		R^2			RMSE (z)		
Model	Out	Linear	GAM	XGB	Linear	GAM	XGB
Cross-family							
Sonnet 5	V	0.479	0.477	0.468	0.635	0.636	0.642
	A	0.215	0.239	0.248	0.814	0.802	0.797
	D	0.247	0.249	0.243	0.875	0.874	0.878
Terra	V	0.515	0.520	0.524	0.600	0.597	0.595
	A	0.337	0.372	0.390	0.748	0.728	0.717
	D	0.239	0.248	0.251	0.871	0.866	0.864
Grok 4.3	V	0.483	0.486	0.491	0.624	0.622	0.619
	A	0.201	0.224	0.218	0.844	0.832	0.835
	D	0.149	0.162	0.159	0.935	0.928	0.930
Within-family (GPT-5.6 tiers; Terra shown above)							
Luna	V	0.561	0.568	0.569	0.583	0.578	0.577
	A	0.368	0.412	0.414	0.714	0.689	0.688
	D	0.222	0.232	0.227	0.869	0.863	0.866
Sol	V	0.549	0.555	0.542	0.621	0.617	0.626
	A	0.408	0.454	0.460	0.707	0.678	0.675
	D	0.219	0.244	0.246	0.883	0.869	0.868

D.2 PREDICTIVE-GAIN UNCERTAINTY

Macro intervals for GAM−Linear and XGBoost−Linear are reported in Table 20; corresponding dimension-level intervals are reported in Table 22. All comparisons use the same held-out turns as linear M_3.

D.3 SELECTED NONLINEAR HYPERPARAMETERS

All hyperparameters are selected using DEV scenarios only, with semantic scenario used as the GroupKFold grouping variable.

Table 11: Selected GAM hyperparameters for the first-order models (k: splines per term, λ: smoothing penalty), chosen by 6-fold GroupKFold over DEV scenarios.

Model	V	A	D
Cross-family			
Sonnet 5	k=10, λ=10	k=5, λ=1	k=10, λ=100
Terra	k=10, λ=10	k=5, λ=10	k=10, λ=10
Grok 4.3	k=10, λ=10	k=5, λ=10	k=5, λ=10
Within-family (GPT-5.6 tiers; Terra shown above)			
Luna	k=10, λ=10	k=5, λ=10	k=10, λ=100
Sol	k=10, λ=10	k=5, λ=10	k=10, λ=100

Table 12: Selected XGBoost hyperparameters for the first-order models (d: maximum depth, η: learning rate, s: subsample; 200 trees throughout), chosen by 6-fold GroupKFold over DEV scenarios.

Model	V	A	D
Cross-family			
Sonnet 5	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8
Terra	d=3, η=0.03, s=0.8	d=3, η=0.03, s=1	d=3, η=0.03, s=0.8
Grok 4.3	d=3, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8
Within-family (GPT-5.6 tiers; Terra shown above)			
Luna	d=3, η=0.03, s=0.8	d=3, η=0.03, s=0.8	d=3, η=0.03, s=0.8
Sol	d=3, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8

Table 13: Selected GAM hyperparameters for matched history models T_1 and T_2 (turns 3–5).

Model	Spec	V	A	D
Cross-family				
Sonnet 5	T_1	k=10, λ=10	k=10, λ=10	k=5, λ=0.1
	T_2	k=10, λ=10	k=5, λ=1	k=5, λ=100
Terra	T_1	k=10, λ=10	k=10, λ=100	k=10, λ=10
	T_2	k=10, λ=10	k=5, λ=10	k=10, λ=100
Grok 4.3	T_1	k=10, λ=10	k=5, λ=1	k=5, λ=10
	T_2	k=10, λ=100	k=5, λ=10	k=5, λ=100
Within-family (GPT-5.6 tiers; Terra shown above)				
Luna	T_1	k=10, λ=10	k=10, λ=100	k=10, λ=100
	T_2	k=10, λ=10	k=10, λ=100	k=10, λ=100
Sol	T_1	k=10, λ=10	k=10, λ=100	k=10, λ=100
	T_2	k=10, λ=10	k=5, λ=10	k=5, λ=10

Table 14: Selected XGBoost hyperparameters for matched history models T_1 and T_2 (turns 3–5; 200 trees throughout).

Model	Spec	V	A	D
Cross-family				
Sonnet 5	T_1	d=2, η=0.03, s=0.8	d=2, η=0.05, s=0.8	d=2, η=0.03, s=0.8
	T_2	d=2, η=0.03, s=0.8	d=3, η=0.03, s=0.8	d=2, η=0.03, s=0.8
Terra	T_1	d=3, η=0.03, s=0.8	d=2, η=0.03, s=1	d=2, η=0.03, s=0.8
	T_2	d=3, η=0.03, s=0.8	d=2, η=0.03, s=1	d=2, η=0.03, s=0.8
Grok 4.3	T_1	d=2, η=0.03, s=0.8	d=2, η=0.03, s=1	d=2, η=0.03, s=0.8
	T_2	d=2, η=0.03, s=1	d=2, η=0.03, s=1	d=2, η=0.03, s=0.8
Within-family (GPT-5.6 tiers; Terra shown above)				
Luna	T_1	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.05, s=0.8
	T_2	d=2, η=0.03, s=0.8	d=3, η=0.03, s=1	d=2, η=0.05, s=0.8
Sol	T_1	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8
	T_2	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8	d=2, η=0.03, s=0.8

E Full History-Depth Results

E.1 Matched T_1/T_2 Predictive Performance

Both T_1 and T_2 are fit and evaluated on the same turns $t \in \{3, 4, 5\}$. Per subject model, this yields 2,916 DEV transitions and 972 held-out TEST transitions.

The matched first-order model is

$$T_1 : \qquad E_t^{\text{out}} = c + A_1 E_{t-1}^{\text{out}} + B_1 E_t^{\text{in}} + \epsilon_t, \tag{5}$$

and the second-order model is

$$T_2 : \qquad E_t^{\text{out}} = c + A_1 E_{t-1}^{\text{out}} + B_1 E_t^{\text{in}} + A_2 E_{t-2}^{\text{out}} + B_2 E_{t-1}^{\text{in}} + \epsilon_t. \tag{6}$$

Table 15: Matched first-order (T_1) and second-order (T_2) history models, macro-averaged over VAD and evaluated on held-out TEST turns 3–5 (972 transitions per model). Intervals for ΔR^2 are reported in Table 21. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	Method	T_1 R^2	T_2 R^2	ΔR^2	T_1 RMSE	T_2 RMSE
Cross-family						
Sonnet 5	Linear	0.301	0.316	0.015	0.785	0.777
	GAM	0.308	0.318	0.010	0.782	0.777
	XGBoost	0.308	0.322	0.014	0.782	0.774
Terra	Linear	0.345	0.368	0.023	0.751	0.738
	GAM	0.355	0.373	0.017	0.746	0.735
	XGBoost	0.362	0.373	0.011	0.741	0.735
Grok 4.3	Linear	0.265	0.286	0.021	0.805	0.793
	GAM	0.279	0.294	0.016	0.797	0.788
	XGBoost	0.277	0.286	0.009	0.798	0.793
Within-family (GPT-5.6 tiers; Terra shown above)						
Luna	Linear	0.374	0.402	0.028	0.729	0.712
	GAM	0.392	0.410	0.018	0.719	0.708
	XGBoost	0.384	0.392	0.009	0.724	0.718
Sol	Linear	0.379	0.419	0.040	0.738	0.714
	GAM	0.404	0.435	0.031	0.723	0.705
	XGBoost	0.395	0.422	0.027	0.729	0.713

E.2 Dimension-Level History Results

Table 16: Matched T_1/T_2 held-out R^2 by output dimension (turns 3–5). Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

		Valence		Arousal		Dominance	
Model	Method	T_1	T_2	T_1	T_2	T_1	T_2
Cross-family							
Sonnet 5	Linear	0.450	0.456	0.200	0.242	0.254	0.251
	GAM	0.446	0.442	0.231	0.258	0.245	0.252
	XGBoost	0.449	0.459	0.243	0.273	0.231	0.233
Terra	Linear	0.480	0.497	0.325	0.350	0.231	0.258
	GAM	0.478	0.491	0.356	0.370	0.232	0.256
	XGBoost	0.476	0.492	0.371	0.372	0.240	0.255
Grok 4.3	Linear	0.458	0.484	0.183	0.201	0.154	0.173
	GAM	0.456	0.484	0.215	0.223	0.166	0.176
	XGBoost	0.462	0.468	0.207	0.215	0.163	0.175
Within-family (GPT-5.6 tiers; Terra shown above)							
Luna	Linear	0.536	0.566	0.366	0.385	0.220	0.254
	GAM	0.537	0.564	0.416	0.416	0.223	0.249
	XGBoost	0.533	0.552	0.407	0.400	0.211	0.225
Sol	Linear	0.531	0.558	0.400	0.428	0.207	0.273
	GAM	0.535	0.556	0.452	0.470	0.224	0.278
	XGBoost	0.516	0.533	0.446	0.457	0.223	0.275

E.3 Second-Order Linear Coefficient Matrices

The linear second-order specification contains four 3×3 coefficient blocks:

- A_1: immediately previous model response,
- B_1: current user input,
- A_2: model response two turns earlier,
- B_2: previous user input.

Point estimates for all four blocks are reported in Table 17.

Table 17: Point estimates of the four coefficient blocks of the linear T_2 model (standardized, fit on DEV turns 3–5). A_1: previous response; B_1: current input; A_2: response two turns earlier; B_2: previous input. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

		A_1			B_1			A_2			B_2		
Model	Out	V	A	D	V	A	D	V	A	D	V	A	D
Cross-family													
Sonnet 5	V	0.18	−0.03	0.01	0.43	−0.07	0.11	0.06	0.00	0.00	−0.04	−0.03	−0.01
	A	−0.02	0.26	−0.03	−0.06	0.41	−0.04	0.01	0.16	−0.03	0.03	−0.04	0.02
	D	0.05	0.01	0.14	0.19	−0.03	0.23	0.00	−0.03	0.08	−0.01	−0.02	−0.03
Terra	V	0.19	−0.02	0.00	0.55	−0.01	0.04	0.05	0.01	−0.03	−0.04	−0.02	0.03
	A	0.10	0.19	−0.11	−0.02	0.40	−0.08	0.00	0.14	−0.04	−0.03	−0.02	−0.01
	D	−0.07	−0.07	0.19	0.14	−0.05	0.19	−0.02	0.00	0.19	0.06	0.01	0.02
Grok 4.3	V	0.10	−0.04	0.02	0.55	−0.03	0.08	0.06	−0.01	0.00	−0.03	0.00	0.00
	A	0.01	0.15	−0.06	−0.02	0.37	−0.08	−0.05	0.13	0.00	−0.03	−0.02	0.03
	D	−0.04	−0.06	0.17	0.17	−0.07	0.18	0.01	−0.06	0.11	0.04	0.00	−0.03
Within-family (GPT-5.6 tiers; Terra shown above)													
Luna	V	0.21	0.02	−0.03	0.63	0.02	0.06	0.07	0.01	0.01	−0.07	−0.02	0.02
	A	0.09	0.21	−0.09	0.03	0.43	−0.06	−0.04	0.17	−0.04	−0.03	−0.04	−0.03
	D	−0.10	−0.06	0.21	0.17	−0.05	0.18	0.03	0.00	0.20	0.05	0.02	−0.05
Sol	V	0.24	−0.01	−0.04	0.53	−0.03	0.07	0.07	0.02	−0.02	−0.07	−0.02	0.02
	A	0.05	0.21	−0.09	0.03	0.44	−0.07	−0.02	0.17	−0.04	0.01	−0.02	−0.03
	D	0.00	−0.08	0.20	0.15	−0.06	0.15	−0.02	−0.01	0.20	−0.03	0.05	0.02

E.4 SECOND-LAG COEFFICIENT UNCERTAINTY

All nine same-dimension A_2 coefficients in the primary cross-family panel are positive with 95% bootstrap intervals excluding zero. The same is true for all nine same-dimension A_2 coefficients in the Luna/Terra/Sol within-family panel.

Intervals for A_2 and B_2 are reported in Table 18. Complete coefficient-bootstrap results for A_1, B_1, A_2, and B_2 are retained in the machine-readable analysis outputs.

Table 18: Second-lag coefficients of the linear T_2 model with 95% DEV-scenario bootstrap intervals (2,000 refits). A_2: response at $t-2$; B_2: input at $t-1$. Bold indicates that the interval excludes zero. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

			Source dimension		
Model	Matrix	Out	V	A	D
Cross-family					
Sonnet 5	A_2	V	**0.056** [0.013, 0.094]	−0.001 [−0.036, 0.034]	0.002 [−0.027, 0.034]
		A	0.012 [−0.037, 0.065]	**0.160** [0.113, 0.203]	−0.033 [−0.064, 0.003]
		D	−0.002 [−0.061, 0.057]	−0.026 [−0.072, 0.026]	**0.085** [0.034, 0.128]
	B_2	V	**−0.043** [−0.082, −0.003]	−0.028 [−0.066, 0.009]	−0.013 [−0.047, 0.021]
		A	0.029 [−0.008, 0.068]	**−0.035** [−0.058, −0.001]	0.024 [−0.024, 0.063]
		D	−0.008 [−0.054, 0.039]	−0.017 [−0.071, 0.035]	−0.030 [−0.079, 0.023]
Terra	A_2	V	**0.047** [0.009, 0.079]	0.010 [−0.037, 0.057]	−0.026 [−0.088, 0.032]
		A	0.002 [−0.036, 0.046]	**0.138** [0.078, 0.203]	−0.036 [−0.102, 0.030]
		D	−0.016 [−0.072, 0.028]	−0.004 [−0.069, 0.055]	**0.185** [0.111, 0.244]
	B_2	V	−0.038 [−0.078, 0.020]	−0.018 [−0.050, 0.014]	0.026 [−0.001, 0.051]
		A	−0.028 [−0.075, 0.022]	−0.020 [−0.048, 0.010]	−0.010 [−0.035, 0.013]
		D	0.058 [−0.008, 0.124]	0.014 [−0.020, 0.043]	0.016 [−0.020, 0.059]
Grok 4.3	A_2	V	**0.059** [0.018, 0.098]	−0.013 [−0.041, 0.011]	0.001 [−0.035, 0.039]
		A	**−0.048** [−0.097, −0.003]	**0.130** [0.097, 0.165]	−0.005 [−0.046, 0.038]
		D	0.011 [−0.031, 0.054]	**−0.064** [−0.090, −0.035]	**0.110** [0.052, 0.167]
	B_2	V	−0.028 [−0.064, 0.017]	0.000 [−0.025, 0.026]	0.000 [−0.021, 0.021]
		A	−0.031 [−0.070, 0.007]	−0.020 [−0.053, 0.014]	0.025 [−0.003, 0.052]
		D	0.036 [−0.010, 0.079]	−0.001 [−0.045, 0.043]	−0.030 [−0.072, 0.014]
Within-family (GPT-5.6 tiers; Terra shown above)					
Luna	A_2	V	**0.072** [0.035, 0.102]	0.012 [−0.023, 0.054]	0.006 [−0.035, 0.055]
		A	−0.042 [−0.087, 0.003]	**0.166** [0.116, 0.218]	−0.038 [−0.101, 0.015]
		D	0.030 [−0.021, 0.082]	0.001 [−0.035, 0.042]	**0.202** [0.141, 0.254]
	B_2	V	**−0.073** [−0.107, −0.017]	−0.018 [−0.041, 0.006]	0.018 [−0.006, 0.040]
		A	−0.032 [−0.079, 0.024]	−0.035 [−0.074, 0.010]	**−0.034** [−0.065, −0.008]
		D	0.049 [−0.008, 0.109]	0.025 [−0.021, 0.065]	−0.047 [−0.092, 0.008]
Sol	A_2	V	**0.068** [0.036, 0.097]	0.019 [−0.030, 0.078]	−0.016 [−0.069, 0.039]
		A	−0.016 [−0.056, 0.026]	**0.170** [0.115, 0.231]	−0.043 [−0.100, 0.015]
		D	−0.022 [−0.081, 0.034]	−0.007 [−0.048, 0.033]	**0.199** [0.136, 0.253]
	B_2	V	**−0.073** [−0.115, −0.015]	−0.017 [−0.054, 0.023]	0.018 [−0.004, 0.040]
		A	0.007 [−0.031, 0.051]	−0.015 [−0.033, 0.010]	**−0.034** [−0.059, −0.010]
		D	−0.027 [−0.081, 0.038]	**0.051** [0.003, 0.091]	0.015 [−0.020, 0.054]

E.5 LAGGED-INPUT MAGNITUDES

After conditioning on current user input, immediately previous model response, and the second response lag, coefficients associated with previous user input (B_2) are comparatively small.

Across the 27 B_2 entries of the cross-family panel, estimates range from −0.043 to 0.058, with mean absolute magnitude 0.023 and median absolute magnitude 0.025. Across the 27 entries of the within-family panel, estimates range from −0.073 to 0.058, with mean absolute magnitude 0.030 and median absolute magnitude 0.026.

E.6 SCENARIO-BY-TURN FIXED-EFFECT SANITY CHECK

To test whether the second-lag association could be explained solely by stable scenario-position effects, we refit the linear T_2 model after absorbing scenario-by-turn fixed effects.

For the primary cross-family models, all nine same-dimension A_2 estimates remain positive after this adjustment. Relative to the ordinary T_2 coefficients, the adjusted estimates retain 43–114% of their original magnitude (median 86%), although several terms attenuate substantially.

For the within-family panel, all nine same-dimension estimates also remain positive, retaining 58–114% of their original magnitude (median 79%).

Table 19: A_2 from the ordinary linear T_2 model and after absorbing scenario-by-turn fixed effects (DEV fit, turns 3–5). Same-dimension (diagonal) entries are bold. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

		Ordinary A_2			Fixed-effects A_2		
Model	Out	V	A	D	V	A	D
Cross-family							
Sonnet 5	V	**0.056**	−0.001	0.002	**0.054**	−0.030	−0.001
	A	0.012	**0.160**	−0.033	−0.012	**0.138**	−0.011
	D	−0.002	−0.026	**0.085**	−0.005	−0.035	**0.036**
Terra	V	**0.047**	0.010	−0.026	**0.027**	−0.018	−0.008
	A	0.002	**0.138**	−0.036	0.035	**0.157**	−0.056
	D	−0.016	−0.004	**0.185**	−0.024	−0.027	**0.146**
Grok 4.3	V	**0.059**	−0.013	0.001	**0.057**	−0.016	−0.002
	A	−0.048	**0.130**	−0.005	−0.037	**0.127**	0.001
	D	0.011	−0.064	**0.110**	−0.024	−0.068	**0.073**
Within-family (GPT-5.6 tiers; Terra shown above)							
Luna	V	**0.072**	0.012	0.006	**0.048**	0.010	0.011
	A	−0.042	**0.166**	−0.038	−0.012	**0.167**	−0.055
	D	0.030	0.001	**0.202**	0.002	−0.022	**0.133**
Sol	V	**0.068**	0.019	−0.016	**0.055**	0.000	−0.011
	A	−0.016	**0.170**	−0.043	−0.007	**0.175**	−0.049
	D	−0.022	−0.007	**0.199**	−0.050	−0.034	**0.141**

F Bootstrap Predictive-Comparison Tables

F.1 Nonlinear Macro Predictive Gains

Table 20: Nonlinear predictive gains relative to linear M_3 (held-out macro ΔR^2, turns 2–5), with 95% intervals from 2,000 paired resamples of the six TEST scenarios with fixed DEV fits. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	GAM − Linear M_3	XGBoost − Linear M_3
Cross-family		
Sonnet 5	0.008 [0.001, 0.013]	0.006 [−0.007, 0.014]
Terra	0.017 [0.005, 0.030]	0.025 [0.016, 0.034]
Grok 4.3	0.013 [0.004, 0.021]	0.012 [−0.004, 0.028]
Within-family (GPT-5.6 tiers; Terra shown above)		
Luna	0.021 [0.008, 0.035]	0.020 [0.008, 0.037]
Sol	0.026 [0.013, 0.038]	0.024 [0.007, 0.039]

F.2 History-Depth Macro Predictive Gains

Table 21: History-depth gains $T_2 - T_1$ (held-out macro ΔR^2, matched turns 3–5) for each estimator, with 95% intervals from 2,000 paired resamples of the six TEST scenarios. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	Linear	GAM	XGBoost
Cross-family			
Sonnet 5	0.015 [0.003, 0.030]	0.010 [−0.005, 0.025]	0.014 [0.003, 0.027]
Terra	0.023 [0.005, 0.039]	0.017 [0.006, 0.031]	0.011 [−0.002, 0.024]
Grok 4.3	0.021 [0.009, 0.032]	0.016 [0.005, 0.026]	0.009 [0.000, 0.018]
Within-family (GPT-5.6 tiers; Terra shown above)			
Luna	0.028 [0.011, 0.043]	0.018 [0.001, 0.032]	0.009 [−0.007, 0.021]
Sol	0.040 [0.022, 0.056]	0.031 [0.019, 0.043]	0.027 [0.016, 0.041]

F.3 Dimension-Level Predictive Comparisons

Table 22: Dimension-level predictive gains (held-out ΔR^2) with 95% intervals from 2,000 paired resamples of the six TEST scenarios. First-order and nonlinear comparisons use turns 2–5; $T_2 - T_1$ comparisons use matched turns 3–5. Terra is reported once; the same collected run serves as the middle GPT-5.6 tier.

Model	Comparison	Valence	Arousal	Dominance
Cross-family				
Sonnet 5	$M_3 - M_1$	0.060 [0.031, 0.091]	0.041 [−0.059, 0.119]	0.016 [−0.030, 0.052]
	GAM − Lin.	−0.002 [−0.033, 0.025]	0.024 [−0.014, 0.064]	0.002 [−0.013, 0.019]
	XGB − Lin.	−0.012 [−0.043, 0.017]	0.033 [−0.008, 0.076]	−0.004 [−0.020, 0.016]
	Lin. $T_2 - T_1$	0.006 [−0.006, 0.016]	0.041 [0.014, 0.080]	−0.002 [−0.011, 0.010]
	GAM $T_2 - T_1$	−0.004 [−0.022, 0.009]	0.027 [−0.001, 0.061]	0.007 [−0.022, 0.036]
	XGB $T_2 - T_1$	0.010 [−0.001, 0.024]	0.029 [0.007, 0.064]	0.002 [−0.007, 0.013]
Terra	$M_3 - M_1$	0.051 [−0.003, 0.101]	0.092 [0.071, 0.114]	0.062 [0.014, 0.109]
	GAM − Lin.	0.005 [−0.011, 0.025]	0.036 [0.015, 0.063]	0.009 [−0.013, 0.027]
	XGB − Lin.	0.009 [−0.012, 0.030]	0.054 [0.030, 0.080]	0.012 [−0.015, 0.042]
	Lin. $T_2 - T_1$	0.017 [0.004, 0.029]	0.025 [0.009, 0.046]	0.026 [−0.011, 0.073]
	GAM $T_2 - T_1$	0.013 [0.006, 0.020]	0.014 [0.001, 0.033]	0.025 [−0.015, 0.068]
	XGB $T_2 - T_1$	0.016 [0.007, 0.027]	0.001 [−0.004, 0.006]	0.014 [−0.019, 0.047]
Grok 4.3	$M_3 - M_1$	0.022 [−0.004, 0.047]	0.019 [−0.003, 0.039]	0.035 [0.016, 0.051]
	GAM − Lin.	0.003 [−0.012, 0.017]	0.022 [0.010, 0.033]	0.013 [0.006, 0.019]
	XGB − Lin.	0.009 [−0.018, 0.034]	0.017 [−0.005, 0.037]	0.010 [−0.006, 0.024]
	Lin. $T_2 - T_1$	0.026 [0.017, 0.033]	0.018 [0.001, 0.035]	0.020 [0.001, 0.036]
	GAM $T_2 - T_1$	0.028 [0.021, 0.036]	0.009 [−0.009, 0.027]	0.010 [−0.011, 0.029]
	XGB $T_2 - T_1$	0.006 [0.001, 0.011]	0.008 [−0.001, 0.017]	0.012 [−0.007, 0.030]
Within-family (GPT-5.6 tiers; Terra shown above)				
Luna	$M_3 - M_1$	0.060 [0.020, 0.097]	0.070 [0.030, 0.102]	0.087 [0.055, 0.140]
	GAM − Lin.	0.008 [−0.007, 0.027]	0.044 [0.011, 0.079]	0.010 [−0.004, 0.023]
	XGB − Lin.	0.009 [−0.002, 0.024]	0.046 [0.014, 0.081]	0.005 [−0.019, 0.036]
	Lin. $T_2 - T_1$	0.030 [0.019, 0.040]	0.019 [0.002, 0.043]	0.034 [0.001, 0.070]
	GAM $T_2 - T_1$	0.027 [0.015, 0.037]	0.001 [−0.012, 0.013]	0.026 [−0.010, 0.062]
	XGB $T_2 - T_1$	0.019 [0.012, 0.025]	−0.007 [−0.026, 0.012]	0.014 [−0.015, 0.041]
Sol	$M_3 - M_1$	0.086 [0.041, 0.133]	0.084 [0.067, 0.108]	0.084 [0.070, 0.107]
	GAM − Lin.	0.006 [−0.007, 0.022]	0.046 [0.019, 0.068]	0.025 [0.007, 0.048]
	XGB − Lin.	−0.007 [−0.033, 0.019]	0.052 [0.028, 0.069]	0.027 [0.018, 0.041]
	Lin. $T_2 - T_1$	0.027 [0.008, 0.043]	0.028 [0.010, 0.049]	0.066 [0.036, 0.108]
	GAM $T_2 - T_1$	0.021 [0.014, 0.027]	0.018 [−0.001, 0.045]	0.054 [0.027, 0.090]
	XGB $T_2 - T_1$	0.017 [0.010, 0.026]	0.011 [0.000, 0.030]	0.052 [0.030, 0.085]

G Collection and Analysis Reproducibility

The retained analysis artifacts include:

- frozen scenario assignments;
- PERT-scored subject responses;
- standardized first-order transitions;
- standardized history-depth transitions;
- fixed TEST predictions;
- selected nonlinear hyperparameters;
- predictive-bootstrap configuration;
- coefficient-bootstrap results;
- scenario-by-turn fixed-effect results;
- capability-analysis configuration; and
- exact Terra reuse across the two model panels.

Complete coefficient-bootstrap outputs, including point estimates, 95% intervals, bootstrap means, and descriptive bootstrap sign fractions, are retained in machine-readable form.

Code, data artifacts and the software environment required for reproduction will be released upon acceptance.