One and Only at SemEval-2026 Task 2: Evaluating Zero-Shot Autonomous LLM Agents and Heuristic Proxies in Ecological Affect Forecasting

Nam Dinh Phuong
University of Information Technology, VNU-HCM
namdp.20@grad.uit.edu.vn

Abstract

This paper presents team *One and Only*'s system for SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time (Soni et al., 2026). We investigate whether zero-shot LLM reasoning can replace fine-tuning for ecological affect forecasting by combining deterministic statistical priors with frozen LLMs (Gemini 3 Pro, Claude Opus 4.6, GPT-5.2). For short-term state changes (Subtask 2A), an OLS mean-reversion anchor is paired with LLM-generated impulses; for long-term disposition changes (Subtask 2B), a Chain-of-Thought prompt drives direct numeric prediction. Our system underperforms fine-tuned approaches on both subtasks. However, post-submission ablation across three LLMs reveals a task-dependent pattern: CoT reasoning substantially improves disposition forecasting (r_V: $-0.185 \rightarrow +0.129$; MAE_V: $0.899 \rightarrow 0.422$), while uncalibrated LLM impulses degrade state-change prediction due to variance collapse ($\sigma_{\text{pred}} = 0.41$ vs. $\sigma_{\text{gold}} = 1.73$). We provide a detailed diagnostic analysis of these failure modes and release all prompts and outputs for reproducibility.

1 Introduction

SemEval-2026 Task 2 (Soni et al., 2026) introduces longitudinal ecological data—daily narrative essays with quantitative Valence and Arousal annotations (Russell, 1980)—for temporal affect forecasting. Existing EMA-based affect forecasting (Stone and Shiffman, 1994; Shiffman et al., 2008) typically fine-tunes Transformer models (Vaswani et al., 2017) on per-user histories, risking overfitting on sparse data. While chain-of-thought prompting (Wei et al., 2022) enables structured multi-step reasoning and sentiment analysis can extract continuous affective dimensions from text (Mohammad, 2021), applying LLMs to *numeric* continuous-valued regression remains challenging. Gruver et al. (2023) showed that frozen LLMs achieve competitive zero-shot time-series forecasting, though safety-tuned models exhibit predictive variance collapse when used as direct numeric regressors (Ouyang et al., 2022).

We hypothesized that a training-free baseline combining statistical priors with zero-shot LLM reasoning could be competitive. Our system pairs an OLS regression anchor with LLM semantic adjustment for Subtask 2A, and uses a CoT prompt for Subtask 2B. Both subtasks underperformed fine-tuned systems; this paper serves as a diagnostic evaluation with two key findings:

- **Variance Collapse:** LLM predictions compress to $\sigma_V = 0.41$ vs. gold $\sigma_V = 1.73$, destroying Pearson correlation in state-change prediction.
- **Task-Dependent Utility:** CoT reasoning helps disposition forecasting ($\Delta r_V = +0.314$) but harms state-change prediction without calibration.

2 Methodology

2.1 Subtask 2A: OLS Anchor + LLM Adjustment

Mean-Reversion Anchor. Affective states exhibit regression toward a physiological mean (Kuppens et al., 2010). We model the predicted shift $\hat{\Delta}X$ for $X \in \{V, A\}$ via OLS regression against the prior state:

$$\hat{\Delta}X_{\text{base}} = \alpha_X \cdot X_{t-1} + \beta_X \tag{1}$$

where $\alpha_X = \rho_{(X_{t-1},\Delta X)}(\sigma_{\Delta X}/\sigma_{X_{t-1}})$ and $\beta_X = \mu_{\Delta X} - \alpha_X \mu_{X_{t-1}}$, computed from training data. The resulting deterministic anchors are:

$$\hat{\Delta}V_{\text{base}} = -0.6626 \cdot V_{t-1} + 0.1573 \tag{2}$$

$$\hat{\Delta}A_{\text{base}} = -0.7624 \cdot A_{t-1} + 0.5749 \tag{3}$$

LLM Semantic Adjustment. A zero-shot Gemini 3 Pro agent reads the recent essays and generates a numeric emotional-shift estimate via CoT

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1205–1211

reasoning (prompt in Appendix B). The raw LLM output is added directly to the OLS anchor without calibration, and clamped to $[-3, 3]$:

$$\Delta V = \text{clamp}\big(\hat{\Delta}V_{\text{base}} + \text{LLM}(T_t),\ [-3, 3]\big) \quad (4)$$

This deliberately uncalibrated design stress-tests the raw LLM signal: if it adds value without calibration, the approach is promising; if it degrades the anchor, the failure mode is informative.

2.2 Subtask 2B: LLM Disposition Forecasting

Subtask 2B targets long-horizon disposition change. The same LLM receives per-user statistics (essay count, historical means, historical disposition change) and five recent essays, applying a 3-step CoT to output $\hat{D}^{(U)}_{\text{future}} \in [-3, 3]$ (prompt in Appendix C). We also report a **trend-continuation heuristic**: $\hat{D}^{(U)}_{\text{heuristic}} = \text{clamp}(D^{(U)}_{\text{historical}},\ [-3, 3])$.

3 Results

Evaluation uses Pearson r and MAE per SemEval-2026 protocols. The LLM component uses **Gemini 3 Pro** via GitHub Copilot Chat, zero-shot with default decoding parameters, CoT reasoning, and JSON output formatting. The OLS component is fully deterministic; the LLM is subject to closed-source inference nondeterminism. All prompts are in Appendix B and C. Tables 2–4 show official submissions; Table 1 and 3 are post-submission ablations on released labels.

3.1 Subtask 2A: Short-Term State Change

Table 1: Ablation Study on Subtask 2A (Test Set, N=46)

Model Config.	Valence		Arousal	
	r	MAE	r	MAE
Linear (Prev Baseline)	0.615	1.168	0.670	0.638
OLS Mean Anchor Only	0.357	**0.907**	0.410	0.752
Hybrid (LLM CoT + Anchor)	−0.009	1.377	−0.114	1.054
LLM Direct (GPT-5.2)	−0.064	1.326	−0.186	0.770

3.2 Subtask 2B: Long-Term Disposition

Table 3 shows an ablation comparing the trend-continuation heuristic against the LLM CoT approach evaluated on the released test labels. Table 4 places our submitted result (heuristic) in the official leaderboard context.

Team / Model	V r	V MAE	A r	A MAE	Avg r
HITSZ-CyberS	**0.698**	–	0.568	–	**0.633**
YNU	0.692	–	**0.647**	–	0.669
Momentum	0.553	–	0.589	–	0.571
linear(prev)	0.615	1.168	0.670	0.638	0.642
One and Only*	*-0.194*	*1.398*	*-0.423*	*0.818*	*-0.308*

Table 2: Official Subtask 2A leaderboard metrics. (*MAE independently evaluated on released labels)

Table 3: Subtask 2B post-submission ablation (released labels, N=46). All LLM rows use v1 prompt except where noted. Post-hoc scaling matches σ_{pred} to σ_{train}.

Method	Valence		Arousal	
	r	MAE	r	MAE
Trend-Continuation Heuristic	−0.185	0.899	+0.016	0.483
LLM CoT v1 (Gemini 3 Pro)	**+0.129**	**0.422**	**+0.094**	**0.306**
LLM CoT v1 (Claude Opus 4.6)	−0.190	0.522	+0.065	0.333
LLM CoT v1 + post-hoc scaling	+0.129	0.606	+0.094	0.419
LLM CoT v2 (enriched prompt)	+0.021	0.687	−0.040	0.420

4 Error Analysis

4.1 Subtask 2A: Failure Modes

The hybrid pipeline yielded negative correlations ($r_V = -0.009$), worse than the pure OLS anchor ($r_V = 0.357$). A separate GPT-5.2 zero-shot run ($r_V = -0.064$) confirms this is not model-specific. Three failure modes explain this:

(1) Variance Collapse: Ground-truth Valence changes have $\sigma = 1.73$; LLM predictions compress to $\sigma = 0.41$ (Figure 1). Safety-tuned LLMs anchor outputs to conservative near-zero values (Ouyang et al., 2022), destroying correlation.

(2) CoT Neutralization Bias: Step-by-step reasoning forces the model to weigh both positive and negative triggers, averaging them into neutralized near-zero outputs that fail to capture impulsive, non-linear affective shifts.

(3) Uncalibrated Noise Injection: Without a calibration layer, raw LLM scalars act as adversarial noise on the stable OLS anchor. Post-hoc linear variance scaling cannot help: Pearson r is scale-invariant ($r(aX + b, Y) = r(X, Y)$), and scaling amplifies misdirected predictions (MAE: $1.377 \rightarrow 1.851$). The binding constraint is direction accuracy, not variance. Representative failure cases are shown in Appendix A.

4.2 Subtask 2B: Disposition Analysis

The LLM CoT achieves $r_V = +0.129$, $\text{MAE}_V = 0.422$ vs. the heuristic's $r_V = -0.185$, $\text{MAE}_V = 0.899$—demonstrating genuine predictive signal. The heuristic fails because copying historical dis-

Team / Model	V r	V MAE	A r	A MAE	Avg r
HITSZ-CyberS	**0.580**	–	0.200	–	**0.390**
linear(prev)	0.434	0.406	0.584	0.286	0.509
UAlberta	0.405	–	**0.602**	–	0.503
One and Only*	*-0.185*	*0.899*	*0.016*	*0.483*	*-0.084*

Table 4: Official Subtask 2B leaderboard (submitted heuristic result). (*MAE independently evaluated on released labels)

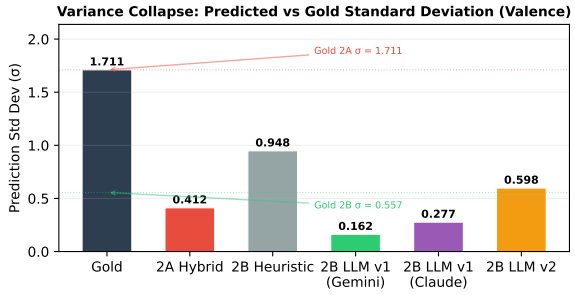

Figure 1: Predicted σ across systems (Valence). The 2A hybrid σ is 4$\times$ smaller than gold, confirming variance collapse. The 2B v2 prompt achieves near-target σ but degrades r by amplifying directional errors.

position change cannot capture trajectory reversals that the LLM identifies by reasoning over recent essays relative to the historical mean. Full scatter plots are in Appendix A.

Post-hoc Calibration. Variance scaling (matching σ_{pred} to σ_{train}) preserves $r_V = +0.129$ by construction (r is scale-invariant) but increases MAE from 0.422 to 0.606, confirming that the binding constraint is directional accuracy.

Prompt Sensitivity. An enriched v2 prompt supplying pre-computed trend and volatility features improved distributional alignment (σ_V: 0.162 $\rightarrow$ 0.598, approaching the target 0.620) but degraded r_V from $+0.129$ to $+0.021$. When direction accuracy is near chance ($\approx 55\%$), amplifying magnitude worsens Pearson r.

Cross-Model Comparison. Running the identical v1 prompt on Claude Opus 4.6 yields $r_V = -0.190$ (vs. Gemini's $+0.129$) despite comparable MAE (0.522 vs. 0.422), demonstrating that zero-shot disposition forecasting is model-sensitive: the same prompt produces qualitatively different correlation structures, suggesting observed correlations reflect model-specific inductive biases rather than robust prompt-driven reasoning.

Takeaway. Zero-shot LLM reasoning provides genuine signal when the task is semantically tractable. Disposition change (slow-moving) benefits from CoT ($\Delta r_V = +0.314$); momentary state change (volatile) requires parametric calibration. LLMs should serve as upstream semantic encoders feeding calibrated regression layers.

5 Conclusion

We evaluated zero-shot LLM agents for longitudinal affect forecasting. LLM CoT reasoning captures disposition-level trends ($r_V = +0.129$ vs. -0.185 for the heuristic) but disrupts state-change prediction via variance collapse. The utility of zero-shot reasoning is gated by signal timescale: slow-moving targets respond to narrative reasoning; volatile targets require parametric calibration. Future work should route LLM semantic signals through calibrated regression layers and evaluate on held-out development partitions (Gruver et al., 2023).

Limitations

The system was evaluated with three LLMs (Gemini 3 Pro, Claude Opus 4.6, GPT-5.2) on a 46-instance test set; Pearson r estimates are noisy without confidence intervals. Cross-model variability is substantial: 2B r_V ranges from -0.190 (Claude) to $+0.129$ (Gemini) under the same prompt, and both Gemini and GPT-5.2 yield negative 2A correlations (-0.009, -0.064), confirming model sensitivity rather than stable prompt-driven reasoning. Post-submission ablation used released labels and was not part of official scoring. Exact replication is constrained by closed-source models, though all prompts and coefficients are disclosed in the appendices.

Acknowledgments

We thank the SemEval-2026 Task 2 organizers for designing this challenging longitudinal affect benchmark.

References

Nate Gruver, Marc Finzi, Micah Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

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

Saif M Mohammad. 2021. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, pages 201–237.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annual review of clinical psychology*, 4:1–32.

Nikita Soni, H. Andrew Schwartz, Ryan L. Boyd, Phi Long Bui, Syeda Mahwish, August Håkan Nilsson, Adithya V Ganesan, Lyle Ungar, Niranjan Balasubramanian, and Saif M. Mohammad. 2026. SemEval-2026 task 2: Predicting variation in emotional valence and arousal over time from ecological essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment. *Annals of behavioral medicine*, 16(3):199–202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

A Supplementary Figures and Tables

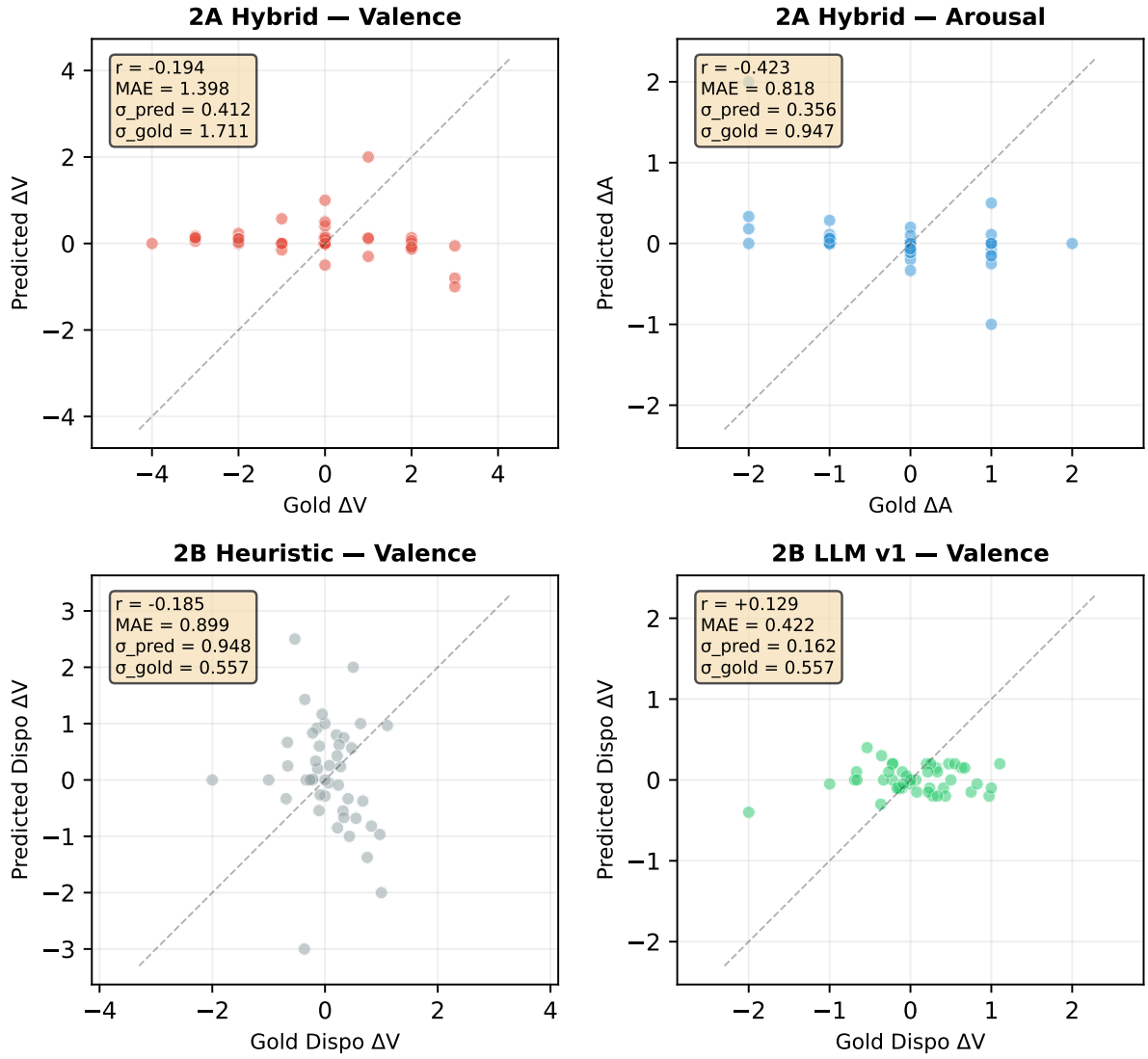

Figure 2: Predicted vs. gold scatter plots (Valence dimension). **Top row:** Subtask 2A hybrid predictions cluster near zero ($\sigma_{\text{pred}} = 0.41$ vs. $\sigma_{\text{gold}} = 1.71$), illustrating severe variance collapse. **Bottom row:** Subtask 2B heuristic (left) shows high spread but negative correlation ($r = -0.185$); LLM v1 (right) achieves positive correlation ($r = +0.129$) despite conservative magnitude.

Table 5: Representative Subtask 2A failure cases (Valence). Context descriptions are derived from LLM reasoning traces.

User	Context summary	Pred	Gold	|Err|
8	Bereavement; LLM flags large negative shift	-1.17	-4.0	2.83
6	Anxious state; LLM adjustment inverts anchor	$+0.69$	-1.0	1.69
50	Surface-positive text; large drop in gold	-1.04	-3.0	1.96
29	Negative keywords; strong positive rebound	$+1.36$	$+3.0$	1.64

B System Prompt Template (Subtask 2A)

The following prompt was provided to the Gemini 3 Pro agent for all Subtask 2A instances. Placeholders in `{braces}` are substituted at inference time with per-instance data.

```
ROLE: You are an expert in quantitative psychology
specialising in longitudinal affect forecasting.
Assess short-term emotional change using the
Russell Circumplex Model:
  Valence: negative <-> positive
  Arousal: calm <-> excited
INPUTS:
  Recent essay history: {ESSAY_HISTORY}
  Current essay:        {CURRENT_ESSAY}
  Prior state: Valence = {PREV_VALENCE}, Arousal = {PREV_AROUSAL}
INPUT TYPE DETECTION:
  Narrative : analyse causal event logic.
  Keywords  : analyse adjective density.
  Noisy/Spam: predict mild arousal decrease (low-motivation signal).
3-STEP CHAIN-OF-THOUGHT REASONING:
Step 1 (Anchor): Identify current V_t and A_t as baseline.
  All predicted changes are relative to this.
Step 2 (Intensity Mapping):
  Large  (+-2.0 to +-3.0): major life events
  Medium (+-0.5 to +-1.5): work, sleep, daily interactions
  Small  (+-0.1 to +-0.4): natural mood fluctuations
Step 3 (Delta Computation): Compute predicted change.
  Regression to the Mean: if V_t near +3, expect negative delta;
  if near -3, expect positive delta.
CONSTRAINTS: Output ONLY valid JSON. Values in [-3, 3].
{"reasoning": "<trace>",
 "pred_state_change_valence": <float>,
 "pred_state_change_arousal": <float>}
```

C System Prompt Templates (Subtask 2B)

We ran two Subtask 2B inference experiments. **Prompt v1** (primary post-submission experiment, Table 3 row 2) supplies raw per-user statistics and 5 recent essays. **Prompt v2** (prompt-sensitivity ablation, row 3) adds pre-computed `recent_trend` and `recent_volatility` signals per dimension and instructs bolder magnitude outputs.

Prompt v1 — Primary Experiment. Input: `subtask2b_llm_inputs.json` (raw statistics + last 5 essays per user).

```
ROLE: Expert in longitudinal affect forecasting (Russell Circumplex).
  Valence: negative <-> positive  (scale: -3 to +3)
```

```
  Arousal: calm <-> excited       (scale: -3 to +3)
TASK: Predict the SHIFT in the user's dispositional BASELINE
  (not momentary state). historical_disposition_change is anchor.
INPUT: n_historical_essays, mean_valence, mean_arousal,
  historical_disposition_change_valence/arousal, recent_essays (last 5).
3-STEP CHAIN-OF-THOUGHT:
Step 1 (Baseline Assessment): Positive/negative/neutral? Trending?
Step 2 (Trajectory Projection): Compare recent vs. historical mean.
  Large  (+/-0.3 to +/-0.8): sustained multi-week trend reversals
  Medium (+/-0.1 to +/-0.3): gradual drift
  Small  (+/-0.0 to +/-0.1): stable baseline
Step 3 (Delta): Output FUTURE disposition change.
  Apply regression-to-mean for extremes.
OUTPUT: Valid JSON array only. All values in [-3, 3].
```

Prompt v2 — Sensitivity Ablation. Input: `subtask2b_llm_inputs_v2.json` (adds `recent_trend` and `recent_volatility` per dimension).

```
ROLE: Expert in longitudinal affect forecasting (Russell Circumplex).
  Valence: negative <-> positive  (scale: -3 to +3)
  Arousal: calm <-> excited       (scale: -3 to +3)
TASK: Same as Prompt v1.
ADDITIONAL INPUT (pre-computed):
  recent_trend_valence/arousal: mean(last 5) - historical mean.
    POSITIVE = recent above baseline (primary directional signal).
  recent_volatility_valence/arousal: std of last 5 essays.
    High volatility (>0.8) = user in active flux.
CALIBRATION: std ~0.6 (valence), ~0.4 (arousal).
  Range: -1.5 to +1.5. Near-0.0 predictions should be RARE.
3-STEP CHAIN-OF-THOUGHT:
Step 1 (Signal Assessment):
  |recent_trend| > 0.5  -> bold   (+/-0.6 to +/-1.5)
  0.2--0.5              -> medium (+/-0.3 to +/-0.6)
  < 0.2                 -> small  (+/-0.1 to +/-0.3)
  Near-zero ONLY if flat recent AND flat hist. change.
Step 2 (Direction & Magnitude):
  Direction = sign(recent_trend).
  Apply 30% regression-to-mean if |hist_dispo_change| > 1.0.
Step 3 (Final Check): Verify direction; clamp to [-3, 3].
OUTPUT: Valid JSON array only. All values in [-3, 3].
```