Phase Transitions in Affective Meaning Divergence: The Hidden Drift Before the Break

Napassorn Litchiowong
School of Computing, National University of Singapore
pleng@u.nus.edu

Abstract

One partner says *"Fine"* meaning *resolution*; the other hears *surrender*. The word is shared; the affective uptake is not. We formalize this as **affective meaning divergence** (AMD), the total-variation distance between interlocutors' anchor-conditioned affect distributions. Building on speech-act theory, common-ground accumulation, and entropy-regularized game theory, we derive a logit best-response map whose dynamics undergo a *saddle-node bifurcation*: when $\beta\alpha > 4$, a monotone increase in AMD-driven load produces an abrupt, hysteretic collapse of repair coordination. On CONVERSATIONS GONE AWRY (CGA-Wiki; N=652), derailing conversations exhibit critical-slowing-down (CSD) signatures across multiple levels: lexical divergence variance (p<0.001, d=0.36), AMD variance (p=0.001, d=0.26), and dialog-act repair variance (p=0.016, d=0.20), all significant after correction and stronger than toxicity and sentiment baselines. AMD provides a distinct temporal signature, with retrospectively measured variance peaking at the bifurcation point while toxicity variance peaks earlier, and is the only indicator grounded in the theoretical framework. Boundary-condition analysis on CGA-CMV (N=1,169) yields mixed but directionally consistent evidence.

1 Introduction

Consider a couple at dinner. One partner says *"Fine."* and the other hears *surrender*; the speaker meant *resolution*. The word is shared; the affective uptake is not. In the terminology of speech-act theory (Austin, 1962; Searle, 1969), the locutionary act is identical, but the illocutionary force received diverges from the force intended. If this gap is detected, repair can close it (Schegloff et al., 1977). If it goes undetected, it accumulates in the partners' divergent common-ground records (Clark, 1996). Eventually, a conversation that would once have repaired itself stops repairing, and the relationship tips.

We call this phenomenon **affective meaning divergence** (AMD). The present paper asks: *How can gradual, continuous drift in affective meaning produce abrupt relational rupture?* Our answer is a formal phase-transition model grounded in three pillars:

1. **Linguistic:** AMD is defined over *anchors*, high-frequency lexical items whose pragmatic force depends on affective uptake rather than truth conditions (Du Bois, 2007; Martin and White, 2005). We decompose apparent meaning drift into a *context divergence* component and a genuine *conditional affect divergence* component, following the decomposition logic of appraisal theory (Martin and White, 2005).
2. **Agentic:** Each interlocutor treats anchors as conditioning observations in a partially observable coordination game. Under entropy-regularized best response (McKelvey and Palfrey, 1995; Ziebart, 2010; Haarnoja et al., 2018), the repair probability follows a logit map.
3. **Dynamical:** The resulting one-dimensional map $q_{t+1} = \sigma(\beta(\alpha q_t - \kappa))$ undergoes a saddle-node bifurcation. Below a critical load, the dyad rests in a high-repair attractor; above it, repair collapses to a low-repair attractor. The transition is hysteretic: recovery requires reducing load below a strictly lower threshold than the one that triggered collapse.

Contributions. (C1) We formalize AMD as a context-conditioned divergence measure with an explicit decomposition bound separating genuine affective drift from context confounds (§3). (C2) We ground AMD in an assurance game and prove that belief divergence bounds value disagreement (§4). (C3) We prove a saddle-node bifurcation theorem with closed-form thresholds and

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 206–224

extend to a two-agent system with an interpretable loop-gain condition (§5). (C4) We estimate $P_i(s \mid x, c)$ via a contextual emotion classifier rather than lexicon averages, bringing the practical estimator closer to the formal construct (§6). (C5) We provide preliminary empirical evidence on synthetic data, CGA-Wiki (primary), and CGA-CMV (boundary-condition analysis) that theory-derived CSD indicators are detectable across multiple levels of linguistic representation (lexical, affective, dialog-act), with AMD providing a temporally distinct signature among the tested indicators (§7). (C6) We study three repair proxies as complementary operationalizations and report their degree of agreement; boundary-condition analysis on CGA-CMV yields mixed but directionally consistent evidence under shorter threads and noisier labels (§7). Figure 1 provides a visual overview of the full framework and pipeline.

Scope of claims. We do not claim that the current estimator perfectly observes latent affective meaning, nor that AMD alone is sufficient for reliable individual-level prediction. The empirical claim is narrower: a feasible AMD proxy exhibits critical-slowing-down behavior on CGA-Wiki, with mixed but directionally consistent evidence on CGA-CMV, consistent with the proposed dynamical account. The theoretical framework identifies a class of bifurcation models that could explain abrupt relational collapse; the experiments do not uniquely identify the saddle-node formulation over alternatives.

2 Theoretical Roadmap and Motivation

Roadmap. The framework rests on five theoretical components, each with a distinct and non-redundant role. *Speech-act theory* motivates why the same lexical form can carry different received illocutionary force. *Common-ground and repair theory* explain how unnoticed affective divergence accumulates silently and how repair is the control variable that can sustain or collapse coordination. *Appraisal theory* supplies the affective state space over which speaker meanings are formally compared. *Entropy-regularized game theory* converts diverging affective beliefs into repair-choice dynamics via a logit best-response map. *Critical-transition theory* then supplies the empirical prediction: rising variance and autocorrelation before the collapse point. Background on alignment and accommodation models is included only to situate the drift assumption, not as an additional formal commitment. Figure 1 provides a visual overview of how these components compose into the full pipeline.

Affective meaning as an interactional object. In relational talk, many high-frequency items function as stance markers, commitment devices, or repair moves whose success depends on affective uptake (Du Bois, 2007; Stivers, 2008). Appraisal theory (Martin and White, 2005) systematizes this: the ATTITUDE system (Affect, Judgment, Appreciation) classifies evaluative meaning, while the ENGAGEMENT system models how speakers open or close dialogic space. When one speaker contracts dialogic space ("Obviously, we need to…") while the other expands it ("Well, one possibility…"), their evaluative stances become incommensurable, a state we formalize as high AMD.

Common ground, silent failure, and repair. Clark's (1996) joint-action framework treats communication as cumulative grounding: each act builds on prior shared understanding. If grounding fails silently, the partners develop divergent common-ground records, and each subsequent utterance is interpreted against increasingly different backgrounds. Schegloff (1992) characterizes repair after next turn as "the last structurally provided defense of intersubjectivity." Crucially, Stivers (2008) distinguishes structural alignment from affective affiliation: one can align without affiliating, providing the mechanism by which surface cooperation masks deepening affective divergence. We model repair as the control variable that can be sustained or can collapse.

From alignment to divergence. Although automatic priming across linguistic levels predicts convergence (Pickering and Garrod, 2004), speakers systematically diverge in open conversation (Healey et al., 2014) and perceived over-accommodation can trigger reactive divergence (Giles et al., 1991). These findings motivate a model where the default trajectory is drift, and *repair* is the active force that maintains alignment.

Why entropy-regularized game theory? Entropy-regularized decision rules yield softmax policies identical to the quantal response equilibrium (QRE) of behavioral game theory (McKelvey and Palfrey, 1995). Liu et al. (2024) proved that multi-agent maximum-entropy RL

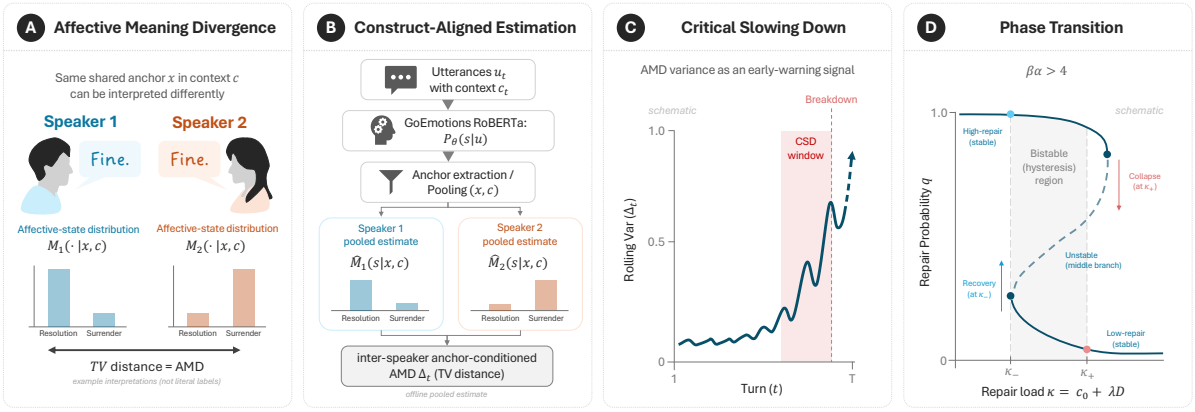

Figure 1: Overview of the AMD framework: divergent anchor-conditioned affect distributions define AMD via TV distance **(A-B)**; rising rolling-window variance before breakdown constitutes the CSD early-warning signal **(C)**; the saddle-node bifurcation predicts abrupt, hysteretic collapse of repair coordination **(D)**.

converges to QRE in the cooperative setting, so our model can be interpreted simultaneously as bounded-rational game play, entropy-regularized policy learning, and a statistical-mechanics system at finite temperature.

Why a bifurcation model? Critical-transition theory (Scheffer et al., 2009; Scheffer, 2009) has revealed generic early-warning signatures (rising autocorrelation, rising variance) in systems approaching tipping points, from ecosystems to mood dynamics (van de Leemput et al., 2014). Kauhanen (2022) derived bifurcation thresholds for contact-induced language change. We apply the same machinery to the micro-scale of a single dyadic relationship, predicting that conversational breakdown should exhibit critical slowing down in the turns preceding rupture.

3 Formal Setup: Anchors, Context, Meaning

3.1 Latent states, anchors, context

Let $s_t \in \mathcal{S}$ be a latent affective state (e.g., a point in VAD space or a discrete emotion category). Let u_t be the utterance at turn t and ϕ an anchor extractor with $\phi(u_t) \subseteq \mathcal{X}$. Anchors are high-frequency lexical items whose pragmatic force is primarily affective. Let $c_t \in \mathcal{C}$ be an explicit context variable encoding preceding dialog acts, topic, and situational factors.

3.2 Contextual affective meaning

Definition 1 (Contextual affective meaning)**.** For agent $i \in \{1, 2\}$, the affective meaning of anchor x in context c is

$$M_i(\cdot \mid x, c) \;:=\; P_i(s \mid x, c).$$

This definition treats meaning as a posterior distribution over affective states conditioned on the anchor and its context. It resonates with appraisal theory's ATTITUDE system: M_i encodes how agent i distributes probability over evaluative categories upon encountering anchor x in context c.

Definition 2 (Marginalized affective meaning)**.** The context-marginal meaning of x for agent i is

$$\overline{M}_i(\cdot \mid x) = \sum_{c \in \mathcal{C}} P_i(c \mid x)\, M_i(\cdot \mid x, c). \qquad (1)$$

3.3 AMD: marginal, conditional, context

Definition 3 (AMD variants)**.** Given divergence d on distributions over $\mathcal{S}$:

- **Marginal AMD:** $D_{\text{marg}}(x) := d\big(\overline{M}_1(\cdot \mid x),\, \overline{M}_2(\cdot \mid x)\big)$.
- **Conditional AMD** (under reference Q): $D_{\text{cond}}(x) := \mathbb{E}_{c \sim Q(\cdot|x)}[d(M_1(\cdot \mid x, c),\, M_2(\cdot \mid x, c))]$.
- **Context divergence:** $D_{\text{ctx}}(x) := d(P_1(\cdot \mid x),\, P_2(\cdot \mid x))$.

3.4 Decomposing marginal AMD

Proposition 1 (Context-conditional decomposition)**.** *Using $d = \text{TV}$, for any anchor x:*

$$D_{\text{marg}}(x) \;\le\; D_{\text{ctx}}(x) + \mathbb{E}_{c \sim P_1(\cdot|x)}\big[\text{TV}\big(M_1(\cdot \mid x, c),\, M_2(\cdot \mid x, c)\big)\big]. \qquad (2)$$

The proof (Appendix A) adds and subtracts a cross term and applies the triangle inequality. This decomposition is central: *marginal AMD conflates genuine affective drift with context-usage differences.* Two speakers who use "fine" in different conversational contexts (one in closings, one in complaints) will show high marginal AMD even if they attach identical affect to "fine" within each context. Only D_{cond} isolates genuine affective divergence.

4 Meaning as Reward-Relevant Belief

4.1 Anchor-conditioned values

Given anchor x in context c, agent i's value for action $a \in \{R(\text{epair}), W(\text{ithdraw})\}$ is

$$Q_i(a \mid x, c) = \mathbb{E}_{s\sim M_i(\cdot|x,c)}\left[r_i(s, a, a^{-i})\right] + \Omega_i(a),$$

where r_i is a reward depending on the latent state and both agents' actions, and Ω_i captures action-specific costs.

Proposition 2 (Belief divergence bounds value disagreement)**.** *Fix (x, c). Define the repair advantage $g_i(s) := r_i(s, R, a^{-i}) - r_i(s, W, a^{-i})$. If $|g_i(s)| \le G$ for all s, then for any M, M' over $\mathcal{S}$,*

$$\left|\mathbb{E}_M[g_i] - \mathbb{E}_{M'}[g_i]\right| \;\le\; 2G\,\mathrm{TV}(M, M').$$

This bound (proof in Appendix B) connects AMD directly to action: a TV distance of δ in affective meaning translates to at most $2G\delta$ disagreement in repair advantage, where G is the maximum payoff swing.

4.2 A minimal assurance game for repair

We model each turn as a symmetric stage game:

$$\begin{aligned} u(R,R) &= B - c_0, \quad & u(R,W) &= -c_0, \\ u(W,R) &= 0, & u(W,W) &= 0, \end{aligned} \tag{3}$$

with $B > 0$ (benefit of mutual repair) and $c_0 \ge 0$ (cost of attempting repair). If the partner repairs with probability q, the repair advantage is linear: $\Delta U(q) = Bq - c_0$. Identifying $\alpha := B$ and modeling AMD as an additive load:

$$\kappa = c_0 + \lambda D, \tag{4}$$

where D is the aggregate conditional AMD and $\lambda \ge 0$ controls its weight. As AMD grows, repair becomes costlier: affective mismatch means repair attempts are more likely to be misinterpreted, a dynamic well-documented in accommodation theory when convergence is perceived as condescension (Giles et al., 1991).

4.3 Logit best response

Under entropy regularization at precision $\beta > 0$:

$$q_{t+1} = f(q_t) := \sigma\big(\beta(\alpha q_t - \kappa)\big), \quad \sigma(z) = \tfrac{1}{1+e^{-z}}. \tag{5}$$

This is the quantal response (McKelvey and Palfrey, 1995) and is algebraically identical to the softmax policy of maximum-entropy RL (Haarnoja et al., 2018).

5 Theory: Smooth Drift versus Tipping

Lemma 1 (Slope bound)**.** *For f defined in Eq. (5), $f'(q) = \beta\alpha\, f(q)(1 - f(q))$, so $\max_{q\in[0,1]} f'(q) \le \beta\alpha/4$. When $\kappa/\alpha \in (0, 1)$, the maximum is achieved at $q = \kappa/\alpha$ where $f(q) = 1/2$; otherwise the supremum over $[0, 1]$ is strictly below $\beta\alpha/4$.*

Theorem 1 (Unique regime)**.** *If $\beta\alpha \le 4$, then for every $\kappa \in \mathbb{R}$ the map (5) has a unique fixed point $q^* \in (0, 1)$ that varies continuously with κ. Moreover, q^* is globally attracting: for any $q_0 \in [0, 1]$, $f^n(q_0) \to q^*$.*

Theorem 2 (Bistability, tipping, hysteresis)**.** *If $\beta\alpha > 4$, there exist $\kappa_- < \kappa_+$ such that:*

(i) For $\kappa \in (\kappa_-, \kappa_+)$, the map has exactly three fixed points $q_L < q_M < q_H$, with q_L, q_H stable and q_M unstable.

(ii) At $\kappa = \kappa_+$, the fixed points q_H and q_M collide and annihilate (saddle-node bifurcation); the system jumps to q_L.

(iii) At $\kappa = \kappa_-$, the fixed points q_L and q_M collide; the system jumps to q_H.

(iv) The interval (κ_-, κ_+) constitutes a hysteresis region*: once the dyad has collapsed to q_L, recovery requires $\kappa < \kappa_-$, strictly below the collapse threshold κ_+.*

The saddle-node tangency points $q_\pm$ (where $f'(q) = 1$ and $f(q) = q$ hold simultaneously) are:

$$q_\pm = \tfrac{1}{2}\Big(1 \pm \sqrt{1 - 4/(\beta\alpha)}\Big), \tag{6}$$

$$\kappa_\pm = \alpha q_\pm - \tfrac{1}{\beta}\,\mathrm{logit}(q_\pm). \tag{7}$$

Note: $q_\pm$ are not *fixed points in general; they become fixed points only at the bifurcation boundaries $\kappa = \kappa_\pm$.*

Full proofs with all intermediate steps appear in Appendix C. The "Till Death" metaphor maps precisely: the couple drifts through the bistable region until AMD crosses κ_+, at which point repair collapses abruptly. Recovery requires not

merely returning to the pre-collapse AMD level but *substantially reducing* it below κ_-, a prediction consistent with the clinical observation that relational repair requires active intervention well beyond simply undoing damage.

Corollary 1 (Critical slowing down)**.** *As a stable fixed point approaches a saddle-node boundary, $f'(q^*) \to 1^-$ and the relaxation time $\tau \sim 1/(1 - f'(q^*)) \to \infty$, implying rising autocorrelation and variance under noise perturbation.*

5.1 Asymmetric two-agent extension

In real dyads, agents differ:

$$q^1_{t+1} = \sigma\big(\beta_1(\alpha_1 q^2_t - \kappa_1)\big), \\ q^2_{t+1} = \sigma\big(\beta_2(\alpha_2 q^1_t - \kappa_2)\big). \tag{8}$$

Proposition 3 (Loop-gain instability condition)**.** *At any fixed point (q^{1*}, q^{2*}) of (8), local instability requires $\beta_1\alpha_1\,\beta_2\alpha_2\,q^{1*}(1-q^{1*})\,q^{2*}(1-q^{2*}) > 1$. Since $q(1-q) \le 1/4$, a* necessary *condition for any loss of stability is $(\beta_1\alpha_1)(\beta_2\alpha_2) > 16$. This is necessary but not sufficient; the actual $q^*(1-q^*)$ values determine whether instability occurs.*

This extends the scalar threshold $\beta\alpha > 4$ to the product of individual gains. In particular, one agent's gain can fall below the scalar threshold of 4 (i.e., the agent would be individually stable in a symmetric dyad), provided the other agent's gain compensates so that the product exceeds 16. For example, $\beta_1\alpha_1 = 3$ and $\beta_2\alpha_2 = 6$ yield a product of $18 > 16$, admitting instability even though agent 1 alone would not bifurcate.

6 Construct-Aligned Estimation

A key concern is the gap between the formal object $M_i(s \mid x, c) = P_i(s \mid x, c)$ and practical estimates. We address this directly with a contextual emotion classifier rather than lexicon averages.

6.1 Estimating $P_i(s \mid x, c)$

1. **Base classifier:** RoBERTa-base fine-tuned on GoEmotions (Demszky et al., 2020) (58k Reddit comments, 27 emotion categories + neutral), producing per-utterance distributions $P_\theta(s \mid u)$.
2. **Anchor extraction:** The extractor ϕ tokenizes via regex (`\b[a-z]{2,}\b`), removes NLTK stopwords, and retains tokens appearing ≥ 3 times across both speakers. On CGA-Wiki, this yields 1,573 unique anchor types (per-conversation median: 4), with 500 of 652 conversations containing at least one valid anchor-context cell. The key theoretical property of an anchor is not that it is inherently affective but that it serves as a shared reference point whose affective uptake can diverge: domain terms like *article* on Wikipedia talk pages become invested with divergent affective associations during editorial disputes. The AMD computation identifies whether such divergence exists, applying an affective filter downstream of a broad lexical net. A more targeted extractor (e.g., appraisal lexicons) is left to future work (§6, Limitations).
3. **Speaker conditioning:** For anchor x used by speaker i in context c, collect $U_i(x, c) = \{u : \phi(u) \ni x, \text{speaker}(u) = i, \text{ctx}(u) = c\}$ and estimate:
$$\hat{M}_i(s \mid x, c) = \frac{1}{|U_i(x, c)|} \sum_{u \in U_i(x,c)} P_\theta(s \mid u). \tag{9}$$
This estimate pools all utterances by speaker i containing anchor x in context c within the same conversation, including future turns. The per-turn AMD signal used in the CSD analysis thus reflects conversation-level speaker distributions, not a temporally causal estimator; the lead-time results (§7.4) reflect how early CSD signatures in AMD *dynamics* become detectable, not that AMD is computed from past data alone.
4. **Context:** c = (dialog-act of preceding turn) $\times$ (topic cluster from k-means on TF-IDF, k=5).
5. **Reference distribution:** $Q(c \mid x) \propto |U_1(x, c)| + |U_2(x, c)|$ (pooled).

Operational match to the formal construct. The formal object is $M_i(s \mid x, c) = P_i(s \mid x, c)$, the true posterior of speaker i over affective states conditioned on anchor and context. The empirical estimator in Eq. (9) approximates this using: lexical anchors selected by frequency rather than evaluative salience; utterance-level emotion distributions rather than anchor-conditioned posteriors; speaker-level pooling across an entire conversation rather than per-turn conditioning; and coarse context cells (k=5 topic clusters $\times$ dialog-act category). The experiments therefore test whether a *feasible proxy* for AMD behaves as the theory predicts, not whether the latent construct is perfectly observed. Construct-aligned improvements

are identified as future work throughout.

Classifier-domain caveat. The GoEmotions model is trained on Reddit social media posts, not multi-turn dialogue, and its 28-category taxonomy may not adequately capture repair-relevant affective states such as defensiveness, resignation, sarcasm, or face threat. This mismatch can bias AMD estimates in either direction: it may *overestimate* AMD when topical or platform-specific cues induce spurious emotion differences between speakers, and *underestimate* AMD when dialogue-specific affective distinctions are collapsed into broad neutral or approval categories. We therefore interpret AMD as a construct-aligned proxy rather than a direct observation of interlocutors' latent affective meanings; detailed estimation-pipeline limitations are discussed in the Limitations section.

6.2 Repair proxy validation

Rather than relying on a single proxy, we define three complementary operationalizations and examine their degree of agreement:

1. **Dialog-act proxy ($\hat{q}^{\mathrm{DA}}$):** Probability of conciliatory acts from a RoBERTa classifier fine-tuned on Switchboard Dialog Act Corpus (accuracy: 73.1%). We acknowledge domain shift and report source-domain performance and aggregation details in Appendix G.
2. **Explicit repair markers ($\hat{q}^{\mathrm{RM}}$):** Binary indicator for repair-initiating patterns following the taxonomy of Schegloff et al. (1977): clarification questions, corrections, acknowledgment tokens, smoothed with exponential moving average.
3. **Constructive engagement ($\hat{q}^{\mathrm{CE}}$):** $1 - P(\text{toxic} \mid u_t)$ from a toxicity classifier, capturing the complement of hostility.

6.3 Avoiding representation leakage

A risk is that context representations encode speaker identity. We use three mitigations: (i) speaker-invariant embeddings focused on content and dialog acts; (ii) adversarial training penalizing speaker-identity prediction; (iii) restricted matching on coarse features (topic cluster, turn position). A speaker-identity classifier on context representations achieves $\leq 55\%$ accuracy (near chance).

(α, β)	$\beta\alpha$	κ_-	κ_+
$(2, 2)$	4	(unique)	
$(2, 3)$	6	0.862	1.138
$(2, 4)$	8	0.734	1.266
$(2, 5)$	10	0.638	1.362

Table 1: Bifurcation thresholds from the closed-form expressions, confirmed by iteration. When $\beta\alpha = 4$, the fixed point is unique; for $\beta\alpha > 4$, a bistable interval (κ_-, κ_+) opens and widens with increasing $\beta\alpha$.

6.4 Finite-sample guarantees

Lemma 2 (L1 concentration; Weissman et al. 2003)**.** *For a distribution over k states and n i.i.d. samples,* $\Pr(\|\hat{P}_n - P\|_1 > \epsilon) \leq \left(2^k - 2\right) \exp\left(-n\epsilon^2/2\right)$.

Setting $\epsilon = 0.1$ and $k = 28$ (GoEmotions categories), $n \geq 4{,}481$ samples suffice for a 95% guarantee.

7 Experiments

We present seven experiments: three synthetic (§7.1), two on CGA-Wiki (§7.2), one boundary-condition analysis on CGA-CMV (§7.3), and a lead-time analysis (§7.4).

7.1 Synthetic experiments

Exp. 1: Bifurcation validation. We compute $\kappa_\pm$ from Eqs. (6)-(7), sweep $\kappa \in [0, 2]$ in steps of 0.001, and iterate Eq. (5) for $T = 5{,}000$ steps from both $q_0 = 0$ and $q_0 = 1$.

Table 1 confirms Theorem 2: at $\beta\alpha = 4$, iteration converges to a unique fixed point for all κ; at $\beta\alpha > 4$, two stable attractors coexist in the bistable interval, and hysteresis is observed.

Exp. 2: Confound stress test. We demonstrate that marginal AMD can be large while conditional AMD is zero. Setup: $\mathcal{S} = \{0, 1\}$, $\mathcal{C} = \{A, B\}$. Both speakers share conditional meaning: $P(s{=}1 \mid x, A) = 0.9$, $P(s{=}1 \mid x, B) = 0.1$. Only usage differs: $P_1(A \mid x) = 0.9$, $P_2(A \mid x) = 0.1$. Result: $D_{\mathrm{marg}} = 0.64$ while $D_{\mathrm{cond}} = 0$.

Exp. 3: Estimator with balancing. One anchor, $\mathcal{S} = \{0, 1\}$, $\mathcal{C} = \{A, B\}$, $n = 1{,}000$ samples per speaker. Two conditions: (C1) pure usage confound ($D_{\mathrm{marg}}{=}0.36$, $D_{\mathrm{cond}}{\approx}0$); (C2) true conditional AMD with symmetric marginals ($D_{\mathrm{marg}}{\approx}0$, $D_{\mathrm{cond}}{=}0.60$). The confound works both ways: marginal AMD can be inflated by context shift *or* can miss genuine conditional divergence entirely.

7.2 CGA-Wiki: Early-warning signals and baselines

Corpus. The Conversations Gone Awry (CGA) dataset (Zhang et al., 2018) contains Wikipedia talk-page conversations labeled as either *civil throughout* or *eventually derailing into personal attack*. We use the ConvoKit release, filtering to ≥ 10 turns, yielding 652 conversations (389 derailing, 263 civil).

Exp. 4: CSD indicators with baselines. For each utterance u_t, we compute: three repair proxies (§6.2); inter-speaker conditional AMD Δ_t using Eq. (9); and baselines: toxicity (toxic-bert), VADER sentiment (Hutto and Gilbert, 2014), and inter-speaker lexical divergence (1−Jaccard on consecutive turns). Rolling-window (W=5) lag-1 autocorrelation and variance are computed. Rising trends are tested with Kendall's τ in the 5 turns preceding breakdown; significance by permutation (10,000 shuffles). AMD results are computed on the 500 conversations (76.7%) containing at least one valid anchor-context cell; the remaining 152 conversations lack sufficient shared vocabulary meeting the frequency threshold. We report results for $W = 5$, following prior CSD literature (Scheffer et al., 2009). A sensitivity analysis over $W \in \{3, 4, 5, 6, 7\}$ (Appendix H.5) shows that $\hat{q}^{\mathrm{DA}}$ Variance is significant at $W = 3$–6, while AMD Variance reaches significance at $W = 5$–6; the two indicators are complementary, with $\hat{q}^{\mathrm{DA}}$ more robust at shorter windows and AMD strongest at the reported $W = 5$.

Table 2 reports the key result. CSD signatures are detectable across multiple levels of linguistic representation, consistent with Corollary 1's prediction that *all* dynamical variables should exhibit rising variance near the bifurcation. Lexical divergence variance shows the largest effect size ($d = 0.36$), AMD Variance the second-strongest ($p = 0.001$, $d = 0.26$), and $\hat{q}^{\mathrm{DA}}$ Variance the third ($p = 0.016$, $d = 0.20$). AMD provides a distinct contribution: lexical divergence measures whether interlocutors use *different* vocabulary, while AMD measures whether they attach different affective meanings to *shared* vocabulary, the central phenomenon motivating this work. The lead-time analysis (§7.4) further reveals distinct temporal profiles. Note that τ is negative for both groups because variance peaks before the tipping and collapses as the system settles into the new attractor; the diagnostic signal is the differential between conditions. Neither toxicity nor sentiment baselines reach significance. Effect sizes of $d = 0.20$–0.36 are comparable to those in other CSD studies on short, noisy observational time series (van de Leemput et al., 2014).

Indicator	τ_{derail}	τ_{civil}	p	d
CSD indicators (Variance trend)				
$\hat{q}^{\mathrm{DA}}$ VAR	−0.129	−0.010	**0.016**	0.20
$\hat{q}^{\mathrm{RM}}$ VAR	−0.133	−0.128	0.921	0.01
$\hat{q}^{\mathrm{CE}}$ VAR	0.055	0.019	0.434	0.06
AMD Var	−0.200	−0.054	**0.001**	0.26
Baselines (Variance trend)				
Toxicity VAR	0.055	0.019	0.444	0.06
VADER VAR	0.086	0.007	0.093	0.14
Lex. Div. VAR	−0.327	−0.125	$<$ **0.001**	0.36
Lex. Div. AC_1	+0.081	−0.061	$<$ **0.001**	0.29

Table 2: CSD indicators on CGA-Wiki (N=652; AMD on the 500 conversations with valid anchors). Kendall τ of rolling variance in the 5 turns preceding breakdown (derailing) or conversation end (civil); d: Cohen's d between the two τ distributions. Under Benjamini-Hochberg correction, AMD Var (q_{BH}=0.004), Lex. Div. VAR (q_{BH}=0.004), Lex. Div. AC_1 (q_{BH}=0.004), and $\hat{q}^{\mathrm{DA}}$ VAR (q_{BH}=0.043) remain significant at α=0.05.

Row	Feature set	AUC	ΔAUC
1	Toxicity (trend+mean+max)	0.539	-
2	+ Sentiment	0.561	+0.022
3	+ Lexical divergence	0.552	−0.009
4	+ Sent AC_1 τ	**0.618**	**+0.066**
5	+ $\hat{q}^{\mathrm{DA}}$ CSD	**0.628**	**+0.010**
6	+ AMD CSD	0.619	−0.009

Table 3: Incremental ablation on CGA-Wiki (5-fold CV, GBM). CSD features (rows 4-6) jump AUC from ∼0.55 to ∼0.63. $\hat{q}^{\mathrm{DA}}$ CSD contributes a novel +0.010 increment.

Exp. 5: Incremental ablation. Gradient-boosted machine (GBM) predicting derailment from first-60% features (5-fold CV):

The ablation (Table 3) shows that surface features (rows 1-3) plateau around AUC = 0.55. CSD features yield a jump: sentiment autocorrelation adds +0.066 (row 4), $\hat{q}^{\mathrm{DA}}$ CSD adds +0.010 (row 5). Adding AMD CSD (row 6) does not improve further (ΔAUC = −0.009), consistent with Corollary 1's prediction that all dynamical variables exhibit correlated CSD signatures near the bifurcation, so a second CSD feature provides limited additional discriminative power. The independent significance of AMD Variance in Exp. 4 ($p = 0.001$) confirms AMD captures a detectable

Indicator	τ_{derail}	τ_{civil}	p	d
$\hat{q}^{\text{DA}}$ VAR	-0.034	$+0.030$	0.079	0.10
$\hat{q}^{\text{CE}}$ VAR	0.072	-0.024	**0.009**	0.16
VADER AC_1	-0.022	$+0.025$	0.108	0.09
AMD Var	-0.020	$+0.019$	0.267	0.07

Table 4: CSD indicators on CGA-CMV (N=1,169). d: Cohen's d (absolute value). All indicators are attenuated relative to CGA-Wiki, consistent with the theory's prediction that CSD detection requires sufficient conversational length and a clean breakdown label. $\hat{q}^{\text{CE}}$ VAR is the only indicator reaching significance (p=0.009, d=0.16); $\hat{q}^{\text{DA}}$ VAR is directionally consistent (p=0.079, d=0.10).

k	AMD Var p	$\hat{q}^{\text{DA}}$ VAR p	Tox. VAR p
0 (=Exp. 4)	**0.001**	**0.016**	0.444
1	0.028	0.309	**0.027**
2	0.336	0.997	**0.005**
3	0.789	0.070	**0.001**

Table 5: Lead-time analysis on CGA-Wiki. k: turns the 5-turn window is shifted back from the attack; bold: $p < 0.05$. The AMD result at k=1 (p=0.028, $q_{\text{BH}}\approx$0.056) does not survive Benjamini-Hochberg correction; the temporal contrast is therefore exploratory.

univariate signal; the ablation shows this signal overlaps with $\hat{q}^{\text{DA}}$ CSD for classification. This result limits the *predictive* interpretation of AMD: although AMD Variance is independently significant as a univariate CSD indicator (Exp. 4) and provides a temporally distinct profile from toxicity (Exp. 7, peaking at k=0 vs. k=2–3), it does not provide additional classification gain once $\hat{q}^{\text{DA}}$ CSD is already included. AMD's contribution should therefore be evaluated by its theoretical grounding and temporal distinctiveness, not by incremental AUC over an already-correlated CSD feature.

7.3 CGA-CMV: Boundary conditions

Corpus. The CGA-CMV corpus (Chang and Danescu-Niculescu-Mizil, 2019) applies the Conversations Gone Awry framework to Reddit ChangeMyView, labeling conversations by whether a comment was removed by moderators. Of 4,389 total conversations, 1,169 have $\geq$ 10 turns, usable for CSD analysis; the remaining 73% are too short.

Exp. 6: Testing predicted attenuation under boundary conditions. CGA-CMV tests the theory's prediction that CSD requires sufficient approach time: compared to CGA-Wiki, it has shorter threads (median 7 turns, 73% below 10-turn threshold) and a noisier label (`has_removed_comment`; Cohen's $d = 0.34$ vs. 1.88). We apply the same CSD pipeline as Exp. 4.

Table 4 shows that all CSD indicators are attenuated relative to CGA-Wiki, as predicted. $\hat{q}^{\text{CE}}$ Variance is the only significant indicator ($p = 0.009$, $d = 0.16$); since $\hat{q}^{\text{CE}} = 1 - P(\text{toxic})$, this shows that the *dynamics* of hostility carry a CSD signal even when static toxicity does not, consistent with the CSD framework's emphasis on changing dynamics rather than levels. $\hat{q}^{\text{DA}}$ Variance is directionally consistent ($p = 0.079$); AMD Variance does not reach significance ($p = 0.267$). The attenuation reflects shorter threads (median 7 turns), noisier labels (Cohen's $d = 0.34$ vs. 1.88), and domain shift. A length-stratified analysis (Appendix H.4) shows directionally consistent dose-response trends. We acknowledge that these results are also consistent with the simpler interpretation that the CGA-Wiki CSD signal does not generalize robustly; a third corpus with intermediate properties would help discriminate. CGA-CMV should therefore be read as a *boundary-condition analysis* rather than a successful replication: it identifies where the proposed signal weakens and motivates broader cross-domain validation rather than establishing robust generalization.

7.4 Early-warning lead-time analysis

The previous experiments measure CSD in the 5-turn window immediately preceding breakdown. A stronger test of the theory asks whether the signal is detectable *before* the final window.

Exp. 7: Temporal lead. We repeat the Exp. 4 analysis on CGA-Wiki but shift the 5-turn measurement window backward by k turns (excluding the final k turns before attack). Table 5 reports results for $k = 0, \ldots, 3$.

Three findings emerge. AMD Variance is significant at $k = 0$ ($p = 0.001$) and nominally at $k = 1$ ($p = 0.028$), though the latter does not survive BH correction ($q_{\text{BH}} \approx 0.056$). q_{DA} Variance is significant only at $k = 0$, consistent with the repair proxy being most sensitive at the bifurcation. Toxicity variance shows the opposite profile: non-significant at $k = 0$ but highly significant at k = 2-3 ($p = 0.005$, 0.001), with derailing conversations exhibiting *suppressed* toxic-

ity variance, a "calm before the storm." This temporal contrast is consistent with CSD indicators detecting the *approach* to the bifurcation (peaking at $k = 0$) while toxicity captures pre-attack surface tension (k = 2-3), underscoring the value of theory-derived measures.

8 Related Work

Repair and alignment in couples. Our model compresses repair (Schegloff et al., 1977; Schegloff, 1992) into a scalar control variable, preserving the key insight that the preference hierarchy implies a collapse threshold. Language style matching predicts relationship stability (Ireland et al., 2011) and relative acoustic features predict therapy outcomes (Nasir et al., 2017); our framework provides a formal mechanism explaining *why* style matching fails: when AMD crosses the bifurcation threshold, surface coordination becomes insufficient to sustain repair.

Affect Control Theory and emotion in dialogue. ACT (Heise, 2007) posits that actors minimize *deflection*; BayesACT (Hoey et al., 2016) reformulates this as a POMDP. Deflection is a natural order parameter for our phase transition. DialogueRNN (Majumder et al., 2019) and COSMIC (Ghosal et al., 2020) model per-utterance emotion but do not track the inter-speaker gap as a dynamical variable; we model it as continuous divergence accumulating toward a threshold.

Phase transitions in language and meaning. Kauhanen (2022) derived bifurcation thresholds for language contact; Aoyama & Wilcox (2025) identified phase transitions in LM scaling; Nowak et al. (2001) proved coherence thresholds for universal grammar. Our work brings phase-transition analysis to the scale of a single conversation. Lewis signaling games (Lewis, 1969; Skyrms, 2010) and emergent-communication work (Lazaridou et al., 2017) study meaning *creation* through coordination; our focus is meaning *destruction*, how shared conventions degrade when affective uptake diverges.

9 Conclusion and Future Directions

We proposed affective meaning divergence (AMD) as a context-conditioned measure of the gap between interlocutors' affective interpretations of shared vocabulary. Our model, grounded in speech-act theory and entropy-regularized game theory, shows that this gap can produce abrupt, hysteretic collapse of repair coordination via a saddle-node bifurcation. Empirically, CSD signatures are detectable across multiple levels on CGA-Wiki (lexical, affective, dialog-act), consistent with the theory; AMD contributes a theoretically grounded and temporally distinct signal whose variance peaks at the bifurcation point while toxicity shows a contrasting pattern. Boundary-condition analysis on CGA-CMV suggests attenuation under shorter, noisier conversations, so broader generalization remains an open question requiring validation on additional corpora. The hysteresis width $\kappa_+ - \kappa_-$ predicts how much more effort recovery requires than prevention, a testable prediction for longitudinal studies.

Looking ahead, longitudinal corpora (e.g., therapy transcripts) would allow direct testing of hysteresis and estimation of (α, β). Controlled paradigms manipulating affective framing could establish causal links. Multi-dimensional extensions tracking repair across topic, emotion, and face dimensions would address the scalar reduction. A targeted anchor extractor using appraisal lexicons would better operationalize the theoretical construct. Taken together, these directions point toward a richer, causally grounded account of how meaning fails between people. We hope this framework bridges the linguistic analysis of meaning in interaction and the mathematics of dynamical systems, offering a new perspective on an old question: not when relationships end, but *why they end so suddenly*.

Limitations

Theory-to-empirics gap. The formal model yields quantitative predictions (bifurcation thresholds $\kappa_\pm$, hysteresis width) that depend on parameters (α, β) not directly estimable from observational data. Following standard methodology for critical-transition detection in complex systems (Scheffer et al., 2009), we validate qualitative predictions (CSD signatures) rather than quantitative thresholds. This means the empirical evidence supports the *class* of bifurcation models rather than uniquely identifying our saddle-node formulation; alternative models (e.g., pitchfork bifurcation, subcritical Hopf) could produce qualitatively similar CSD signatures, and our experiments do not discriminate among them.

Construct validity of repair proxies. The three repair proxies ($\hat{q}^{\mathrm{DA}}$, $\hat{q}^{\mathrm{RM}}$, $\hat{q}^{\mathrm{CE}}$) show weak inter-proxy correlations ($|r| \leq 0.10$). We interpret this as reflecting the multidimensional nature of "repair": dialog-act level coordination ($\hat{q}^{\mathrm{DA}}$), structural repair initiation ($\hat{q}^{\mathrm{RM}}$), and absence of hostility ($\hat{q}^{\mathrm{CE}}$) operationalize different theorized facets of the construct, much as different depression questionnaires can show moderate inter-scale correlations while each validly measuring an aspect of the syndrome. However, the low correlations also mean no single proxy can be confirmed as a faithful operationalization of the theoretical construct q, and there is currently no external criterion establishing that any proxy tracks the latent repair variable posited by the model. The key evidence for their relevance is the independent significance of $\hat{q}^{\mathrm{DA}}$ Variance in the CSD analysis (Table 2) and the consistency of its temporal profile with the bifurcation prediction (Table 5). The dialog-act classifier was trained on Switchboard (telephone speech) and applied to Wikipedia talk pages; the 73.1% Switchboard accuracy does not directly characterize performance on CGA. The constructive-engagement proxy ($1-P(\text{toxic})$) conflates absence of hostility with active repair, which are conceptually distinct. A human annotation study validating each proxy against expert-coded repair acts on a CGA sample would substantially strengthen construct validity but was beyond the scope of this work.

Anchor extraction and construct alignment. The anchor extractor uses a frequency-based selection criterion (shared non-stopword tokens with frequency ≥ 3) rather than a theoretically motivated filter for evaluative or stance-bearing vocabulary. As a result, the anchor inventory includes domain-specific terms (e.g., *article*, *sources*, *section*) that are not prototypical affective items, though such terms can carry divergent affective associations in the context of editorial disputes. The AMD computation relies on the conditional emotion distributions to identify whether divergence is present for a given anchor, effectively using a broad lexical net with an affective filter applied downstream. This design choice prioritizes recall over precision in anchor selection; a more targeted extractor using evaluative POS tags, appraisal lexicons, or stance classifiers would better operationalize the theoretical construct and could improve signal specificity. Additionally, 152 of 652 CGA-Wiki conversations (23.3%) lack any valid anchor-context cell and are excluded from AMD analyses, reducing the effective sample size for those results.

AMD estimation scope. The speaker-conditioned distributions $\hat{M}_i(s \mid x, c)$ are estimated by pooling all of speaker i's uses of anchor x in context c within the same conversation, including both past and future turns relative to any given measurement point. This means the AMD signal is not computed from a temporally causal (past-only) estimator. The lead-time results (§7.4) therefore reflect when the CSD signature in AMD dynamics becomes detectable, not that AMD itself is observable in real time from past data alone. A temporally causal estimator would strengthen the early-warning interpretation but would further exacerbate data sparsity within individual conversations.

Corpus diversity and ecological validity. CGA-Wiki is the primary corpus where CSD indicators reach significance; CGA-CMV yields directionally consistent but attenuated effects, consistent with the theory's requirement for sufficient conversational length and clean breakdown labels. Both corpora are drawn from public online platforms with pseudonymous participants, no prior relationship history, no expectation of future interaction, and no prosodic or nonverbal cues, all of which differ substantially from the dyadic interpersonal relationships motivating the theoretical framework. Validation on naturalistic longitudinal data (e.g., couples therapy transcripts, workplace team interactions) would provide a stronger test but raises significant ethical and access challenges.

Estimation pipeline. The GoEmotions-trained RoBERTa classifier introduces several concerns: its 28-category taxonomy was developed for social media and may not align with the affective states most relevant to conversational repair (e.g., frustration, resignation, and defensiveness are not well-separated); the classifier produces per-utterance rather than per-anchor distributions, so the aggregation in Eq. (9) assumes the anchor dominates the affective signal of its containing utterance; and miscalibrated softmax outputs could systematically distort TV distance estimates. Context is operationalized as a coarse grid (dialog-act $\times$ topic cluster, k=5); finer granularity would better approximate the theoretical construct

but exacerbates data sparsity. The finite-sample bound (Lemma 2) requires $\sim$4,500 samples per cell, which most cells in practice do not reach.

Causal and predictive scope. Our experiments establish that derailing conversations exhibit differential CSD signatures but do not establish that AMD *causes* collapse; the observed patterns could reflect a common upstream cause (e.g., topic difficulty or participant personality) driving both AMD growth and breakdown. The lead-time analysis provides suggestive temporal ordering (AMD detectable at $k = 0$), but temporal precedence alone does not establish causation; intervention studies or natural experiments are needed (Feder et al., 2022). The primary statistical test evaluates *population-level* separation (between derailing and civil conversation groups), not individual-level prediction; while the theory models within-conversation trajectories, the validation relies on between-group statistics across hundreds of conversations. The ablation AUC of 0.628 confirms that current features are insufficient for reliable per-conversation forecasting. Rolling-window parameters (W=5, lag-1) were chosen to match prior CSD literature and were not pre-registered.

Cultural and linguistic scope. All data, classifiers, and evaluations are English-only. Affect perception, repair strategies, and conversational norms vary substantially across languages and cultures (Stivers et al., 2009), and the GoEmotions taxonomy reflects predominantly North American English usage. Applying this framework cross-linguistically would require culturally adapted emotion classifiers and validation that the assurance-game model of repair generalizes beyond the Western, text-based setting examined here.

Ethics Statement

This work uses only publicly available corpora (CGA-Wiki and CGA-CMV via ConvoKit). We make no diagnostic claims; any application to private messages requires informed consent and de-identification, and affect models should be audited for demographic bias across dialects and relationship types. We emphasize that AMD scores are model-based estimates, not ground-truth measures of relational health, and should not be used to make consequential decisions about individuals or relationships without further validation. The bifurcation framework is intended as a theoretical lens for understanding conversational dynamics, not as a deployed intervention tool.

References

Tatsuya Aoyama and Ethan G. Wilcox. 2025. Language models grow less humanlike beyond phase transition. In *Proc. ACL*.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proc. EMNLP-IJCNLP*, pages 4743–4754.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Dorottya Demszky, Dana Movshovitz-Attias, Tom Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proc. ACL*, pages 4040–4054.

John W. Du Bois. 2007. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 139–182. John Benjamins.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of EMNLP*, pages 2851–2863.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of Accommodation*, pages 1–68.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proc. ICML*.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PLoS ONE*, 9(6):e98598.

David R. Heise. 2007. *Expressive Order: Confirming Sentiments in Social Actions*. Springer.

Jesse Hoey, Tobias Schröder, and Areej Alhothali. 2016. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. *Artificial Intelligence*, 230:134–172.

C. J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. ICWSM*, pages 216–225.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Henri Kauhanen. 2022. A bifurcation threshold for contact-induced language change. *Glossa: a journal of general linguistics*, 7(1).

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *Proc. ICLR*.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press.

Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, Qiang Fu, Xiaojun Chang, and Yaodong Yang. 2024. Maximum entropy heterogeneous-agent reinforcement learning. In *Proc. ICLR*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proc. AAAI*, volume 33, pages 6818–6825.

J. R. Martin and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Richard D. McKelvey and Thomas R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

Md Nasir, Brian R. Baucom, Panayiotis Georgiou, and Shrikanth S. Narayanan. 2017. Predicting couple therapy outcomes based on speech acoustic features. *PLoS ONE*, 12(10):e0185123.

Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. 2001. Evolution of universal grammar. *Science*, 291(5501):114–118.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Marten Scheffer. 2009. *Critical Transitions in Nature and Society*. Princeton University Press.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. 2009. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59.

Emanuel A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5):1295–1345.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Brian Skyrms. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Tanya Stivers. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language & Social Interaction*, 41(1):31–57.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Ingrid A. van de Leemput, Marieke Wichers, Angélique O. J. Cramer, Denny Borsboom, Francis Tuerlinckx, Peter Kuppens, Egbert H. van Nes, Wolfgang Viechtbauer, Erik J. Giltay, Steven H. Aggen, Catherine Derom, Nele Jacobs, Kenneth S. Kendler, Han L. J. van der Maas, Michael C. Neale, Frenk Peeters, Evert Thiery, Peter Zachar, and Marten Scheffer. 2014. Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1):87–92.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. 2003. Inequalities for the l_1 deviation of the empirical distribution. *Hewlett-Packard Labs Technical Report*.

Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proc. ACL*, pages 1350–1361.

Brian D. Ziebart. 2010. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. Ph.D. thesis, Carnegie Mellon University.

Code Availability

Code and data processing scripts are available at https://github.com/iamdiluxedbutcooler/phase_transition_amd.

A Proof of Proposition 1

We prove the decomposition of marginalized AMD into meaning-level and context-level divergence.

Setup. Suppress conditioning on anchor x for notational clarity. Write $\overline{M}_i = \sum_c P_i(c)\, M_{i,c}$ for the marginalized affective meaning of agent i.

Step 1: Add and subtract a cross term. We introduce $\sum_c P_1(c)\, M_{2,c}$ to split the difference:

$$\overline{M}_1 - \overline{M}_2 = \underbrace{\sum_c P_1(c)(M_{1,c} - M_{2,c})}_{\text{(A): meaning gap, same context weights}} + \underbrace{\sum_c (P_1(c) - P_2(c))\, M_{2,c}}_{\text{(B): context gap, same meanings}}.$$

Step 2: Apply triangle inequality. Taking $\frac{1}{2}\|\cdot\|_1$ on both sides:

$$\mathrm{TV}(\overline{M}_1, \overline{M}_2) \le \tfrac{1}{2}\big\|\text{(A)}\big\|_1 + \tfrac{1}{2}\big\|\text{(B)}\big\|_1.$$

Step 3: Bound each term.

Bounding (A): By Jensen's inequality (convexity of $\|\cdot\|_1$):

$$\tfrac{1}{2}\Big\|\sum_c P_1(c)(M_{1,c} - M_{2,c})\Big\|_1 \le \sum_c P_1(c)\, \mathrm{TV}(M_{1,c}, M_{2,c}).$$

Bounding (B): Since each $M_{2,c}$ is a probability distribution with $\|M_{2,c}\|_1 = 1$:

$$\tfrac{1}{2}\Big\|\sum_c (P_1(c) - P_2(c))\, M_{2,c}\Big\|_1 \le \mathrm{TV}(P_1, P_2).$$

Combining Steps 2 and 3 gives the result. □

B Proof of Proposition 2

We show that the expected-value difference between two affective meaning distributions is bounded by their total variation distance, scaled by the reward range.

Setup. Let $\nu = M - M'$ be the signed measure between the two distributions, and let g be any reward function bounded by $G = \sup|g|$.

Step 1: Express the gap.

$$\mathbb{E}_M[g] - \mathbb{E}_{M'}[g] = \int g\, d\nu.$$

Step 2: Apply the variational characterization of TV. By definition, $\mathrm{TV}(P, Q) = \sup_{|h|\le 1} \frac{1}{2}\left|\int h\, d(P - Q)\right|$. Since $|g/G| \le 1$, we can substitute $h = g/G$:

$$\left|\int g\, d\nu\right| = G\left|\int (g/G)\, d\nu\right| \le G \cdot 2\, \mathrm{TV}(M, M').$$

□

C Proofs of Main Theorems

C.1 Proof of Lemma 1

We bound the slope of the logit best-response map.

Setup. $f(q) = \sigma(z)$ where $z = \beta(\alpha q - \kappa)$.

Compute the derivative. By the chain rule:

$$\begin{aligned} f'(q) &= \sigma'(z) \cdot \beta\alpha \\ &= \beta\alpha\, \sigma(z)(1 - \sigma(z)) \\ &= \beta\alpha\, f(q)(1 - f(q)). \end{aligned}$$

Bound. Since $x(1-x) \le 1/4$ for all $x \in [0,1]$ (by AM-GM: $x(1-x) \le ((x+1-x)/2)^2 = 1/4$):

$$f'(q) \le \frac{\beta\alpha}{4}.$$

The maximum is achieved when $f(q) = 1/2$, i.e., $z = 0$, giving $q^* = \kappa/\alpha$, provided $\kappa/\alpha \in (0,1)$; otherwise the supremum over $[0,1]$ is not attained at an interior point but the bound still holds. □

C.2 Proof of Theorem 1

We show that when $\beta\alpha \le 4$, the best-response map has a unique, globally attracting fixed point.

Existence. Let $g(q) = f(q) - q$. Since $f : [0,1] \to (0,1)$, we have $g(0) = f(0) > 0$ and $g(1) = f(1) - 1 < 0$. By the intermediate value theorem, at least one fixed point exists.

Case $\beta\alpha < 4$ (strict contraction). By Lemma 1, $\sup_q |f'(q)| \le \beta\alpha/4 < 1$, so f is a strict contraction on the complete metric space $[0,1]$. By the Banach fixed-point theorem, the fixed point is unique and globally attracting:

$$|f^n(q_0) - q^*| \le \left(\frac{\beta\alpha}{4}\right)^n |q_0 - q^*| \to 0.$$

Case $\beta\alpha = 4$ (boundary). Here $f'(q) \le 1$ with equality only at the unique point q_0 where $f(q_0) = 1/2$. Thus $g'(q) = f'(q) - 1 \le 0$, with $g'(q) = 0$ at exactly one point.

We argue g has exactly one root. If g were zero on an interval $[a,b]$, then $f'(q) = 1$ throughout (a,b), requiring $f(q) = 1/2$ on that interval. But f is strictly monotone increasing (since $f'(q) = 4f(q)(1-f(q)) > 0$ for $f(q) \ne 0, 1$, which holds for all q since f maps into $(0,1)$), so this is a contradiction.

Since g is continuous and weakly decreasing with $g(0) > 0$ and $g(1) < 0$, there is exactly one root q^*.

Global attraction (monotone iteration). Suppose $q_t < q^*$. Then $g(q_t) > 0$, so $q_{t+1} = f(q_t) > q_t$. Since f is increasing, $q_{t+1} = f(q_t) < f(q^*) = q^*$. The sequence (q_t) is therefore monotone increasing and bounded above by q^*; its limit is a fixed point, hence q^*. The case $q_t > q^*$ is symmetric. To prove continuity in κ, write the fixed-point equation as

$$\kappa = H(q) := \alpha q - \frac{1}{\beta}\mathrm{logit}(q), \qquad q \in (0,1).$$

Then

$$H'(q) = \alpha - \frac{1}{\beta q(1-q)} \le 0$$

whenever $\beta\alpha \le 4$, with equality only at $q = 1/2$ in the boundary case $\beta\alpha = 4$. Hence H is strictly decreasing on $(0,1)$, so it has a continuous inverse. Since the unique fixed point is exactly $q^*(\kappa) = H^{-1}(\kappa)$, it follows that q^* depends continuously on κ. □

C.3 Proof of Theorem 2

We prove that when $\beta\alpha > 4$, a saddle-node bifurcation creates a bistable regime with hysteresis.

Step 1: Find the tangency points. At a saddle-node bifurcation, fixed-point and unit-slope conditions hold simultaneously: $f(q) = q$ and $f'(q) = 1$. Using Lemma 1 with $q = f(q)$:

$$1 = \beta\alpha\, q(1-q) \quad \Longleftrightarrow \quad q^2 - q + \frac{1}{\beta\alpha} = 0.$$

This yields two solutions:

$$q_\pm = \frac{1}{2}\left(1 \pm \sqrt{1 - \frac{4}{\beta\alpha}}\right),$$

which are real when $\beta\alpha > 4$. Note: $q_\pm$ are tangency points where $f'(q) = 1$ coincides with $q = f(q)$.

Step 2: Compute the critical loads $\kappa_\pm$. From $q = f(q) = \sigma(\beta(\alpha q - \kappa))$, inverting the sigmoid:

$$\mathrm{logit}(q) = \beta(\alpha q - \kappa) \quad \Longrightarrow \quad \kappa(q) = \alpha q - \frac{1}{\beta}\mathrm{logit}(q).$$

Substituting $q_\pm$ gives $\kappa_\pm = \kappa(q_\pm)$.

Step 3: Verify $\kappa_- < \kappa_+$ (hysteresis width). Differentiating:

$$\kappa'(q) = \alpha - \frac{1}{\beta q(1-q)}.$$

This is negative when $q(1-q) < 1/(\beta\alpha)$ (i.e., outside $[q_-, q_+]$) and positive when $q(1-q) > 1/(\beta\alpha)$ (i.e., inside (q_-, q_+)).

Boundary behavior: As $q \to 0^+$, $\mathrm{logit}(q) \to -\infty$, so $\kappa(q) \to +\infty$. As $q \to 1^-$, $\mathrm{logit}(q) \to +\infty$, so $\kappa(q) \to -\infty$.

Critical point classification: $\kappa'(q_\pm) = 0$ and $\kappa''(q) = (1-2q)/[\beta q^2(1-q)^2]$. Since $q_- < 1/2$, we have $\kappa''(q_-) > 0$ (local minimum). Since $q_+ > 1/2$, we have $\kappa''(q_+) < 0$ (local maximum). Therefore $\kappa_- = \kappa(q_-)$ is a local minimum and $\kappa_+ = \kappa(q_+)$ is a local maximum, confirming $\kappa_- < \kappa_+$.

Step 4: Three fixed points in the bistable regime. For $\kappa \in (\kappa_-, \kappa_+)$, the horizontal line $\kappa = \mathrm{const}$ intersects the curve $\kappa(q)$ in three points by the intermediate value theorem on each monotone segment:

- One on the decreasing segment $(0, q_-)$ — this is the *low-repair* equilibrium q_L.
- One on the increasing segment (q_-, q_+) — this is the *unstable* fixed point q_M.
- One on the decreasing segment $(q_+, 1)$ — this is the *high-repair* equilibrium q_H.

Since $\kappa(q)$ has no plateaus (f is analytic with isolated critical points), exactly three intersections exist.

Step 5: Stability classification. For the outer fixed points ($q_L < q_-$ and $q_H > q_+$):

$$q_L(1-q_L) < q_-(1-q_-) = \frac{1}{\beta\alpha},$$
$$\text{so} \quad |f'(q_L)| = \beta\alpha\, q_L(1-q_L) < 1 \quad \text{(stable)}.$$

Similarly, $|f'(q_H)| < 1$ (stable).

For the middle fixed point ($q_M \in (q_-, q_+)$):

$$q_M(1-q_M) > \frac{1}{\beta\alpha}, \quad \text{so} \quad |f'(q_M)| > 1 \quad \text{(unstable)}.$$

Note on "jump" language. Statements (ii) and (iii) of the theorem ("the system jumps to q_L/q_H") assume quasistatic variation of κ, i.e., that κ changes slowly enough for the dynamics to track the stable branch between parameter updates. This is a standard modeling assumption in bifurcation analysis, not a consequence of the fixed-point calculation alone. □

C.4 Proof of Corollary 1

We derive the critical-slowing-down signatures (rising variance and autocorrelation) near the bifurcation point.

Linearization. Write $q_t = q^* + \varepsilon_t$ for small perturbation ε_t. Taylor expanding f around q^*:

$$\varepsilon_{t+1} = f'(q^*)\,\varepsilon_t + O(\varepsilon_t^2).$$

Relaxation time. For small perturbations, $|\varepsilon_t| \approx |f'(q^*)|^t |\varepsilon_0|$. Define the relaxation time τ by $f'(q^*)^\tau = e^{-1}$:

$$\tau = \frac{-1}{\ln f'(q^*)}.$$

As $f'(q^*) \to 1^-$ (approaching the bifurcation): $\ln f'(q^*) \approx -(1 - f'(q^*))$, so $\tau \sim 1/(1 - f'(q^*)) \to \infty$.

Variance and autocorrelation. Under additive noise $\eta_t \sim N(0, \sigma^2)$, the stationary variance of ε_t is:

$$\mathrm{Var}(\varepsilon) = \frac{\sigma^2}{1 - f'(q^*)^2} \approx \frac{\sigma^2}{2(1 - f'(q^*))} \to \infty,$$

and the lag-1 autocorrelation $\to 1$. Both are the hallmark CSD signatures. □

C.5 Proof of Proposition 3

We extend the analysis to asymmetric 2D repair dynamics.

Setup. The two-player system (8) is $q^1_{t+1} = h(q^2_t)$, $q^2_{t+1} = g(q^1_t)$ where $h(q) = \sigma(\beta_1(\alpha_1 q - \kappa_1))$ and $g(q) = \sigma(\beta_2(\alpha_2 q - \kappa_2))$.

Jacobian. The Jacobian of the one-step map at equilibrium (q^{1*}, q^{2*}):

$$J = \begin{pmatrix} 0 & h'(q^{2*}) \\ g'(q^{1*}) & 0 \end{pmatrix}$$

where $h'(q^{2*}) = \beta_1\alpha_1\, q^{1*}(1 - q^{1*})$ (since $q^{1*} = h(q^{2*})$) and $g'(q^{1*}) = \beta_2\alpha_2\, q^{2*}(1 - q^{2*})$.

Characteristic polynomial. $\lambda^2 = h'(q^{2*}) \cdot g'(q^{1*})$. Instability requires $|\lambda| > 1$, i.e., $h'g' > 1$.

Necessary condition. Maximizing: $q(1 - q) \le 1/4$, so $h'g' \le (\beta_1\alpha_1)(\beta_2\alpha_2)/16$. For $h'g' > 1$ to be achievable: $(\beta_1\alpha_1)(\beta_2\alpha_2) > 16$. This is necessary but not sufficient; the actual equilibrium values determine whether the product exceeds 1. □

D Channel Robustness

Proposition 4 (Channel robustness)**.** *Let $\tilde{M}_i = TM_i$ for a stochastic channel T. If $\|T - I\|_{1\to 1} \le \eta$, then for any M, M':*

$$\mathrm{TV}(TM, TM') \le \mathrm{TV}(M, M'),$$
$$\mathrm{TV}(M, M') \le \mathrm{TV}(TM, TM') + \eta.$$

Proof. **Part (i):** The first inequality follows directly from the data-processing inequality for total variation distance: applying a stochastic channel cannot increase divergence.

Part (ii): By the triangle inequality:

$$\mathrm{TV}(M, M') \le \mathrm{TV}(M, TM) + \mathrm{TV}(TM, TM') + \mathrm{TV}(TM', M').$$

For the first error term:

$$\mathrm{TV}(M, TM) = \tfrac{1}{2}\|(I - T)M\|_1 \le \tfrac{1}{2}\|I - T\|_{1\to 1} \underbrace{\|M\|_1}_{=1} \le \eta/2.$$

Similarly, $\mathrm{TV}(TM', M') \le \eta/2$. Combining: $\mathrm{TV}(M, M') \le \mathrm{TV}(TM, TM') + \eta$. □

E Sensitivity: Alternative AMD Couplings

Additive load $\kappa = c_0 + \lambda D$ is a modeling choice. We verify that qualitative tipping persists under alternatives.

AMD erodes coupling. If $\alpha(D) = \alpha_0 - \rho D$ with fixed β, the bistability threshold $\beta\alpha(D) = 4$ gives $D^* = (\alpha_0 - 4/\beta)/\rho$, provided $\alpha_0 > 4/\beta$.

AMD increases noise. If $\beta(D) = \beta_0/(1 + \xi D)$ with fixed α, the threshold $\beta(D)\alpha = 4$ gives $D^* = (\beta_0\alpha/4 - 1)/\xi$, provided $\beta_0\alpha > 4$.

In both cases, qualitative tipping persists whenever an effective gain crosses 4.

F KL Extension of Decomposition

For KL divergence, an analogous upper bound holds:

$$\mathrm{KL}(\overline{M}_1 \| \overline{M}_2) \leq \mathrm{KL}(P_1 \| P_2) + \mathbb{E}_{c\sim P_1}\big[\mathrm{KL}(M_{1,c} \| M_{2,c})\big].$$

This provides a KL analogue of Proposition 1. The bound is not tight in general: when $M_{1,c}$ varies across contexts, the marginal mixture $\overline{M}_1$ can have lower KL divergence from $\overline{M}_2$ than the average of the conditional KL terms, due to the concavity of marginal mixing.

G Repair Proxy Validation Details

Dialog-act classifier. RoBERTa-base fine-tuned on Switchboard Dialog Act Corpus (46 acts), 3 epochs, lr 2×10^{-5}. Test accuracy: 73.1%. Repair-adjacent acts aggregated: `aa` (agree/accept), `bk` (backchannel), `br` (signal-non-understanding), `ba` (appreciation).

Inter-proxy correlations. $r(\hat{q}^{\mathrm{DA}}, \hat{q}^{\mathrm{RM}}) = -0.100$ ($p < 0.001$); $r(\hat{q}^{\mathrm{DA}}, \hat{q}^{\mathrm{CE}}) = 0.031$ ($p = 0.010$). Repair marker coverage: 84.4% of conversations contain at least one repair marker. The weak but significant inter-proxy correlations indicate that the three proxies measure related but distinct aspects of conversational repair.

H Experiment Details

H.1 CGA-Wiki preprocessing

ConvoKit v2.5 release (Wikipedia version). Conversations with fewer than 10 turns are excluded, yielding 652 conversations (389 derailing, 263 civil). The repair proxy classifier is RoBERTa-base fine-tuned on the Switchboard Dialog Act Corpus (46 tags) for 3 epochs with learning rate 2×10^{-5} (test accuracy 73.1%). We aggregate dialog-act probabilities for conciliatory acts (agree/accept, backchannel, signal non-understanding, appreciation) into a single repair score $\hat{q}_t$.

H.2 CGA-CMV preprocessing

The CGA-CMV corpus (Chang and Danescu-Niculescu-Mizil, 2019) applies the Conversations Gone Awry labeling framework to Reddit ChangeMyView threads. Of 4,389 total conversations, 1,169 have $\geq$ 10 turns and are usable for CSD analysis; the remaining 73% are too short. The breakdown label (`has_removed_comment`) is noisier than CGA-Wiki's annotated personal attack: Cohen's $d = 0.34$ (CGA-CMV) vs. 1.88 (CGA-Wiki). The corpus-wide median thread length is 7 turns.

H.3 Asymmetric 2D simulation

With $\alpha_1 = \alpha_2 = 2$, $\beta_1 = \beta_2 = 3$ (product $= 36 > 16$) and $\kappa_1 = \kappa_2 = \kappa$, trajectories under symmetry reduce to the 1D case. For $\kappa_1 = 1.0$, $\kappa_2 = 0.8$ (moderate asymmetry), we observe multi-attractor behavior with the basin boundary shifted toward the lower-load speaker. Full 2D bifurcation classification is left to future work.

H.4 Dose-response analysis on CGA-CMV

We test whether the CSD effect size increases with conversation length on CGA-CMV by stratifying conversations into length bins, as predicted by the theory's requirement that CSD detection needs sufficient approach time to the bifurcation.

No individual stratum reaches significance, consistent with the noisy breakdown label (Cohen's $d = 0.34$); the directional trend nonetheless supports the theory's prediction that longer conversations allow more time for CSD signatures to manifest before breakdown.

H.5 Window-size sensitivity

We re-run the Exp. 4 CSD analysis on CGA-Wiki with rolling-window sizes $W \in \{3, 4, 5, 6, 7\}$, keeping all other parameters identical (Kendall τ on variance trends, permutation test with 10,000 shuffles). Conversations are excluded if they either (i) have fewer than $W + 5$ turns or (ii) yield fewer than 3 non-NaN variance values in the pre-breakdown window for Kendall τ computation. The latter filter accounts for conversations where the attack occurs early enough that the adjusted variance window has too few valid points.

The results confirm that CSD signals are robust across $W = 3$–6: at least one theory-derived indicator is significant in every case. AMD Variance achieves the strongest separation at the reported $W = 5$ (and $W = 6$), while $\hat{q}^{\mathrm{DA}}$ Variance is more robust at shorter windows ($W = 3$–4), likely because the dialog-act proxy is a per-utterance measure that requires less smoothing to differentiate trends. The complementary pattern supports the interpretation that both indicators reflect genuine CSD dynamics, with AMD requiring a slightly longer observation window to accumulate a detectable signal in the emotion-distribution space.

Stratum	N	$\Delta\tau(q_{\mathrm{DA}})$	p	$\Delta\tau$(AMD)	p
5-6	1612	*not analyzed (<10 turns)*			
7	649	$+0.092$	0.109	-0.021	0.702
8-9	959	$+0.011$	0.782	$+0.010$	0.800
10-12	765	-0.045	0.267	-0.006	0.883
13+	404	-0.032	0.529	$+0.010$	0.846

Table 6: Length-stratified dose-response analysis on CGA-CMV. $\Delta\tau = \tau_{\text{derail}} - \tau_{\text{civil}}$ for variance trends. No individual stratum reaches significance for either indicator, consistent with the noisy `has_removed_comment` label (Cohen's $d = 0.34$). The q_{DA} VAR $\Delta\tau$ transitions from positive (short conversations) to negative (long), directionally consistent with the theoretical prediction that CSD signals strengthen with conversation length. Strata with < 10 turns are excluded from the CSD computation; strata at 7 and 8-9 turns are marginal but yield at least 3 variance points with W=5.

W	AMD Var p	AMD Var d	q_{DA} VAR p	q_{DA} VAR d	N
3	0.484	-0.057	**0.039**	-0.166	650
4	0.054	-0.157	**0.005**	-0.219	650
5	**0.001**	-0.257	0.016	-0.199	652
6	**0.003**	-0.342	0.039	-0.244	300
7	0.734	-0.096	0.554	-0.167	65

Table 7: Window-size sensitivity analysis on CGA-Wiki. Bold: smallest p among CSD indicators for that W. At $W = 3$–4, $\hat{q}^{\mathrm{DA}}$ Variance dominates; at $W = 5$–6, AMD Variance is strongest. At $W = 7$, the sample size ($N = 65$) is too small for either indicator to reach significance. The two CSD indicators are complementary across window sizes. The small variation in N across $W = 3$–5 (650, 650, 652) reflects 2 conversations where the attack occurs early enough that the pre-breakdown variance window has < 3 valid points at smaller W.