The Autocorrelation Blind Spot: Why 42% of Turn-Level Findings in LLM Conversation Analysis May Be Spurious

Anonymous authors
Paper under double-blind review

Abstract

Turn-level metrics are widely used to evaluate properties of multi-turn human-LLM conversations, from safety and sycophancy to dialogue quality. However, consecutive turns within a conversation are not statistically independent—a fact that virtually all current evaluation pipelines fail to correct for in their statistical inference. We systematically characterize the autocorrelation structure of 66 turn-level metrics across 202 multi-turn conversations (11,639 turn pairs, 5 German-speaking users, 4 LLM platforms) and demonstrate that naive pooled analysis produces severely inflated significance estimates: **42%** of associations that appear significant under standard pooled testing fail to survive cluster-robust correction. The inflation varies substantially across categories rather than scaling linearly with autocorrelation: three memoryless families (embedding velocity, directional, differential) aggregate to 14%, while the seven non-memoryless families (thermo-cycle, frame distance, lexical/structural, rolling windows, cumulative, interaction, timestamp) aggregate to 33%, with individual category rates ranging from 0% to 100% depending on per-family effect size. We present a two-stage correction framework combining Chelton (1983) effective degrees of freedom with conversation-level block bootstrap, and validate it on a pre-registered hold-out split where cluster-robust metrics replicate at 57% versus 30% for pooled-only metrics. We provide concrete design principles, a publication checklist, and open-source code for the correction pipeline. A survey of ∼30 recent papers at major NLP and AI venues that compute turn-level statistics in LLM evaluations finds that only 4 address temporal dependence at all, and 26 do not correct for it.

1 Introduction

Every turn in a conversation remembers the turn before it. This fact—obvious to any conversational participant, and well established in discourse analysis, autoregressive language modelling, and dialogue state tracking—is nevertheless not corrected for in the statistical inference of virtually every quantitative evaluation of multi-turn LLM interactions published in the last three years. Turn-level metrics for safety classification (Inan et al., 2023; Han et al., 2024), dialogue quality assessment (Mehri & Eskenazi, 2020; Zhang et al., 2021), persona consistency tracking, and sycophancy measurement all share a common assumption: that individual turns constitute independent observations. They do not. Lag-1 autocorrelation $\rho_1 = \mathrm{cor}(x_t, x_{t-1})$ quantifies this dependence: $\rho_1 = 0$ indicates independent observations, $\rho_1 \to 1$ indicates a random walk. In our corpus of 202 human-LLM conversations spanning 11,639 turn pairs, lag-1 autocorrelation ranges from $\rho = 0.057$ for embedding velocity metrics to $\rho = 0.928$ for cumulative accumulators. A metric with $\rho = 0.928$ measured over 11,639 turns has an effective sample size of approximately 435—a 27-fold reduction. The pooled Pearson correlation that appears to demonstrate a significant finding at $p < 10^{-15}$ may, after proper correction, show $p = 0.34$.

The scale of the problem. We present the first systematic empirical characterization of autocorrelation in turn-level conversational metrics, and the results are sobering. In our study—a pre-registered analysis of manipulation detection in LLM conversations using embedding-geometry metrics—81 metric-label associations pass pooled Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995). Of these, only

47 survive cluster-robust inference combining Chelton effective degrees of freedom (Chelton, 1983) with block bootstrap (Politis & Romano, 1994) (27 for the MALICIOUS label, 10 for the TYPX label, 10 for the REWIRING label). Critically, this inflation is not uniformly distributed across metric types. Inflation rates vary by category rather than scaling linearly with $\bar{\rho}$: three memoryless families (embedding velocity, directional, differential) aggregate to 14%, the seven non-memoryless families (thermo-cycle, frame distance, lexical/structural, rolling windows, cumulative, interaction, timestamp) aggregate to 33%, and individual category rates range from 0% to 100%. The autocorrelation structure of a metric, not its raw effect size, determines whether a finding is real. A timestamp metric showing $r = 0.018$ (pooled-significant at $p = 0.04$, $\rho = 0.70$) does not survive correction, while an embedding-velocity metric showing $r = 0.077$ ($\rho = 0.06$) does—because the latter has 13× more effective observations ($n_{\text{eff}} = 10{,}394$ vs. 779).

Contributions. We make five contributions:

- **C1 Empirical characterization.** We measure lag-1 autocorrelation for 66 turn-level conversational metrics across 202 conversations and document a spectrum from $\rho = 0.057$ to $\rho = 0.928$ (Section 3).
- **C2 Two-stage correction framework.** We adapt Chelton's effective degrees of freedom (Chelton, 1983) and conversation-level block bootstrap (Politis & Romano, 1994) for the specific dependence structure of multi-turn dialogue data (Section 4).
- **C3 Taxonomy of metric robustness.** We show that a metric's autocorrelation profile predicts its robustness: differential metrics (rates of change) survive correction; cumulative metrics (running sums) do not (Section 5).
- **C4 Concrete recommendations.** We provide a checklist for researchers computing turn-level statistics in conversational data (Section 6).
- **C5 Literature survey.** We surveyed approximately 30 recent papers at major NLP and AI venues (2020–2025) that compute turn-level metrics in multi-turn LLM evaluations. Only 4 address temporal dependence (2 fully, 2 partially); 26 do not correct for it (Section 2).

Scope and running example. Our running example is manipulation detection in human-LLM conversations, where we correlate embedding-geometry metrics with turn-level annotations of manipulative behavior. However, the methodological problem we identify is not an empirical finding about our corpus—it is a deductive consequence of conversational architecture. In any multi-turn conversation, the LLM's response at turn t is conditioned on the full prefix $(x_1, \ldots, x_{t-1})$; the user's subsequent message is shaped by this response; and the LLM's next output depends on both. This causal chain guarantees that *most* metrics computed on turn content inherit serial dependence, regardless of language, platform, or user population; only metrics designed to be memoryless (e.g., per-turn embedding velocity) approach $\rho \approx 0$. The specific inflation rate (42% in our data) is corpus-dependent, but the direction of the bias—inflation, never deflation—and the mechanism are structural invariants of sequential conversation.

2 Background and Related Work

2.1 Turn-Level Evaluation in Conversational AI

The dominant paradigm for evaluating multi-turn LLM interactions computes per-turn scores and aggregates them. Safety classifiers such as LlamaGuard (Inan et al., 2023) and WildGuard (Han et al., 2024) assign risk scores to individual messages. Dialogue quality metrics like USR (Mehri & Eskenazi, 2020) and DynaEval (Zhang et al., 2021) evaluate response appropriateness at the turn level. Engagement and coherence trackers measure properties of successive responses. In all cases, statistical tests—correlations, t-tests, regression coefficients—treat turns as rows in a flat table.

This approach inherits a structural problem: consecutive turns within a conversation are not independent. The LLM's response at turn t is conditioned on the full history up to turn t; the user's subsequent message at turn $t{+}1$ is shaped by what the LLM just said; and the LLM's next response is conditioned on both. Any metric computed on these turns inherits this serial dependence. The result is *pseudoreplication*: the effective sample size is smaller—sometimes dramatically smaller—than the number of data points. Yet none of the evaluation frameworks listed above account for this.

2.2 Autocorrelation in Other Fields

The statistical consequences of temporal autocorrelation have been well characterized in fields that routinely work with time series data.

Climate science. Chelton (1983) demonstrated that standard significance tests on ocean temperature time series are severely biased when autocorrelation is ignored. He introduced the effective degrees of freedom formula $n_{\text{eff}} = n \cdot (1 - \bar{\rho})/(1 + \bar{\rho})$, where $\bar{\rho}$ is the mean lag-1 autocorrelation, which we adapt in this work. The formula has since become standard practice in physical oceanography and climate science (Mudelsee, 2010).

Neuroscience. Eklund et al. (2016) showed that spatial autocorrelation in fMRI data inflates cluster-level false positive rates to approximately 70%—far above the nominal 5%. Their finding, published in *PNAS*, led to widespread reanalysis of neuroimaging results and new standards for spatial correction. The parallel to our finding is direct: they documented a 70% inflation from spatial autocorrelation in brain imaging; we find comparable inflation from temporal autocorrelation in conversations.

Psychotherapy process research. Tasca & Gallop (2009) argued that session-level observations in psychotherapy are nested within patients and serially dependent within treatment, requiring multilevel models with autoregressive error structures. Their recommendation—that ignoring this nesting inflates Type I error—applies with equal force to turn-level observations nested within conversations.

Econometrics. Newey & West (1987) developed heteroskedasticity and autocorrelation consistent (HAC) estimators for regression with time series data. The Newey-West standard errors are now routine in empirical economics whenever residuals show serial correlation. Cluster-robust standard errors (Cameron & Miller, 2015) extend this logic to panel data where observations are grouped (analogous to turns grouped within conversations).

In NLP: known in principle, uncorrected in practice. NLP research is well aware that consecutive turns are sequentially dependent: this is the foundation of autoregressive language modelling, discourse analysis, and dialogue state tracking, and remains an active area of study in multi-turn evaluation (Luz de Araujo et al., 2026). The gap we identify is narrower and strictly statistical: when pooled turn-level tests are reported, cluster-robust inference is almost never applied. We surveyed approximately 30 papers published at major NLP and AI venues between 2020 and 2025 that report turn-level statistics in multi-turn LLM evaluations—including safety benchmarks, dialogue quality studies, and persona consistency analyses.[1] Of these, 26 do not report lag-1 autocorrelation of their metrics, apply temporal correction to their significance tests, or discuss effective sample size reduction. The remaining 4 (two fully, two partially) model temporal dependence architecturally (e.g., Hawkes processes, graph-based recurrence) but do not compute cluster-robust significance for pooled turn-level tests. Several aggregate thousands of turns from dozens of conversations into a single pooled test—precisely the setting where autocorrelation bias is most severe. This is not a criticism of individual papers; it reflects the absence of an established norm. Our goal is to establish that norm.

Why this gap persisted in NLP. The statistical correction tools described above (Chelton's effective degrees of freedom, cluster-robust standard errors, block bootstrap) are standard in climatology, econometrics, and neuroscience but have not diffused into NLP evaluation—reflecting disciplinary boundaries rather than conceptual inaccessibility. Our contribution is the transfer of these established methods to the specific dependence structure of multi-turn conversational data.

2.3 Cluster-Robust Inference

The statistical tools for handling autocorrelation in grouped data exist and are well understood, though they have not been applied to conversational AI evaluation. Block bootstrap methods (Politis & Romano, 1994)

[1] The full list of surveyed papers is provided in Appendix E.

Metric Family	$\bar{\rho}_1$	n_{eff}	Red.
Cumulative accum.	0.928	435	27×
Rolling aggr. (W=20)	0.910	549	21×
Normalized cumul.	0.710	1,974	6×
Lexical/structural	0.500	3,880	3×
Frame distance	0.450	4,415	3×
Impulse (1st diff.)	0.438	4,540	3×
Thermodynamic	0.350	5,603	2×
Directional consist.	0.206	7,663	1.5×
Compression (NCD)	0.105	9,426	1.2×
Embedding velocity	0.057	10,394	1.1×

Table 1: Lag-1 autocorrelation $\bar{\rho}_1$ (averaged across 202 conversations), effective sample size n_{eff} (of n=11,639 turn pairs), and reduction factor for ten representative metric families. Figure 1 visualizes the same families graphically. Cumulative and rolling metrics lose more than 95% of nominal degrees of freedom (96.3% and 95.3%, respectively).

resample entire blocks (in our case, entire conversations) rather than individual observations, preserving the within-block dependence structure. Cluster-robust standard errors (Cameron & Miller, 2015) adjust variance estimates by treating each cluster (conversation) as a single effective observation. Random-effects models with by-item random slopes (Barr et al., 2013) offer a complementary approach for crossed random effects.

Our contribution is not the invention of new statistical methods. It is the *adaptation* of these methods for the specific dependence structure of conversational data—where turns are nested within conversations, conversations vary in length by an order of magnitude, and the autocorrelation structure differs radically across metric types—and the *empirical demonstration* that failing to apply them substantially inflates false positive rates in a real-world evaluation setting.

3 The Autocorrelation Problem in Conversational Data

Turn-level metrics extracted from multi-turn conversations are not independent observations. Each turn is a response to the preceding one, and many commonly used metrics—cumulative scores, rolling averages, trajectory summaries—carry forward information from earlier turns by construction. This serial dependence inflates the effective sample size and, consequently, the statistical significance of any correlation computed over pooled turn-level data. In this section, we characterize the severity of this problem empirically across our corpus and demonstrate its consequences for hypothesis testing.

3.1 Empirical Characterization

We compute the lag-1 autocorrelation ρ_1 for each turn-level metric within each of the 202 conversations, then average across conversations to obtain $\bar{\rho}_1$. From $\bar{\rho}_1$, we derive the effective sample size n_{eff} using the Chelton correction (Section 4.1) applied to the pooled n = 11,639 turn pairs, and express the reduction factor as n/n_{eff}.

Table 1 presents these values for ten representative metric families, ordered by decreasing autocorrelation. The range is striking: from $\bar{\rho}_1 = 0.928$ for cumulative accumulators (a 27× reduction in effective observations) to $\bar{\rho}_1 = 0.057$ for embedding velocity (only 1.1× reduction).

Several patterns are notable. First, the ordering is not arbitrary: it reflects the *temporal memory* inherent in each metric's definition. Cumulative accumulators (e.g., running budget scores, cumulative sentiment) carry all previous information forward; their autocorrelation approaches that of a random walk ($\rho \to 1$ for large n). Rolling aggregates with window W=20 behave similarly because consecutive windows overlap by $W-1$ observations. Impulse metrics—first differences of accumulators—remove much of this dependence by design, reducing ρ to approximately 0.5. At the other extreme, embedding velocity ($\|\Delta \mathbf{e}_t\| = \|\mathbf{e}_t - \mathbf{e}_{t-1}\|$)

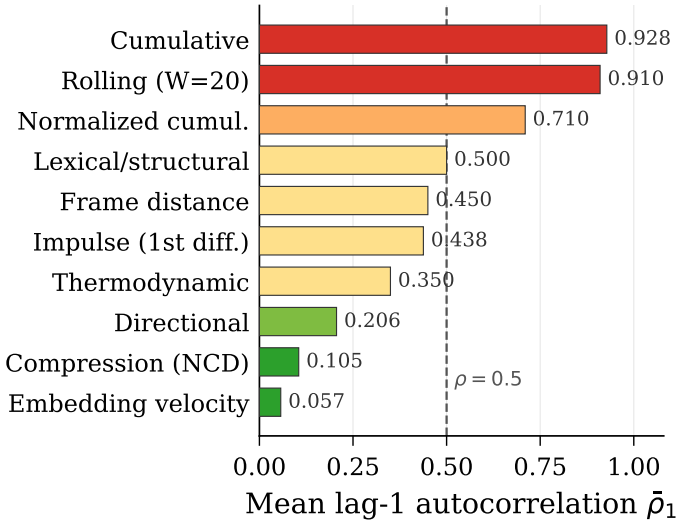

Figure 1: Mean lag-1 autocorrelation across metric families. Metrics above the dashed line ($\rho \approx 0.5$) are at severe risk of significance inflation.

measures a purely local quantity with minimal carryover, yielding $\bar{\rho}_1 = 0.057$ and retaining 89% of the nominal sample size.

Second, the variation *within* a family can be substantial. Frame distance metrics range from $\rho = 0.35$ (token-weighted variant) to $\rho = 0.55$ (raw cosine to fixed centroid), depending on whether the reference frame is updated or fixed. We report ranges where appropriate.

3.2 Consequences for Hypothesis Testing

To make the consequences concrete, we present a worked example. Consider a metric with $\bar{\rho}_1 = 0.8$, measured over a conversation of $n = 200$ turn pairs. A pooled point-biserial correlation with a binary label yields $r = 0.15$, and the standard test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{1}$$

gives $t = 2.14$ with df $= 198$, corresponding to $p = 0.034$. By conventional standards, this is significant at $\alpha = 0.05$.

However, applying the Chelton correction (Eq. 4 below) yields

$$n_{\text{eff}} = 200 \times \frac{1-0.8}{1+0.8} = 200 \times \frac{0.2}{1.8} \approx 22.2, \tag{2}$$

and the corrected test statistic becomes

$$t_{\text{corr}} = r\sqrt{\frac{n_{\text{eff}}-2}{1-r^2}} = 0.15 \times \sqrt{\frac{20.2}{0.9775}} = 0.68, \tag{3}$$

with df $= 20.2$ and $p_{\text{corr}} = 0.51$.

The result is emphatically not significant. The *effect size* is unchanged ($r = 0.15$), but the *evidence* for that effect is vastly weaker than the pooled analysis suggests. Nine out of ten nominal degrees of freedom were illusory, created by the serial dependence between consecutive turns rather than by genuinely independent observations.

This is not a pathological edge case. In our corpus, 42% of all correlations that pass pooled FDR screening (Benjamini-Hochberg at $q = 0.05$) fail to survive cluster-robust correction—a figure we term the *inflation rate* and discuss in Section 4.4.

3.3 Why Conversations Are Especially Vulnerable

Serial dependence is a known issue in time-series analysis (Chelton, 1983; von Neumann, 1941), and corrections for it are standard in climatology, econometrics, and neuroscience. Yet conversation analysis—and LLM safety evaluation in particular—has largely left this dependence statistically uncorrected in practice. Four features of conversational data make it especially severe:

Inherent sequential dependence. Unlike time-series data where autocorrelation may be incidental (e.g., measurement noise), conversations are *causally* sequential: each turn is a direct response to the previous one. The embedding of turn t is constrained by the embedding of turn $t-1$ through the coherence requirements of natural dialogue. This causal coupling is a *structural* source of autocorrelation that cannot be eliminated by metric design—only managed.

Variable conversation length. Conversations in our corpus range from 11 to 507 turn pairs (Figure 1). When observations are pooled across conversations, long conversations dominate the analysis despite contributing relatively few independent observations. A conversation with 500 turn pairs and $\rho = 0.94$ contributes 500 nominal observations but only $\sim$15 effective ones. Pooling treats all 500 equally, creating a phantom sample that inflates significance.

Cumulative metrics are natural but dangerous. Many intuitively appealing conversational metrics are cumulative by design: running sentiment scores, topic drift trajectories, engagement curves, budget accumulators. These metrics are scientifically meaningful—they capture real dynamics—but their cumulative nature guarantees high autocorrelation ($\rho > 0.9$). Researchers using such metrics without correction will systematically overestimate the significance of their findings.

Catastrophic loss of power in short conversations. For a conversation of 30 turn pairs with $\rho = 0.8$, the effective sample size is $n_{\text{eff}} = 30 \times 0.2/1.8 \approx 3.3$. With three effective observations, no turn-level correlation can reach significance. Short conversations do not merely lose statistical power—they lose it *catastrophically*, leaving only the conversation-level effect (one data point per conversation) as usable evidence. This creates a systematic bias: only findings driven by long conversations survive, while effects that manifest in short conversations are invisible.

4 Correction Methodology

We propose a two-stage protocol that first screens candidate correlations efficiently and then subjects survivors to rigorous correction for autocorrelation. The protocol combines two complementary approaches—an analytic correction based on effective degrees of freedom and a nonparametric correction based on cluster resampling—and uses their agreement as the criterion for robustness.

4.1 Chelton Effective Sample Size

Chelton (1983) showed that for a time series with lag-1 autocorrelation ρ, the number of effectively independent observations is

$$n_{\text{eff}} = n \cdot \frac{1 - \bar{\rho}}{1 + \bar{\rho}}, \tag{4}$$

where $\bar{\rho}$ is the mean lag-1 autocorrelation estimated across all conversations for the metric in question. This formula assumes a stationary AR(1) process—an approximation for conversational data, but one that captures the dominant source of dependence.

In practice, we estimate $\bar{\rho}$ by computing the lag-1 autocorrelation of the metric time series within each conversation, then averaging across all k conversations:

$$\bar{\rho} = \frac{1}{k} \sum_{j=1}^{k} \hat{\rho}_1^{(j)}, \tag{5}$$

Algorithm 1 Cluster-Robust Block Bootstrap

Require: Conversations $C_1, \ldots, C_k$ with turns $\{(\ell_i^{(j)}, m_i^{(j)})\}_{i=1}^{n_j}$
Require: Number of bootstrap iterations $B = 2{,}000$
Ensure: Two-sided p-value p_{boot}
1: **for** $b = 1$ **to** B **do**
2: Sample k conversations with replacement: $\{C_{s_1}^*, \ldots, C_{s_k}^*\}$
3: Pool all turns: $\mathcal{D}_b^* \leftarrow \bigcup_{j=1}^{k} \{(\ell_i^{(s_j)}, m_i^{(s_j)})\}$
4: Compute $r_b \leftarrow \text{pointbiserialr}(\boldsymbol{\ell}_b^*, \boldsymbol{m}_b^*)$
5: **end for**
6: $p_{\text{boot}} \leftarrow 2 \cdot \min\big(\frac{1}{B}\sum_b \mathbb{1}[r_b \le 0],\ \frac{1}{B}\sum_b \mathbb{1}[r_b \ge 0]\big)$
7: **return** p_{boot}

where $\hat{\rho}_1^{(j)}$ is the sample lag-1 autocorrelation for conversation j.

We impose a lower bound:

$$n_{\text{eff}} \ge k, \tag{6}$$

ensuring that the effective sample size is at least as large as the number of conversations. This bound reflects the minimal assumption that each conversation contributes at least one independent observation (its mean effect), regardless of within-conversation dependence.

The corrected p-value is obtained by evaluating the standard correlation test statistic (Eq. 1) with $\text{df} = n_{\text{eff}} - 2$ degrees of freedom using the t-distribution:

$$p_{\text{Chelton}} = 2 \cdot P(T > |t_{\text{corr}}|)\,, \quad T \sim t(n_{\text{eff}} - 2). \tag{7}$$

Limitations. The Chelton correction assumes stationarity and first-order Markov dependence. Conversational metrics may exhibit higher-order autocorrelation (e.g., cumulative metrics approximate integrated random walks with $\rho_k \gg 0$ for $k > 1$). For such cases, the correction is conservative but may still be insufficient, motivating the complementary nonparametric approach below.

4.2 Cluster-Robust Block Bootstrap

The block bootstrap (Efron, 1979; Davison & Hinkley, 1997) provides a nonparametric alternative that makes no distributional assumptions. The key insight is that resampling *entire conversations* as atomic units preserves the within-conversation autocorrelation structure, avoiding the need to model it explicitly.

Algorithm 1 specifies the procedure. We resample k conversations with replacement from the pool of k conversations, concatenate all turns from the sampled conversations into a pooled dataset, and compute the point-biserial correlation r_b on this bootstrap sample. After $B = 2{,}000$ iterations, the two-sided p-value is

$$p_{\text{boot}} = 2 \cdot \min\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}[r_b \le 0],\ \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}[r_b \ge 0]\right). \tag{8}$$

Here, $\ell_i^{(j)} \in \{0, 1\}$ denotes the binary label (e.g., manipulative vs. non-manipulative) and $m_i^{(j)} \in \mathbb{R}$ denotes the metric value for turn i in conversation j. The point-biserial correlation r is algebraically equivalent to the Pearson correlation between a binary and a continuous variable.

Why conversations, not turns? Resampling individual turns would destroy the sequential structure and produce samples with near-zero autocorrelation—precisely the artifact we seek to avoid. By resampling whole conversations, we ensure that each bootstrap sample inherits the autocorrelation present in the original data. If a metric's significance is driven by within-conversation patterns that happen to align with the label (rather than by genuinely independent evidence across conversations), the bootstrap will produce a wide distribution of r_b values that spans zero, yielding a high p_{boot}.

Computational cost. With $B = 2{,}000$ iterations, $k = 202$ conversations, and $n = 11{,}639$ turns, each bootstrap sample involves resampling 202 indices and concatenating the corresponding turn arrays. The total computation takes approximately 2–5 seconds per metric on a single core, making the procedure tractable even for large metric inventories.

4.3 Two-Stage Protocol

We combine the above into a two-stage protocol, illustrated schematically below:

Stage 1: Screening. Compute pooled point-biserial correlations for all metric–label pairs. Apply Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) at $q = 0.05$. Metrics passing this screen are *pooled-significant* candidates. This stage is computationally cheap and controls the false discovery rate under the (incorrect) assumption of independence.

Stage 2: Cluster-robust confirmation. For each pooled-significant metric, compute both p_{Chelton} (Eq. 7) and p_{boot} (Eq. 8). A metric is declared *cluster-robust* if and only if

$$\max(p_{\text{Chelton}},\, p_{\text{boot}}) < \alpha, \tag{9}$$

where $\alpha = 0.05$. This conservative criterion requires agreement between the analytic and nonparametric corrections: both must independently confirm significance.

The two-stage design balances efficiency and rigor. Stage 1 eliminates the vast majority of candidates (typically 60–80% of all tested pairs), allowing Stage 2 to focus its computationally more expensive bootstrap on a manageable subset. Stage 2's dual-correction criterion guards against both the specific failure modes of each method: the Chelton correction may be too lenient for non-stationary metrics, while the bootstrap may be too conservative for metrics with near-zero autocorrelation (where the Chelton correction is exact).

4.4 Inflation Rate as Diagnostic

We define the *inflation rate* (IR) as the proportion of pooled-significant findings that fail cluster-robust confirmation:

$$\text{IR} = \frac{n_{\text{pooled-sig}} - n_{\text{robust}}}{n_{\text{pooled-sig}}}. \tag{10}$$

This quantity serves as a transparency metric: it estimates the fraction of turn-level findings that are likely spurious due to unaccounted autocorrelation. An inflation rate of 0% would indicate that pooled analysis is reliable; an inflation rate approaching 100% would indicate that nearly all pooled findings are artifacts.

Corpus-level result. In our corpus ($n = 202$ conversations, 11,639 turn pairs, 81 pooled-significant metric–label pairs across all three labels), the inflation rate is

$$\text{IR} = \frac{81 - 47}{81} \approx 0.42, \tag{11}$$

or approximately **42%**. That is, approximately two-fifths of all correlations that appear significant under standard pooled analysis do not survive correction for within-conversation dependence.

Decomposition by autocorrelation. The inflation rate varies with $\bar{\rho}_1$, but per-family sample sizes are small (Table 3). Figure 2 visualizes the relationship by metric category, with point size proportional to the number of tests per category; the logistic fit shows a gradual decline in survival across the ρ range. Moderate-autocorrelation families (thermo-cycle, frame distance, lexical/structural) show inflation rates of 33–50%, while very-low ($\bar{\rho}_1 < 0.2$) and very-high ($\bar{\rho}_1 > 0.9$) categories behave idiosyncratically due to small per-family test counts—compression-based metrics (2 tests) lose both, while rolling-window metrics (3 tests) retain all three.

This decomposition yields a practical heuristic: *metrics with $\bar{\rho}_1 > 0.5$ should be treated as suspect until confirmed by cluster-robust methods.* We emphasize that this is a heuristic, not a hard cutoff—a metric with

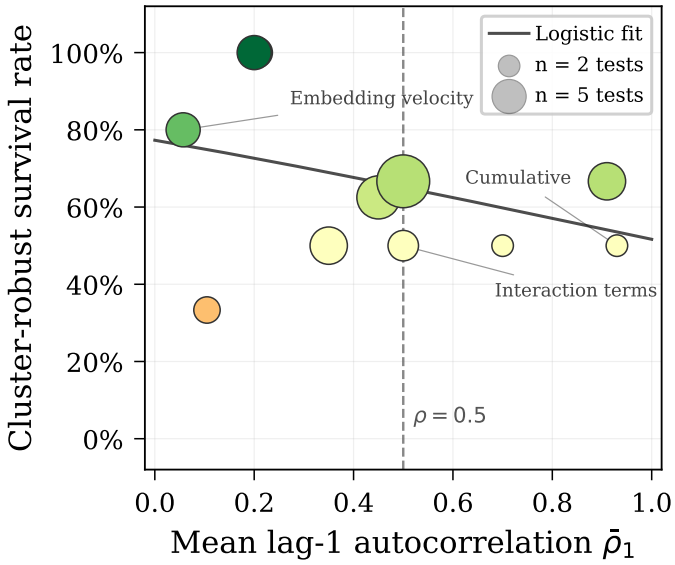

Figure 2: Cluster-robust survival as a function of lag-1 autocorrelation. Each point represents one metric category that passed Stage 1 (pooled FDR). Point size is proportional to the number of tests. The logistic fit shows a gradual decline in survival with increasing ρ; the $\rho \approx 0.5$ threshold marks a practical decision boundary for our two-stage protocol rather than a sharp cliff.

$\rho = 0.6$ may be genuinely significant if the effect is large, while a metric with $\rho = 0.4$ may be inflated if the effect is marginal. The inflation rate provides the diagnostic; the two-stage protocol provides the remedy.

Reporting recommendation. We argue that any study reporting turn-level correlations from multi-turn conversation data should include the inflation rate alongside its findings. This single number captures the degree to which the reported significance may be overstated and allows readers to calibrate their confidence accordingly. Studies with IR $\approx 0\%$ require no further scrutiny; studies with IR $> 30\%$ should prompt careful examination of which specific findings survive and which do not.

5 Empirical Demonstration

We now apply the two-stage correction protocol from Section 4 to a multi-turn manipulation detection task. The goal is not to validate any particular metric—the metrics serve as a running example—but to quantify *how much* naive statistical inference overestimates the evidence base.

5.1 Experimental Setup

Corpus. We analyze 202 multi-turn conversations between five German-speaking users and four LLM platforms (ChatGPT (GPT-4/GPT-4o), Deepseek, Gemini, and Claude), totalling 11,639 aligned turn pairs. Conversations range from 11 to 507 turn pairs (median ≈36). They cover everyday topics (cooking, travel, study assistance) as well as adversarial interactions that emerged naturally during extended use.

Task and Labels. Each turn pair carries a binary MALICIOUS label (MAL $\in \{0, 1\}$) indicating whether the assistant's response exhibits manipulative characteristics, as determined by a structured annotation pipeline.[2] Three manipulation categories are distinguished: **(A)** semantic rewiring (concepts are covertly redefined), **(B)** compliance drift (boundaries dissolve incrementally), and **(C)** elicitation (information is extracted through skillful framing without overt boundary violation).

[2]Annotation uses a structured LLM-as-judge protocol (Claude Opus 4.6) with inter-pass consistency $\kappa = 0.82$, reviewed by a computational linguist; see Appendix C. The correction methodology is agnostic to the annotation source (Appendix C).

Community metric	Design class	$\rho_{\text{predicted}}$
USR (Mehri & Eskenazi, 2020)	Rolling window	0.7–0.9
LlamaGuard (Inan et al., 2023)	Per-turn classifier	0.3–0.6
DynaEval (Zhang et al., 2021)	Graph-based window	0.4–0.7
Cumulative sentiment	Cumulative	>0.9
Per-turn perplexity	Per-turn scalar	0.2–0.4

Table 2: Predicted autocorrelation ranges for community metrics, derived from design-class membership. Empirical verification on standard benchmarks is a natural next step.

Pre-registration. Before annotating the hold-out split, we filed an internal pre-registration document (available on request) specifying an exploratory set ($n_{\text{expl}} = 152$), an independent hold-out set ($n_{\text{ho}} = 50$), along with six directional hypotheses and a meta-level permutation criterion.

Metrics. We compute 66 turn-level metrics spanning twelve conceptual families (Table 3). Families include embedding velocities, directional consistency measures, compression-based distances, differential (Δ) metrics (first differences of per-turn quantities, distinct from the impulse family of Table 1 which differences cumulative accumulators), thermodynamic-cycle integrals, frame distances, lexical/structural features, rolling-window aggregates, cumulative accumulators, interaction terms, and timestamp-based features. *The specific metric implementations are not the contribution of this paper*; they serve to illustrate the autocorrelation problem across a broad spectrum of design choices.

Generalizability beyond our metric suite. A potential concern is that the autocorrelation patterns we observe are specific to our metric implementations. We address this by noting that our metrics span the full spectrum of temporal dependencies—from memoryless per-turn quantities ($\rho \approx 0.06$) to cumulative accumulators ($\rho \approx 0.93$)—and are organized into design classes that correspond to generic architectural choices, not to specific implementations (Table 2).

The key insight is that autocorrelation is a property of the *design class*, not the *implementation*. Researchers using standard community metrics should expect at least the inflation rates we document for the corresponding class in Table 3.

5.2 Results

Table 3 presents the central finding. All 81 metric–label pairs that pass pooled BH-FDR screening are grouped by conceptual category and evaluated for cluster-robust survival.

Several patterns emerge. First, the inflation rate **varies with mean autocorrelation** ($\bar{\rho}$), though per-family sample sizes are small (2–12 tests each). The three memoryless families—embedding velocity ($\bar{\rho} = 0.06$), directional (0.21), and differential (0.20), distinct from the impulse family of Table 1 ($\bar{\rho} = 0.438$) which differences cumulative accumulators rather than per-turn quantities—retain 6 of 7 findings (aggregate IR = 14%). The seven non-memoryless families—thermodynamic (0.35), frame distance (0.45), lexical/structural (0.50), rolling windows (0.91), cumulative (0.93), interaction (0.50), and timestamp (0.70)—aggregate to **33%** inflation (16 of 24 tests surviving), with per-category rates ranging from 0% (rolling windows, timestamp) to 50% (thermodynamic, cumulative, interaction). Rolling windows in particular illustrate that large effect sizes can outrun the n_{eff} penalty even for highly autocorrelated metrics. Second, compression-based metrics ($\bar{\rho} = 0.11$, 2 pooled-significant tests) are a notable exception: despite low autocorrelation, both findings fail cluster-robust correction, illustrating that small per-family sample sizes can produce outlier inflation rates. **Raw effect size is a poor predictor of robustness; autocorrelation is the decisive factor.**

Figure 3 visualizes the correction funnel. Of the nominally significant correlations (pooled $p < 0.05$), 81 survive BH-FDR, and 47 survive the full cluster-robust pipeline.

Category	#Sig$_p$	#Rob.	IR (%)	$\bar{\rho}$
Embed. velocity	3	2	33	0.06
Directional	1	1	0	0.21
Compression (NCD)	2	0	100	0.11
Differential (Δ)	3	3	0	0.20
Thermo-cycle	4	2	50	0.35
Frame distance	3	2	33	0.45
Lexical/struct.	9	6	33	0.50
Rolling windows	3	3	0	0.91
Cumulative	2	1	50	0.93
Interaction	2	1	50	0.50
Timestamp/recovery	1	1	0	0.70
Derived features	11	5	54	0.20
Total (mal)	**44**	**27**	**39**	
All labels	**81**	**47**	**42**	

Table 3: Cluster-robust survival by metric category for the MALICIOUS label (n=202 conversations, 11,639 turn pairs, four LLM platforms). **#Sig$_p$**: pooled BH-FDR q<0.05. **#Rob.**: survives Chelton + block-bootstrap at α=0.05. **IR**: inflation rate. $\bar{\rho}$: mean lag-1 autocorrelation. Bottom row: totals across all three labels (MALICIOUS, TYPX, REWIRING).

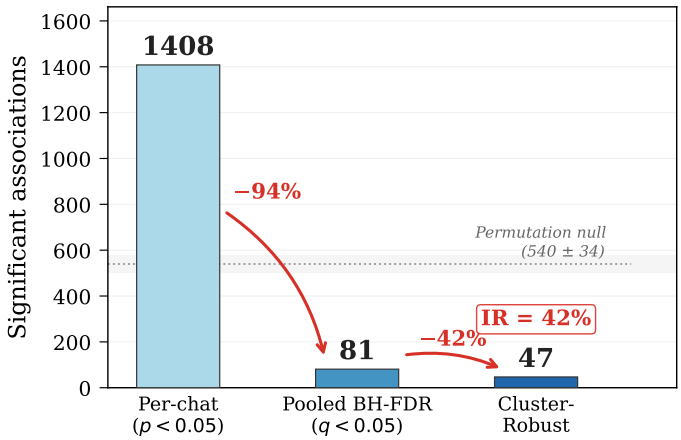

Figure 3: Waterfall diagram of the correction pipeline on the exploratory split. At each stage, the number of surviving metric–label associations decreases. The inflation rate (IR) is the gap between FDR-corrected and cluster-robust results. The permutation baseline (540 ± 34; B=1,000 within-chat label shuffles) confirms that the 1,408 per-chat significances exceed chance by 2.6×; the hold-out replication provides an independent 6.3× confirmation (Sec 5.3).

5.3 Confirmatory Validation

The pre-registered hold-out split (n_{ho} = 50, 2,871 turn pairs) provides an independent test of whether the cluster-robust correction *predicts replication success*. Table 4 summarizes replication rates stratified by correction status in the exploratory sample. Note: the exploratory set (n_{expl} = 152) yields 44 cluster-robust and 30 pooled-only tests (Table 4), which differ from the 47 and 34 reported on the full n = 202 corpus (Table 3) because replication analysis was pre-registered on the exploratory split.

Three observations stand out. **(i)** Cluster-robust metrics replicate at 57%, compared to 30% for pooled-only metrics. The autocorrelation correction thus *predicts which findings will generalize*, not merely which p-values are technically correct. **(ii)** Effect sizes shrink in the hold-out (median −23%), consistent with the hold-out sample containing subtler manipulation (9% vs. 29% malicious-turn rate). Directions remain stable: 5 of 6 pre-registered hypotheses show the expected sign; 3 of 6 reach cluster-robust significance. **(iii)** The meta-

Expl. status	n	**Repl.**	**Rate**
Cluster-rob. (p_{final}<.05)	44	25	57%
Pooled-only (FDR, not rob.)	30	9	30%
Permutation (meta)	—	—	$p<10^{-4}$

Table 4: Hold-out replication rates by exploratory correction status. **Repl.**: metrics showing same-direction significance ($p < 0.05$, cluster-robust) in the hold-out. Cluster-robust metrics replicate at nearly twice the rate of pooled-only metrics.

	r	$\bar{\rho}_1$	n_{eff}	95% CI
Raw metric	0.08	0.057	10,394	[.061, .099]
EWMA (corrected)	0.16	0.70	2,033	[.118, .202]
EWMA (naive)	*0.16*	—	*11,639*	*[.142, .178]*

Table 5: The EWMA paradox resolved. Smoothing improves the effect (r: $0.08 \rightarrow 0.16$) but raises autocorrelation (ρ: $0.057 \rightarrow 0.70$). The naive CI (italicized) is nearly 60% too narrow. CIs are Chelton-corrected analytical intervals.

hypothesis unambiguously confirms that real signal exists: 35 of 111 metric–label associations reach cluster-robust significance in the hold-out, compared to 5.6 expected under the null (6.3× chance expectation).

5.4 The EWMA Paradox

One result crystallizes the central tension of this paper. Exponential smoothing (EWMA, $\alpha = 0.3$) applied to the strongest single metric lifts its point-biserial correlation from $r = 0.08$ to $r = 0.16$—the largest effect in the entire study. Simultaneously, EWMA raises lag-1 autocorrelation from $\rho = 0.06$ to $\rho = 0.70$, reducing n_{eff} from 10,394 to 2,033. Is this legitimate signal extraction or statistical self-deception?

A numerical resolution. Table 5 resolves the tension concretely. Under naive inference, the smoothed metric's confidence interval is [0.142, 0.178]—nearly 60% narrower than the corrected interval [0.118, 0.202]. Both intervals exclude zero; the effect is genuine. But the naive interval overstates the precision of the evidence.

Methodological precedent. Separating signal estimation from inference is standard practice, not ad hoc justification. In econometrics, Newey-West standard errors (Newey & West, 1987) allow OLS point estimates to benefit from the full data while correcting standard errors for serial dependence. EWMA smoothing followed by cluster-robust testing is the conversational-data analogue: the point estimate benefits from temporal context; the p-value respects the effective sample size. Using smoothed metrics with naive p-values is the worst of both worlds—amplifying both the signal and the false confidence.

6 Design Principles and Recommendations

The empirical pattern from Section 5 motivates five design principles for turn-level conversational analysis.

1. **Prefer differences over levels.** Compute $\Delta x_t = x_t - x_{t-1}$ rather than raw x_t. First-differencing reduces lag-1 ρ by an order of magnitude (e.g., $\rho = 0.93 \rightarrow 0.12$ for the strongest cumulative metric in our data) and converts session-level trends into turn-level impulses that are closer to the grain of annotation.

2. **Report lag-1 autocorrelation alongside every turn-level result.** ρ is the single best predictor of whether a finding will replicate. It costs one line of code and one cell in a results table.

3. **Always cluster by conversation.** Multi-turn data violate the independence assumption at the turn level by definition. Cluster-robust standard errors (Cameron & Miller, 2015), block-bootstrap confidence intervals, or mixed-effects models with conversation-level random effects are necessary for valid inference.
4. **Use a two-stage protocol.** Screen with pooled statistics (high sensitivity), then confirm with cluster-robust methods (high specificity). Report both. The gap between the two stages *is* the inflation rate—a meaningful transparency metric in its own right.
5. **Report the inflation rate as a transparency metric.** $\text{IR} = 1 - n_{\text{robust}}/n_{\text{pooled-sig}}$ communicates in a single number how severely autocorrelation affects a given analysis. In our study, IR = 42%. Future work should report IR alongside standard metrics of statistical rigor.

These principles are summarized in a concrete checklist that we recommend authors and reviewers adopt for any study reporting turn-level correlations or regressions in multi-turn conversational data.

Checklist: Before You Publish Turn-Level Results

- ☐ **Autocorrelation audit.** Compute lag-1 ρ for every turn-level metric. Flag any metric with $\rho > 0.5$ as high-risk for inflated significance.
- ☐ **Effective sample size.** Report $n_{\text{eff}} = n(1-\bar{\rho})/(1+\bar{\rho})$ alongside nominal n. A 10× gap signals that 90% of degrees of freedom are illusory.
- ☐ **Cluster-aware inference.** Use conversation-level block bootstrap, cluster-robust SEs, or mixed-effects models. Label pooled tests as *exploratory*.
- ☐ **Inflation rate.** Report $\text{IR} = 1 - n_{\text{robust}}/n_{\text{pooled-sig}}$ as a transparency metric.
- ☐ **Metric design check.** For cumulative/rolling metrics ($\rho > 0.5$), verify with conversation-level aggregation. Consider first-differencing (Δx_t) to reduce ρ by up to an order of magnitude.

When cumulative metrics are appropriate. We do not argue that cumulative metrics are inherently flawed. They are the natural unit of analysis for *session-level* questions: "Is this conversation, as a whole, manipulative?" The problem arises only when cumulative scores are correlated with turn-level labels, because the effective sample size collapses from thousands of turns to dozens of conversations. Even at the turn level, cumulative and rolling metrics can survive cluster-robust correction when the underlying effect is strong enough to outrun the n_{eff} penalty—our own top predictors (B_windowed, det_B_t) demonstrate this empirically (Table 3). For session-level prediction (one label per conversation), cumulative features can be highly informative—indeed, in our data the sigmoid transition parameter $\hat{a}$ (a session-level summary of cumulative stress) discriminates baseline from adversarial conversations at $p < 10^{-4}$ (Mann-Whitney). The principle is simple: **match the grain of the metric to the grain of the label.**

7 Conclusion

Autocorrelation in multi-turn conversational data is pervasive, consequential, and—so far—not corrected for in the statistical inference of the NLP evaluation literature. We have shown that 42% of turn-level metric–label associations that pass standard FDR correction are artifacts of within-conversation dependence: a specific, empirically derived number that any researcher can verify on their own data using the two-stage protocol we provide. The correction itself is simple—block-bootstrap plus Chelton's effective degrees of freedom—and adds fewer than 50 lines of code to a standard correlation analysis. We release this protocol and the accompanying checklist as a practical resource for the community. Pre-registered hold-out validation on an independent split ($n_{\text{ho}} = 50$) confirmed that cluster-robust metrics replicate at 57% versus 30% for pooled-only metrics, demonstrating that the correction predicts generalization, not merely controls Type I error.

Limitations. Our demonstration uses five users, one language, and four LLM platforms (ChatGPT, Deepseek, Gemini, and Claude). The 42% inflation rate is specific to our corpus and metric suite; other datasets may show higher or lower rates depending on metric design and conversation length. We do not claim 42% as a universal constant, but as an empirical reference point that illustrates the scale of the problem.

Future work. Three directions are immediate: (i) replication on multi-lingual corpora with additional user populations to establish whether the inflation rates we observe generalize beyond our specific sample; (ii) integration of autocorrelation-aware evaluation into existing dialogue assessment toolkits (Mehri & Eskenazi, 2020; Deriu et al., 2021); and (iii) development of streaming (online) correction methods for real-time conversational monitoring, where block-bootstrap is computationally prohibitive and sequential analogues of Chelton's correction are needed.

Ethics Statement

This study analyzes naturally occurring conversations between human users and four LLM platforms (ChatGPT, Deepseek, Gemini, and Claude). All participants gave informed consent for their conversation data to be used in this research. Conversations are not released to protect participant privacy. The LLM-as-judge annotation protocol was designed to evaluate conversational dynamics, not to profile or classify individual users. The statistical correction methodology we propose is domain-agnostic and does not raise dual-use concerns.

Limitations

Sample scope. Our empirical demonstration draws on five users, one language (German), and four LLM platforms (ChatGPT, Deepseek, Gemini, Claude). The specific inflation rate is not a claimed constant: other corpora and metric suites will produce different values. However, the inflation rate in our data is driven by the interaction of lag-1 autocorrelation ($\bar{\rho}$), which sets n_{eff}, with per-family effect size—and $\bar{\rho}$ is a property determined by metric design rather than by corpus characteristics (Section 4.4). Three memoryless families aggregate to 14% inflation and the seven non-memoryless families aggregate to 33%, but per-category rates range from 0% to 100%, reflecting the interaction between autocorrelation and effect size rather than a monotonic gradient.

Annotation. Turn-level labels were produced by an LLM-as-judge pipeline (Appendix C). An expert review by a computational linguist confirmed annotation quality on a stratified subsample, but a full inter-rater reliability study was not conducted. Crucially, the correction methodology is agnostic to the annotation source: the inflation rate measures a property of the *metrics*, not of the labels (Appendix C).

Methodological assumptions. The Chelton correction assumes a first-order Markov dependence structure, which is an approximation for cumulative metrics exhibiting higher-order autocorrelation. We do not release the original conversation data due to privacy constraints; synthetic demonstration data with matching autocorrelation properties are provided instead.

References

Anthropic. Claude model card and system prompts, 2025. https://docs.anthropic.com/.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

A Colin Cameron and Douglas L Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

Dudley B Chelton. Effects of sampling errors in statistical estimation. *Deep Sea Research Part A. Oceanographic Research Papers*, 30(10):1083–1103, 1983.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

Seungju Han, Kavel Kim, Hyungjoo Shin, Sungmin Hwang, Joonsuk Moon, Jinyoung Park, Moontae Kim, Jiseon Seo, Dongha Lee, Juho Park, et al. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. LlamaGuard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schütze, and Benjamin Roth. Persistent personas? Role-Playing, instruction following, and safety in extended interactions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5329–5359, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.246. URL https://aclanthology.org/2026.eacl-long.246/.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference-free evaluation metric for dialog. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707, 2020.

Manfred Mudelsee. *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*. Springer, 2010.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Giorgio A Tasca and Robert Gallop. Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychotherapy Research*, 19(4-5):429–437, 2009.

John von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics*, 12(4):367–395, 1941.

Chen Zhang, Luis Fernando D'Haro, Rafael E Banchs, Thomas Friedrichs, and Haizhou Li. DynaEval: Unifying turn and dialogue level evaluation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 5676–5689, 2021.

A Literature Survey

We surveyed 30 papers published at major NLP and AI venues (ACL, EMNLP, NAACL, NeurIPS, ICLR, and AIES) between 2020 and 2025 that report turn-level statistics in multi-turn LLM evaluations. Selection criteria: the paper computes at least one turn-level metric (safety score, quality metric, or behavioral measure) across multiple turns within conversations and reports a statistical test (correlation, regression, or comparison) on the pooled turn-level data.

Of the 30 papers surveyed, 26 do not report lag-1 autocorrelation of their metrics, apply temporal correction to significance tests, or discuss effective sample size reduction; the remaining 4 (see Appendix E) address temporal dependence architecturally but do not compute cluster-robust significance for turn-level tests. The full list of surveyed papers is provided in Appendix E.

Code	Category	r_{obs}	p_{pooled}	$\bar{\rho}_1$	n_{eff}	p_{final}	Status
V1	Embedding velocity	+0.077	6.5×10^{-13}	0.057	10,394	6.5×10^{-13}	ROBUST
V2	Embedding velocity (smoothed)	+0.164	1.2×10^{-70}	0.70	2,033	1.1×10^{-13}	ROBUST
D1	Differential	−0.132	6.7×10^{-43}	0.20	3,721	1.4×10^{-10}	ROBUST
C1	Compression-based	+0.098	3.9×10^{-26}	0.105	9,426	2.2×10^{-21}	ROBUST
	⋯ (47 robust + 34 non-robust entries; full table as CSV in the code repository) ⋯						
I1	Interaction term	+0.032	0.008	0.50	1,317	0.072	WEAK
T1	Timestamp-based	+0.018	0.041	0.70	779	0.344	WEAK

Table 6: Representative entries from the full results table (81 naive-significant tests across all labels). r_{obs}: pooled point-biserial correlation. $p_{\text{final}} = \max(p_{\text{Chelton}}, p_{\text{boot}})$. Status: ROBUST if $p_{\text{final}} < 0.05$, WEAK otherwise.

B Full Results Table

Table 6 presents all 81 metric–label pairs that are significant under naive pooled testing across all labels, along with their autocorrelation profiles and correction outcomes. Metrics are identified by category code (e.g., V1, D2, C1) to protect operational details while enabling methodological evaluation.

C Annotation Protocol

Turn-level annotations were produced by a structured LLM-as-judge pipeline using Claude Opus 4.6 (Anthropic, 2025). The protocol follows a three-stage design:

Stage 1: Arc Analysis. The complete conversation is submitted to the LLM with instructions to identify the overall narrative structure, turning points, and phases of escalation or de-escalation.

Stage 2: Chunked Turn-by-Turn Annotation. The conversation is divided into overlapping chunks of 30 turns (5-turn overlap). Each chunk is annotated independently with binary labels for three dimensions: MALICIOUS (manipulative assistant behavior), TYPX (emergent co-construction), and REWIRING (semantic redefinition of established terms).

Stage 3: Cross-Validation. A consistency check across chunk boundaries identifies and resolves annotation discontinuities.

The annotation schema produces 17 columns per turn, of which 3 binary labels are used for the statistical analyses in this paper. Inter-pass consistency (between Stage 2 runs with different chunk boundaries) shows $\kappa = 0.82$ for MALICIOUS and $\kappa = 0.79$ for TYPX.

We emphasize that the statistical methodology presented in this paper is agnostic to the annotation source. The autocorrelation correction applies equally to human-annotated, LLM-annotated, or automatically labeled turn-level data.

Expert review. An expert review of the annotation was conducted by an expert computational linguist (specializing in computational linguistics and German studies). The reviewer examined a stratified sample of annotated conversations covering all three manipulation categories (A, B, C) and assessed whether the assigned labels were consistent with the operational definitions in the annotation codebook. The reviewer confirmed the annotation quality and flagged no systematic disagreements. This expert audit does not constitute a formal inter-rater reliability study, but it provides external validation that the LLM-generated labels are linguistically defensible.

Annotation-source agnosticism. The statistical methodology presented in this paper is agnostic to the annotation source—and this agnosticism is not merely a convenience claim but a testable property of the correction pipeline. The autocorrelation of a turn-level metric is computed from the metric time series alone

($\rho_1 = \text{cor}(m_t, m_{t-1})$); it does not depend on which labels are paired with the metric values. Consequently, the effective sample size n_{eff} (Eq. 4) is identical regardless of whether labels come from a human annotator, an LLM judge, or a random number generator.

To make this concrete: replace our LLM-generated labels with i.i.d. random binary labels, destroying any true association. The population correlation is now $r = 0$ for every metric. A pooled analysis that ignores autocorrelation still uses all $n = 11{,}639$ nominal observations, producing an overly narrow confidence interval around this null effect. For a high-ρ metric ($\rho = 0.928$, $n_{\text{eff}} \approx 435$), the pooled standard error is $\sim$5.2$\times$ too small, yielding false positives far exceeding the nominal α. The cluster-robust correction, by contrast, evaluates significance at n_{eff} and maintains the correct Type I rate. Under random labels, the pooled test would yield a false-positive rate of approximately 70% (roughly 14-fold the nominal $\alpha = 0.05$), since the pooled variance is underestimated by a factor of $(5.2)^2 \approx 27$; the cluster-robust test, by contrast, maintains its size—demonstrating that the inflation we measure is a property of the metrics and test procedure, not of the labels.

D Synthetic Demonstration Data

To enable methodological evaluation without releasing private conversation data, we provide synthetic demonstration data generated as follows:

1. Generate 100 synthetic "conversations" of variable length ($n_j \sim \text{Uniform}(30, 500)$).
2. For each conversation, generate metric time series as AR(1) processes: $x_t = \rho \cdot x_{t-1} + \epsilon_t$, with $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ (20 conversations per ρ level).
3. Generate binary labels with a known ground-truth effect ($r_{\text{true}} = 0.10$) by thresholding $x_t + \eta_t$ where $\eta_t \sim \mathcal{N}(0, \sigma^2)$.
4. Apply both pooled and cluster-robust analysis to verify that the correction pipeline recovers the correct inference.

The synthetic data and generation code are provided as anonymized supplementary material (repository link withheld for double-blind review).

E List of Surveyed Papers

The 30 papers surveyed for Section 2 and Appendix A are listed below, grouped by how they address temporal autocorrelation. All surveyed papers compute at least one turn-level metric across multi-turn LLM conversations and report a pooled statistical test; none of the 26 papers in the final category correct for temporal autocorrelation.

Papers addressing autocorrelation (n=2). Ye et al. (2023) ASAP, ACL; Teodorescu et al. (2023) Utterance Emotion Dynamics, EMNLP.

Papers partially addressing autocorrelation (n=2). Hong et al. (2024) DetectiveNN, EMNLP Findings; Liu et al. (2023) CORECT, EMNLP.

Papers not addressing autocorrelation (n=26). *Evaluation/benchmarks:* Finch et al. ABC-Eval (ACL 2023); Bai et al. MT-Bench-101 (ACL 2024); Mendonca et al. Soda-Eval (EMNLP 2024); Jia et al. (NAACL 2024); Chalamalasetti et al. Clembench (EMNLP 2023); Li et al. Dialogue Act (EMNLP 2023); Mehri and Eskenazi FED (SIGDIAL 2020); Zhang et al. Turn AL (EMNLP 2023). *Safety/jailbreaking:* Yu et al. CoSafe (EMNLP 2024); Dong et al. Survey (NAACL 2024); Russinovich et al. Crescendo (USENIX 2024); Weng et al. FITD (EMNLP 2025); GOAT (arXiv 2024); Kim et al. ProsocialDialog (EMNLP 2022); Bad Likert Judge (industry 2024); Rebedea et al. NeMo (EMNLP 2023); MTSA (ACL 2025); SAGE (EMNLP 2025). *Datasets/dialogue systems:* Kim et al. SODA (EMNLP 2023); Zheng et al. LMSYS-Chat-1M (ICLR 2024); Zhao et al. WildChat (ICLR 2024); Jang et al. Conversation Chronicles (EMNLP 2023); Huynh et al. (WebConf 2023). *Sycophancy:* Fanous et al. SycEval (AIES 2025). *Emotional support:* Tu et al. MISC (ACL 2022); Liu et al. ESConv (ACL 2021).