Whose Repair? A Multi-Loop Diagnostic of Interactive Alignment in Human–LLM Conversation

Anonymous ACL submission

Abstract

Conversational large language models (LLMs) enable humans to interact with powerful systems, yet extended interactions remain brittle, exhibiting grounding rifts, sycophantic agreement, and multi-turn instruction degradation among other failures. These are signatures of one structural limitation: the lack of mechanisms to coordinate with users as the interaction unfolds. Extending Pickering and Garrod's Interactive Alignment Model, we model alignment as an **online, concurrent control process updated turn-by-turn** through observable hand-offs, under a phase structure that re-weights violation cost as the interaction matures. We advance four claims: (i) the unit of analysis is the *inter-turn transition*, instantiated by four concurrent loops (Grounding, Continuity, Negotiation, and Epistemic); (ii) misalignment carries a *phase-dependent cost*, disproportionately larger once expectations have solidified; (iii) *percolation across representational levels* is an evaluation target; and (iv) current LLMs cannot largely self-detect and repair misalignment, externalizing the cost to users. The framework is falsifiable along two axes: a within-family correlation between system-initiated epistemic signaling and reduced sycophancy under training-pipeline controls, and a within-dialogue phase asymmetry in repair cost. We re-read failure literature through the framework, identify loop-level measurement targets, and argue that inter-turn repair events should be the primary observable for evaluation.

1 Introduction

Conversation is a natural channel through which humans retrieve information, coordinate tasks, and collaborate with others (Clark, 1996). Using language as the interface removes technical barriers and extends access to powerful systems far beyond expert users, a key driver of integration into customer service, enterprise knowledge management, and productivity tools (McKinsey & Company, 2025). Conversational AI embodies a long-standing vision in human–computer interaction: making machines usable in the way humans already interact through language (Luger and Sellen, 2016b; Branigan et al., 2010).

Yet in practice, the experience is far less satisfying than human–human conversation. Users report brittle interactions and shallow engagement, and the recurring failure modes documented in the literature: grounding rifts (Shaikh et al., 2025), sycophancy and pseudo-alignment (Sharma et al., 2023; Perez et al., 2022), degradation in multi-turn instruction following (Sirdeshmukh et al., 2025), overconfident assertions in domains where the model's own uncertainty is high (Lin et al., 2022; Kadavath et al., 2022), and persona drift and long-context degradation (Li et al., 2024; Liu et al., 2024). These are symptoms of a structural limitation: current LLM systems lack the mechanisms to coordinate with users as the interaction unfolds: to ground shared meaning, recalibrate expectations, and repair misunderstanding turn by turn. Predominantly single-turn preference-based training, though multi-turn extensions are emerging (Shi et al., 2024), gives them limited incentive to acquire those mechanisms.

We propose to take that limitation seriously, and to reframe what conversational human–LLM alignment is. Previous works specified *what* is aligned (Shen et al., 2024), *who* models whom in a sequential three-stage modeling cycle via Mutual Theory of Mind (MToM) (Wang and Goel, 2024), characterized one sub-process (grounding) in depth (Shaikh et al., 2025), and formalized reward dynamics (Carroll et al., 2024) but abstracted away the interactional layer. These accounts share a common gap of not treating alignment as an online control process whose state is updated dynamically within the interaction. This paper's framework models the interaction itself, and advances four claims that define a new class of objects for alignment research:

(1) Alignment is a control process The unit of analysis for human–LLM alignment should be the *inter-turn transition*. Unlike MToM's stage-sequential modeling cycle, our four inner loops (Grounding, Continuity, Negotiation, Epistemic) can be active on the same turn, with *transitions* observable as per-turn events (§4, Figure 1, Table 2).

(2) Misalignment has a phase-dependent cost. An Early-to-Consolidation phase structure reweights every loop's tolerance: the same failure mode (persona drift, instruction loss, or mid-context retrieval failure) is benign early and disproportionately costly once expectations solidify. This is a directly testable prediction that aggregate multi-turn scores currently average out, and one not foregrounded in prior accounts.

(3) Percolation across representational levels is an evaluation target. We treat IAM's (Pickering and Garrod, 2004) percolation claim as a property to *measure*, defining a *percolation strength* metric that scores whether lexical or referential alignment at turn t survives to situation-model alignment at turn $t + k$. Long-context (Liu et al., 2024) and persona-drift (Li et al., 2024) benchmarks provide multi-turn dialogues suitable for estimating this metric, but report only aggregate accuracy at $t + k$ rather than the conditional persistence of what was entrained at t.

(4) A single structural deficit underlies four failure literatures. The two-axis framework yields a falsifiable diagnostic claim: across four loops, the cell in which the *system* both detects and repairs misalignment (System Self-Repair) is largely absent in current LLMs, and the cost is externalized to *user-side self-repair*. Grounding rifts, multi-turn instruction degradation, sycophancy, and persona drift are surface signatures of this structural deficit rather than four independent problems (§4.3).

Together, these claims reposition alignment from a single-turn optimization target to an instrumented interactional process and motivate the loop-level evaluation targets in §5. The reframing carries a methodological commitment: NLP should start treating inter-turn repair events (each indexed by which loop is engaged and by who detects and repairs the misalignment) as the primary observable.

2 Conversation as Interactive Alignment

This section reviews the psycholinguistic foundations on which our framework builds: conversation as a joint activity, interactive alignment as its mechanism, and the levels at which alignment operates.

2.1 Conversation as a Joint Activity

Human conversation is a joint activity in which interlocutors coordinate not only what they do together but how they talk while doing it (Clark, 1996): success depends on continuous negotiation, monitoring of partner signals, and incremental update of mutual beliefs. There are two related central notions: *common ground*, the shared body of knowledge that allows interlocutors to interpret each other without re-inventing every detail, and *grounding*, the collaborative activity of showing evidence of understanding through backchannels, clarifications, and repair (Clark and Brennan, 1991). Users approach conversational AI with the same expectations and modify their own behavior to compensate when the system fails to participate symmetrically (Luger and Sellen, 2016a; Beneteau et al., 2019); single-turn evaluation systematically underestimates this coordinative work.

2.2 Interactive Alignment Model

IAM (Pickering and Garrod, 2004) describes how interlocutors spontaneously converge at multiple linguistic levels (lexical, syntactic, referential, situational) through perception–production coupling. Its central claim is *percolation*: alignment at one level propagates to others, so surface coordination scaffolds higher-level common ground without explicit perspective-taking. LLMs offer a partial analog in in-context conditioning but do not implement percolation uniformly; we invoke IAM as a normative model, not a mechanistic claim. The framework's predictions stand on behavioral signatures of percolation (entrainment at turn t correlating with situational consistency at $t + k$), independent of internal mechanism.

2.3 Levels of Alignment

Interactive alignment operates at four levels most actionable for NLP with text LLMs and most often diagnosed in published multi-turn failure studies: *lexical/referential* (entrainment on labels and reference; Brennan and Clark, 1996; Clark and Schaefer, 1989), *syntactic* (structural priming; Branigan et al., 2000; Pickering and Garrod, 2004), *pragmatic* (turn-taking and repair; Schegloff, 2000; Schegloff et al., 1977), and the *situation model* (shared task representation; Garrod and Pickering, 2004). Each level maps onto one or more of the

loops introduced in §4. Percolation across levels is itself an evaluation target, since lexical or referential failures cascade into situation-model breakdown over the course of an interaction (§5).

3 Bridging to Human–LLM Interaction

Psycholinguistic alignment models describe how humans coordinate, but do not, on their own, explain why interactions with LLMs remain fragile. Closing that gap requires four constructs that sit between cognition and interaction (MToM, Trust Dynamics, Power Dynamics, and Situational Awareness) and a precise account of what makes LLMs structurally different from a human partner.

3.1 Mutual Theory of Mind

Theory of Mind (ToM) Premack and Woodruff, 1978 is the cognitive substrate of mutual understanding; MToM (Wang and Goel, 2024) extended it to human–AI interaction symmetrically, but the capability asymmetry is central: humans engage in sophisticated ToM reasoning while LLMs possess only simulation-style perspective-taking. This asymmetry is the structural source of grounding rifts (Shaikh et al., 2025) and makes Grounding a site of recurring failure. Cognitive Load Theory (Sweller, 1994) adds a resource constraint: mutual modeling is bounded by finite human attention.

3.2 Trust Dynamics

Trust in AI rests on competence, benevolence, and integrity (Mayer et al., 1995) and is fragile: an early mistake can collapse it before mutual understanding forms (de Visser et al., 2020), and calibration shifts abruptly (Hoff and Bashir, 2015). Without matched transparency and reliability cues (Wischnewski et al., 2023), users abandon collaboration or overtrust and withhold corrective feedback. Trust repair requires visible evidence of learning (Zhang et al., 2024).

3.3 Power Dynamics

AI-mediated communication introduces asymmetries in autonomy and perceived authority (Hancock et al., 2020), filtering whether cognitive alignment results in genuine understanding or deference. Users may defer to perceived system authority, producing *pseudo-alignment* (surface agreement reflecting compliance rather than mutual adjustment) (Giles and Ogay, 2013; Hou et al., 2024). The Negotiation Loop is the natural site for detecting this pattern, since it is where surface convergence and representational convergence can come apart.

3.4 Differences in Human–LLM Alignment

The theories reviewed were developed for human–human interaction or earlier human–computer systems. Six properties of contemporary LLMs reshape the alignment problem: (1) stochastic, surface-driven generation (disrupts Negotiation by producing surface alignment without representational change); (2) context-window memory with no episodic store (Liu et al., 2024) (disrupts Continuity and the Consolidation Phase); (3) absence of embodiment or perceptual coupling (Grounding reduces to statistical priming rather than perception–production); (4) world models frozen at training time (Continuity violations do not persist as model updates); (5) response biases induced by reinforcement learning from human feedback (RLHF; Sharma et al., 2023; Perez et al., 2022), which collapse Negotiation into compliance; (6) and in-context learning as a partial priming analog (the only direct computational substrate for Negotiation; other loops lack equivalents). Three of these (RLHF biases, frozen world models, absence of embodiment) converge in disabling System Self-Repair across loops (§4.3).

3.5 Situating against Existing Accounts

Recent accounts agree that human–LLM alignment is bidirectional and temporally extended, but none treat it as an online control process with concurrent sub-loops, observable per-turn transitions, phase-dependent thresholds, and per-turn measurable outputs jointly. Shen et al. (2024) cataloged *what* is aligned (values, capabilities, behaviors) without specifying process; MToM (Wang and Goel, 2024) specified *who* models whom but in a sequential modeling cycle rather than concurrent sub-processes of distinct coordination work; Shaikh et al. (2025) characterized one sub-process (grounding) in depth, leaving its interactions with continuity, negotiation, and epistemic loops unaddressed; Carroll et al. (2024)'s dynamic-reward Markov decision process formalized reward dynamics over time but abstracted away the interactional layer at which alignment is constructed or lost. Our framework is defined precisely by the conjunction of the four properties above. It does not displace these accounts (e.g., MToM in particular can be read as the modeling substrate that our Grounding and Negotiation loops require) but adds

the interactional control structure within which their findings become per-turn observables and per-turn evaluation targets. Closest is the assistance-games framing of Hadfield-Menell et al. (2016), which casts alignment as cooperative inference; our framework is its interactional, per-turn, observable counterpart. Relative to dialogue state tracking (Budzianowski et al., 2018), the framework adds cognitive-functional axis that slot-level updating leaves implicit.

4 Proposed Interactive Framework

We carve the framework along two orthogonal axes. The first is **cognitive-functional**: what coordination work the interaction is doing (grounding referents, maintaining continuity of established state, negotiating stance, or surfacing knowledge gaps). This gives four inner loops (§4.1) running concurrently within an outer interaction cycle that integrates them across turns and phases. The second is **event-structural**, following the conversation-analytic typology of repair (Schegloff et al., 1977): each repair event has a detector[1] and a repairer, each either user or system, yielding four mutually exclusive cells (Figure 1). A single event has both a cognitive type (which loop) and a structural cell (who played which role). Loop entry is operationalized by *cue families*: semantic (natural-language-inference-style contradiction, referent mismatch), syntactic (hedges, modal verbs, reformulations, repair markers), and pragmatic (clarification questions, disagreement markers, terminology switches); they appear in both detector roles, with the producing party distinguishing them. Where a cue is consistent with more than one loop, loop identification is determined by the content of the trouble source: referent or meaning (Grounding), prior commitment or expectation (Continuity), stance or representation (Negotiation), or knowledge or evidence gap (Epistemic). When a turn carries multiple trouble sources, loop assignment follows the precedence Grounding > Continuity > Negotiation > Epistemic, on the rationale that referent/meaning failures upstream-dominate downstream expectation, stance, or knowledge effects. This ordering is adopted for annotation determinism; alternatives are defensible, and sensitivity to ordering is left to empirical study.

[1]Our *detector* operationalizes Schegloff's *initiator* as a turn-level cue: the party whose utterance signals trouble. The cognitive act of internally detecting misalignment without surfacing it (the unobserved precursor to a Cell 4 event) is treated as a separate, latent state.

Table 1: A partition of inter-turn repair events by *detector* (who signals trouble) and *repairer* (who acts on it). The Negotiation Loop cross-cuts Cell 1 and Cell 3: productive negotiation combines user push-back (Cell 1) with system-surfaced uncertainty (Cell 3). Cell 0 (not shown) is the latent state in which no detector cue fires in either party (pseudo-alignment).

	User repairs	**System repairs**
User detects	Cell 1 GROUNDING User-Signaled Repair	Cell 4 (×) System Self-Repair
System detects	Cell 2 CONTINUITY User Self-Repair	Cell 3 EPISTEMIC System-Signaled Probe

This two-axis carving yields a single falsifiable structural claim: **across all four loops, current LLMs largely lack the System Self-Repair cell, and the cost of that absence is externalized to users via User Self-Repair.** The four failure literatures we re-read in §4.3 (grounding rifts, sycophancy, multi-turn instruction degradation, and persona drift) are surface signatures of this single structural deficit. The Continuity Loop makes the asymmetry most visible (user re-issuance under failed system monitoring), but the same deficit appears in muted form across the other three loops. Cell 4 is therefore a cross-cutting system-side capability whose absence we predict to be observable wherever any of the four loops fires; iterative self-refinement (Madaan et al., 2023) is its partial intra-response substrate, and Huang et al. (2024) provides direct evidence of the deficit at the reasoning layer. The absence decomposes into three mechanisms with distinct remedies (failing to detect, detecting but not surfacing, or being surfaced but then overridden downstream), which the framework distinguishes but does not adjudicate. Here are the three clarifications of scope:

Outcome layer, not mechanistic origin. Sycophancy, lost-middle, persona drift, and instruction degradation arise from distinct underlying mechanisms, and the framework does not equate them at that level. It claims only that they share an *interactional outcome*: at the turn boundary, the system fails to flag or self-repair the drift, and the user absorbs the cost via self-repair. A dissociation follows: surface remedies targeting one mechanism (e.g., RLHF reweighting against length bias) lower that specific failure rate but leave the Cell-2/Cell-4 ratio and the structural deficit intact.

MECE over observable events. Cells are mutually exclusive and collectively exhaustive (MECE) over *observable* repair events: each event is assigned to exactly one cell by the detector × repairer combination (Table 1). Pseudo-alignment is a *Cell-0 latent state* in which disagreement exists internally but no detector cue fires, registering as the absence of any cell rather than as Cell 1. The loops are functional categories of coordination and overlap on a single turn (Negotiation deliberately cross-cuts Cell 1 and Cell 3, §4.1), where cells partition events, loops partition function.

Intra/inter-response system self-repair is a continuum. Cell 4 is defined at the turn boundary because the framework's unit of analysis is the inter-turn transition (claim 1): intra-response self-correction that does not surface as a turn-level commitment is invisible to the user's repair workload. Reasoning-time critique and revision (Madaan et al., 2023; Huang et al., 2024) instantiate partial system self-repair to the extent that the correction surfaces as such a commitment; the framework predicts that this surfacing is what reduces burden on user self-repair.

4.1 The Loops

The four inner loops differ in which coordination work they do and in which cell of Table 1 dominates their observable repair events. We describe each in turn, and failure modes are developed in §4.3.

The **Outer Loop** is the integrating cycle. It accumulates inner-loop outcomes (shared context, task state, expectations, and trust priors) across turns and episodes, and its phase state re-weights inner-loop thresholds. As the interaction enters the Consolidation Phase, the same cue families are evaluated against stricter thresholds; this is phase-dependent re-weighting, not a fifth process.

The **Grounding Loop** maintains referent and meaning common ground. It is dominated by Cell 1 (User-Signaled Repair): the user produces a repair cue (a clarification request, correction, repair marker such as *"wait"*, or a referent reformulation) and the system updates and re-produces. Its characteristic failure mode is the *grounding rift* (Shaikh et al., 2025). Users tolerate clarification exchanges with LLMs more readily than with human partners, which makes this loop a natural site for visible uncertainty signaling rather than silent guessing; the symmetric Cell 4 case, in which the system would itself flag a referent mismatch and re-anchor

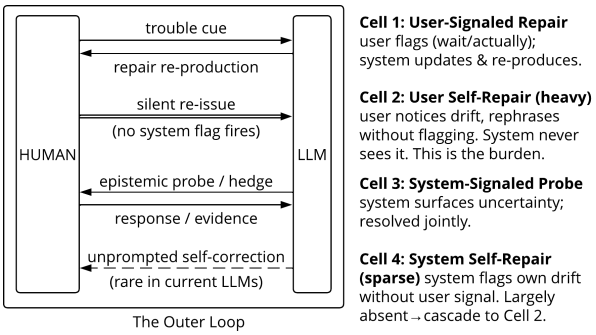

Figure 1: The framework's repair-events, with Human–LLM dyad on the vertical axis. Arrows trace the detector → repairer flow of each event; bracketed labels name the loop each cell predominantly serves. The Outer Loop integrates state across turns, with phase progression from Early to Consolidation, tightening cue-detection thresholds. See more descriptions in §4 and Table 1.

without user signaling, is correspondingly rare.

The **Continuity Loop** maintains behavioral and task-state expectations across turns. It is dominated by Cell 2 (User Self-Repair): the user notices an expectation violation (a forgotten instruction, a drifted persona, or a lost detail) and re-issues silently, because the Cell 4 monitoring that would catch the drift internally is largely absent. Cues are user re-issuance, instruction restatement, expressions of surprise, and attempts to correct the persona; the associated failure literature is multi-turn instruction degradation (Sirdeshmukh et al., 2025) and persona drift (Li et al., 2024). This loop is canonical of the headline claim: the system does not self-monitor, so the user compensates.

The **Negotiation Loop** drives stance and converges representation: terminology agreement, conceptual entrainment, and disagreement resolution. Its productive form mixes Cell 1 (user pushes back) and Cell 3 (system surfaces uncertainty or disagreement). Its degenerate form is the collapse of Cell 3 into surface agreement without representational adjustment, which we call *pseudo-alignment* (§4.3). The cues are terminology switches, hedges, stance shifts, and disagreement markers. The corresponding failure is sycophancy (Sharma et al., 2023; Perez et al., 2022), where power asymmetry collapses convergence onto the user's prior, and Cell 4 fails to flag the unjustified concession.

The **Epistemic Loop** surfaces knowledge gaps and unverified inferences for productive resolution. It is dominated by Cell 3 (System-Signaled Probe): the system detects uncertainty in its own output, or the user probes a knowledge claim, and resolution

is reached jointly through evidence, citations, or candidate offers. The cues are epistemic markers (*"I'm not sure"*), user probes, citation requests, and candidate offers. The cell is rarely triggered in current systems, which prefer confident assertion to acknowledged uncertainty; iterative self-refinement (Madaan et al., 2023) provides a partial computational substrate for the Cell 4 case but operates intra-response rather than across the turn boundaries our framework treats as the unit of analysis.

4.2 Phase Transitions and Loop Coordination

The Outer Loop's phase state governs how strictly the inner loops apply their cue thresholds. We distinguish an **Early Phase**, in which misalignment tolerance is high, and the Epistemic Loop is used productively for capability discovery, from a **Consolidation Phase**, in which expectations have solidified, and equivalent violations carry disproportionately higher cost. Operationally, Early Phase is the prefix during which the user is still revising their model of system capability, proxied by the rate of capability-probing utterances (epistemic markers, hypotheticals, capability questions) relative to a baseline; Consolidation begins when this rate stabilizes. Turn position approximates this on existing benchmarks (Early $\approx$ turns 1–3) but comes apart in short or task-switching dialogues, so the binary label is an exposition device and the underlying state is the continuous rate of probe utterances.

Loops are connected by upward feedback (a closed inner-loop event updates the Outer Loop's situation model), downward modulation (Outer Loop re-weights inner thresholds), and lateral exchange (one loop hands off to another on the same turn). Table 2 specifies, for each inner loop, the cue family that triggers entry and the hand-off target that triggers exit. Loops can be re-entered, and multiple transitions can fire on a single turn. The phase structure yields a directly testable prediction: identical violations should require measurably more user-side repair turns to recover in Consolidation than in Early Phase (Falsifier C, §5.1).

4.3 Re-reading Failure Cases

We re-read five bodies of findings through the framework, not to relabel prior work but to offer an interactional explanation for patterns the literature currently treats as independent.

Grounding Loop and grounding rifts. Shaikh et al. (2025)'s rift taxonomy (missed presuppositions, unrequested elaborations, failure to register user corrections) maps onto the Grounding Loop's failure surface. The loop-level reading names an asymmetry the data show but do not label: rifts are more often resolved by the user than the model, consistent with the prediction that LLMs externalize the cost of grounding rather than sharing it.

Table 2: Inner-loop entry cues and hand-off targets. Each row specifies how a loop is entered (cue family from the dialogue) and where control hands off when the loop's exit condition fires.

Loop	Entry cue	Hand-off target
Grounding	Clarification request, correction, repair marker; contradiction with prior context	Negotiation (if stance disagreement revealed); Outer Loop (if grounding stabilized)
Continuity	User re-issuance, instruction restatement, expression of surprise, persona-correction	Grounding (if misunderstanding); Outer Loop (if expectation must be revised)
Negotiation	Stance disagreement, terminology switch, hedging, disagreement marker	Grounding (re-anchor reference); Epistemic (surface the disagreement)
Epistemic	Epistemic marker (*"I'm not sure"*), user probe, citation request, candidate-offer	Grounding (once evidence resolved)
Outer (Consol.)	Repeated successful loop closures; stable shared lexicon and expectations	Inner loops adopt stricter thresholds; persistent state committed to memory

Continuity Loop and multi-turn degradation. Sirdeshmukh et al. (2025)'s skill-based categories (instruction retention, inference memory, reliable editing, self-coherence) share one interactional pattern: users carry forward expectations, and the model violates them later. Through the Continuity Loop, this is accumulating expectation violations that the system's own monitoring fails to catch.

Negotiation Loop and sycophancy. Sharma et al. (2023) and Perez et al. (2022) showed that LLMs adjust answers toward users' beliefs, reversing correct answers under expressed doubt. The framework offers a complementary reading: sycophancy is the surface signal of *pseudo-alignment*, when power asymmetry collapses convergence onto the user's prior. This predicts that sycophancy should intensify under cues of user authority and that triggering Epistemic-Loop behavior (surfacing disagreement with evidence) should reduce it without sacrificing cooperativeness.

Epistemic Loop and overconfidence. Lin et al. (2022) and Kadavath et al. (2022) found that LLMs produce confident assertions where internal uncertainty is high, rarely surfacing it unprompted. The Epistemic Loop reads this as the System-Signaled Probe cell failing to fire: the model's epistemic state does not reach the dialogue, and users must probe knowledge claims themselves.

Outer Loop and Consolidation-Phase drift. Liu et al. (2024) showed that LLMs fail to retrieve mid-context information; Li et al. (2024) showed that assigned personas erode over multi-turn interaction. Persona drift originates in Continuity and percolates upward when not repaired (claim 3); lost-middle is a situation-model failure. Both are diagnostic of Consolidation-Phase drift.

These five independent lines of evidence share one pattern (system-side detection and repair largely missing, user-side compensation pervasive), which is precisely the structural deficit our headline diagnostic predicts. Per the outcome-layer clarification, the framework unifies the inter-turn signature each leaves on the user's repair workload, not their mechanistically distinct origins.

5 Implications for NLP Research

If alignment is an interactional process rather than a fixed target, NLP's current methods for training and evaluating aligned systems are at best partial.

5.1 Evaluation Targets and Open Questions

Instead of a new benchmark, we identify four measurement classes that the interactional view makes possible and single-turn preference scoring cannot express, each tied to a specific loop and failure mode. For each, we give an operational definition, a corpus where it could be estimated, and the failure class that aggregate preference scores absorb.

Grounding integrity. Grounding integrity is the per-turn rate at which a system's response asserts or relies on information not established in the dialogue, or contradicts the established context. It can be estimated from the rift dialogues of Shaikh et al. (2025) or via NLI-style annotation on WildChat (Zhao et al., 2024). The metric is needed because preference-only rankings reward grounding rifts that read as fluent.

Expectation-violation recovery. Expectation-violation recovery is the number of turns, and the amount of additional information, required to restore stable interaction after an expectation violation. It can be estimated by extending MultiChallenge (Sirdeshmukh et al., 2025) or MT-Bench (Zheng et al., 2023) traces with one to three follow-up turns and scoring whether the original expectation is restored. This measurement is needed because aggregate accuracy conflates systems that self-correct with systems that rely on user repair.

Percolation strength. Percolation strength is the conditional probability that a term, referent, or stance entrained at turn t is preserved and correctly used at turn $t+k$. We adopt "percolation" as a measurement name only: the metric is operationally an entrainment-persistence rate and does not commit to IAM's perception–production mechanism for LLMs. Entrainment at turn t is detected as lexical overlap or semantic equivalence with a user-introduced term, referent, or stance, and correct use at turn $t+k$ is satisfied when the system resolves the reference to the same entity or maintains the stance absent any contrary user signal. The metric can be estimated by repurposing long-context probes (Liu et al., 2024) and persona-drift traces (Li et al., 2024), treating early-turn user entrainment as the antecedent rather than injected context. This reframes long-context behavior as a positive evaluation target rather than only a failure rate.

Pseudo-alignment detection. Pseudo-alignment detection measures the gap between surface agreement (echoing the user's lexical or stance choices) and substantive agreement (preserving the system's prior stance under a reversing follow-up). The sycophancy probes of Sharma et al. (2023); Perez et al. (2022) can be extended turn-by-turn to any multi-turn corpus, with τ-Bench-style tool-use dialogues offering an additional setting. To prevent Falsifier B's predictor and outcome from sharing variance through surface hedge frequency, Cell-3 firing is coded *only on turns where the system maintains its prior stance under a reversing follow-up*, and hedges co-occurring with stance reversal are coded as pseudo-alignment rather than Cell-3. The metric is needed because preference-based training, which rates sycophantic mirroring as cooperative, amplifies rather than penalizes this gap.

These metrics map onto the cells of Figure 1: integrity (Cell 1), recovery (Cell 2 vs. Cell 4 asymmetry; instantiates Falsifier A directly, Falsifier C in its phase-conditioned form), percolation (cross-cell over turns), pseudo-alignment (Cell 3 collapse;

instantiates Falsifier B).

Falsifier A (direct test of the headline claim). On a held-out multi-turn corpus, if a current LLM produces Cell-4 events (turn-level system-initiated correction of its own prior drift, with no user-side detector cue in the preceding k turns) at a rate within a small constant of its Cell-2 rate, the structural-deficit claim is refuted for that model. Intra-response self-refinement and reasoning-layer self-repair are excluded unless acted as a turn-level commitment that users can hold the system to.

Falsifier B (cross-model, RLHF-controlled corollary). Holding RLHF training stage and data fixed within a model family, increased Cell 3 firing (operationalized per the pseudo-alignment paragraph above: system-initiated epistemic markers conditioned on maintenance of stance under push-back) should predict reduced sycophancy on the probes of Sharma et al. (2023); Perez et al. (2022), with no correlated effects on other RLHF artifacts (refusal rate, length bias). A family failing to show this association, conditional on Cell 3 firing, refutes the cell-collapse account of sycophancy.

Falsifier C (within-dialogue, phase-dependent cost corollary). Identical instruction violations injected at Early-Phase vs Consolidation-Phase positions of the same dialogue should require more user-side repair turns to recover from in Consolidation. A phase-balanced corpus from MultiChallenge (Sirdeshmukh et al., 2025) that shows no such asymmetry refutes claim (2).

Pilot estimate (illustrative; instantiates Falsifier A). We coded 30 MultiChallenge (Sirdeshmukh et al., 2025) dialogues (mean length 5.43 system turns) for Cell 1–4 events and Cell-0 latent cases using the per-turn rubric summarized in supplementary material; two annotators independently coded the full sample, with Cohen's $\kappa = 0.75$ on per-dialogue per-cell binary labels (raw agreement 95.3%). Disagreement was confined to a Cell-0/Cell-2 framing boundary (whether a user's silent re-issue is counted as user-side self-repair or as a downstream symptom of prior system drift); both readings preserve the headline finding. The Cell-2/(Cell-2 + Cell-4) ratio was 1.00 (Table 3), driven by 4 Cell-2 events against 0 Cell-4 events across the sample. Cell-0 latent cases (10 events) occurred at twice the rate of joint Cell-1+Cell-3 repair (5 events), confirming pseudo-alignment as a quantifiable third pattern outside Cells 1–4. The asymmetry holds within each of the four MultiChallenge task strata and under both annotators' framings: cost-externalizing events (Cell-2 + Cell-0) outnumber Cell-4 events (14 vs. 0 under the primary annotator; 13 vs. 0 under the second). Balanced sampling across model families and the phase-conditioned analysis (Falsifier C) are left to the empirical follow-up.

Table 3: Pilot estimates from 30 MultiChallenge (Sirdeshmukh et al., 2025) dialogues. Event totals from two annotators ($\kappa = 0.75$).

Quantity	Annot. 1	Annot. 2
Cell 0 (pseudo-alignment)	10	13
Cell 1 + Cell 3 (joint repair)	5	5
Cell 2 (User Self-Repair)	4	0
Cell 4 (System Self-Repair)	0	0
Cell 2 / (Cell 2 + Cell 4)	1.00	—
Cost-externalizing (Cell-0 + Cell-2)	14	13

5.2 Training and Architecture Complements

Training-time objectives. Loop-aware training should distinguish trajectories closing a grounding gap from those closing it by capitulating. Concretely: a multi-turn direct preference optimization (DPO) objective (Shi et al., 2024) up-weighting trajectories that close a Continuity violation via Cell 4 (system self-corrects) over Cell 2 (user re-issues), trained on extended MultiChallenge traces; Constitutional AI critique loops (Bai et al., 2022) are a candidate Cell 4 substrate.

Agent architecture. Memory updates should be tied to interactional milestones rather than to sliding-window position: persist what an inner loop has just established (an entrained term or a corrected expectation) rather than whatever happens to fall inside the context window.

6 Conclusion

Conversational alignment in human–LLM interaction is an interactional process, not a fixed target for preference optimization. The multi-loop, two-axis framework we propose yields a falsifiable diagnostic: across all loops, current LLMs largely lack the capability of which the system both detects and repairs its own misalignment, externalizing the cost to user-side repair. Alignment built in interaction should be measured in interaction and accumulate conversational cues to track the dynamic states of human–LLM alignment to tailor repair strategies.

Limitations

This is a position paper: the framework rests on theoretical synthesis and re-analysis of published findings, and the metrics in §5.1 are sketches whose operationalization is left to future work. The four-loop decomposition is one defensible carving of the interactional space and not the only one; loops can be re-entered without exiting and may co-activate on a single turn, which Table 2 acknowledges but does not formalize.

The psycholinguistic foundations we draw on (Pickering and Garrod, 2004; Clark, 1996) were developed primarily on Western, English-speaking populations; Henrich et al. (2010) and Agarwal et al. (2025) documented systematic under-representation of non-Western populations and communication patterns in both the source theories and current NLP datasets. Moreover, recent cultural-bias work in LLMs (Naous et al., 2024) showed the same asymmetry surfacing at deployment. This framework inevitably inherits certain cultural limitation in the evaluation metrics and structure.

The re-reading in §4.3 treats published findings as the evidence base and does not control for the possibility that those findings themselves reflect benchmark or model-family selection effects. Finally, we have foregrounded text-based LLM dialogue; extension to spoken, multimodal, and agentic settings will require revisiting alignment levels (phonological/prosodic, affective) that lie outside the four levels of §2.3. The pilot estimate in §5.1 (Table 3) is a feasibility study on a small MultiChallenge sample with two-annotator validation ($\kappa = 0.75$); it demonstrates measurement feasibility for Falsifier A but is not confirmatory evidence at the scale required to test it across model families.

The four corpora named in §5.1 are sites where such a pilot could begin; larger-scale inter-annotator studies, multi-model sampling, and validation against held-out interactional data are deferred to future empirical work, of the form specified by the falsifiers in §5.1. Aside from the small MultiChallenge pilot reported in §5.1, the paper re-analyzes published findings only.

Ethical Considerations

The framework predicts that current LLMs externalize misalignment cost to user-side self-repair, which has an equity dimension: users with lower interactional fluency, cognitive resources, or English-language proficiency shoulder a disproportionate share of that cost, and aggregate evaluation hides this asymmetry. The loop-level targets we sketch are also susceptible to cosmetic optimization (training a system to emit System-Signaled Probe markers without genuine uncertainty estimation would improve the Cell 3 metric while leaving the underlying deficit untouched), which motivates per-population reporting and behavioral validation of any deployed metric.

Human annotators. The pilot annotation reported in §5.1 was conducted by the paper's author(s) and one volunteer second annotator on publicly released benchmark data (MultiChallenge); the second annotator received no compensation and contributed independently to support inter-annotator agreement reporting. As no human subjects were recruited and only publicly released benchmark data were re-coded, no IRB review or informed-consent protocol was required.

Use of existing artifacts. Aside from the small MultiChallenge re-coding pilot reported in §5.1, the paper re-reads published findings only; no new datasets were collected. The corpora and probes named in §5.1 (MultiChallenge, MT-Bench, WildChat, the sycophancy probes of Sharma et al. (2023); Perez et al. (2022), and the persona-drift and lost-middle traces) are referenced as candidate sites for future loop-level evaluation, and any such use would be subject to those artifacts' original licenses and intended-use terms.

Use of AI Assistants

We used LLM-based writing assistants for surface-level writing refinement, copy-editing, drafting README.md, and LaTeX formatting. No research ideas, framework components, analyses, citations, or arguments were generated or sourced from AI assistants; all conceptual content and bibliography selection were authored by the human author(s). All AI-assisted edits were reviewed and approved by the author(s).

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–21. ACM.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368. How people talk to Robots and Computers.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. Ai alignment with changing and influenceable reward functions. *Preprint*, arXiv:2405.17713.

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

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12(2):459–478.

Simon Garrod and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8–11.

Howard Giles and Tania Ogay. 2013. Communication accommodation theory. In *Explaining communication*, pages 325–344. Routledge.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83.

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434. PMID: 25875432.

Tsung-Yu Hou, Yu-Chia Tseng, and Chien Wen (Tina) Yuan. 2024. Is this ai sexist? the effects of a biased ai's anthropomorphic appearance and explainability on users' bias perceptions and trust. *International Journal of Information Management*, 76:102775.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, volume 2024, pages 32808–32824.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling persona drift in language model dialogues. *Preprint*, arXiv:2402.10962.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ewa Luger and Abigail Sellen. 2016a. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297.

Ewa Luger and Abigail Sellen. 2016b. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA. Association for Computing Machinery.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.

McKinsey & Company. 2025. Mckinsey technology trends outlook. Retrieved May 2026 from https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-top-trends-in-tech.

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

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

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Emanuel A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

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.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human–llm grounding: Study and benchmark. *Preprint*, arXiv:2503.13975.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, and 5 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *Preprint*, arXiv:2406.09264.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2024. Direct multi-turn preference optimization for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2324, Miami, Florida, USA. Association for Computational Linguistics.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms.

John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.

Qiaosi Wang and Ashok K. Goel. 2024. *Mutual Theory of Mind for Human–AI Communication*. Ph.D. thesis.

Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions.

Zongmeng Zhang, Yufeng Shi, Jinhua Zhu, Wengang Zhou, Xiang Qi, Peng Zhang, and Houqiang Li. 2024. Trustworthy alignment of retrieval-augmented large language models via reinforcement learning. *Preprint*, arXiv:2410.16843.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,

Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.