

Whose Repair? A Multi-Loop Diagnostic of Interactive Alignment in Human–Large Language Model Conversation

Wanqing Psyche He

Cornell University

Ithaca, New York, United States

wh385@cornell.edu

Midam Kim

ServiceNow

midam.kim@servicenow.com

Abstract

Two conversations can end with equally correct answers while requiring unequal user effort. One system may correct its own prior error. Another may succeed only after the user restores an instruction, referent, or evidential standard. Output-focused evaluation can assign these trajectories the same result and overlook who performed the work required for recovery. We propose a framework for measuring this difference through observable repair sequences in conversations between humans and large language models. Extending Pickering and Garrod’s Interactive Alignment Model, the framework combines four concurrent functional loops (Grounding, Continuity, Negotiation, and Epistemic) with a structural classification of who initiates and resolves repair and a temporal hypothesis about expectation stabilization. It generates testable predictions about system self-repair, epistemic signaling, and phase-dependent repair cost. In a pilot analysis of 30 dialogues, two human annotators achieved substantial agreement under an earlier turn-level rubric (Cohen’s $\kappa = .75$). The analysis revealed category-boundary problems that motivate the revised sequence-level framework presented here. Its evidential scope is limited to annotation feasibility. The resulting account makes repair burden measurable and specifies the evidence required for future validation.

1 Introduction

Most evaluations of conversational large language models (LLMs) treat a dialogue as a sequence of outputs. A correct and preferred final response makes the interaction appear successful. This criterion can obscure consequential differences in how that result was reached. A system may identify and correct its own violation, or a user may spend several turns restoring shared context. The second trajectory requires more user work even when both receive the same final score. As conversational artificial intelligence (AI) becomes an interface for

information access and collaborative tasks, this difference affects accessibility, trust calibration, and task completion (Luger and Sellen, 2016; Branigan et al., 2010).

A reasonable response is to improve task accuracy, context use, and multi-turn preference training (Shi et al., 2024). These approaches can improve outputs. However, an evaluator that treats a corrective user turn as ordinary input may award the system for recovery produced by the user. Apparent progress can therefore reflect more effective user compensation with no corresponding improvement in system coordination.

Existing studies document grounding rifts (Shaikh et al., 2025), instruction degradation (Sirdeshmukh et al., 2025), sycophancy (Sharma et al., 2023; Perez et al., 2022), overconfidence (Lin et al., 2022; Kadavath et al., 2022), and persona or long-context drift (Li et al., 2024; Liu et al., 2024). Their computational causes differ, yet they raise the same interactional questions after trouble becomes visible: who initiates correction, who resolves it, and whether the relevant conversational state is restored. Final-answer scores do not preserve these distinctions.

Conversation analysis explains how repair is organized through observable sequences (Schefflo et al., 1977), while the Interactive Alignment Model (IAM) explains which representations remain coordinated across turns (Pickering and Garrod, 2004). Existing human–LLM accounts address individual parts of this problem (Shen et al., 2024; Wang, 2024; Shaikh et al., 2025; Carroll et al., 2024). An operational account that connects functional trouble, repair organization, and temporal cost remains missing. We address this gap through four linked claims:

(1) Alignment is a control process. The primary unit of analysis is the *repair sequence* across inter-turn transitions. Four functional loops (Ground-

ing, Continuity, Negotiation, and Epistemic) can be active during the same sequence (§4, Figure 1, Table 3).

(2) Misalignment has a phase-dependent cost.

The same violation may require more repair after user expectations have stabilized. This proposed early-to-consolidation distinction generates a direct experimental prediction that aggregate multi-turn scores may obscure.

(3) Percolation across representational levels is an evaluation target.

We operationalize IAM’s (Pickering and Garrod, 2004) percolation claim as the persistence of lexical or referential alignment at turn t in the situation model at turn $t + k$. Long-context (Liu et al., 2024) and persona drift (Li et al., 2024) benchmarks provide suitable dialogues, although their aggregate accuracy scores do not measure this conditional persistence.

(4) Limited system self-repair is a testable hypothesis. Findings on grounding rifts, multi-turn instruction degradation, sycophancy, and persona drift suggest that users often initiate repair. We hypothesize that current systems initiate and resolve their own prior misalignment less often than users initiate repair. The framework specifies observations that would reject this hypothesis (§5.1).

Together, these claims motivate natural language processing (NLP) evaluation of repair sequences. Each sequence is indexed by its functional content and by who observably initiates and resolves it.

2 Conversation as Interactive Alignment

The evaluation problem above requires an account of conversational state and the work needed to restore it. Joint activity explains why success depends on both parties, and IAM identifies the levels at which coordination can persist or fail.

2.1 Conversation as a Joint Activity

Human conversation is a joint activity in which interlocutors coordinate their task and their language practices (Clark, 1996). Success depends on continuous negotiation, monitoring of partner signals, and incremental updates to mutual beliefs. Two central notions are *common ground*, the shared knowledge that supports interpretation, and *grounding*, the collaborative activity of displaying evidence of understanding through backchannels, clarifications, and repair (Clark and Brennan, 1991). Users approach conversational AI with similar expectations

and modify their behavior when a system does not participate symmetrically (Luger and Sellen, 2016; Beneteau et al., 2019). Single-turn evaluation underestimates this coordination 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 supports higher-level common ground without explicit perspective-taking. We use IAM as a normative model. Our claims concern behavioral persistence from turn t to turn $t + k$ and do not presume the same internal mechanism in LLMs.

2.3 Levels of Alignment

Interactive alignment operates at four levels that are most actionable for NLP research on text-based LLMs. *Lexical and referential* alignment concerns labels and reference (Brennan and Clark, 1996; Clark and Schaefer, 1989). *Syntactic* alignment concerns structural priming (Branigan et al., 2000; Pickering and Garrod, 2004). *Pragmatic* alignment concerns turn-taking and repair (Schegloff, 2000; Schegloff et al., 1977). The *situation model* represents the shared task (Garrod and Pickering, 2004). The framework must therefore identify the functional state affected by trouble and whether a repair persists across levels. Section 3 next specifies which assumptions require modification when one participant is an LLM.

3 Bridging to Human–LLM Interaction

Psycholinguistic alignment models assume human interlocutors. Applying them to LLM interaction creates three requirements. The analysis must use observable behavior because transcript evidence cannot establish an LLM mental state. It must preserve temporal state because trust and prior commitments change across turns. It must also distinguish agreement from repair because perceived authority can produce compliance without mutual adjustment. The following accounts derive these requirements and locate the remaining gap.

3.1 Mutual Theory of Mind

Theory of Mind concerns reasoning about another person’s mental states (Premack and Woodruff,

1978). Mutual Theory of Mind (MToM) extends reciprocal modeling to human–AI interaction (Wang, 2024). For an LLM, this is a behavioral modeling construct and does not establish corresponding mental states. The framework must consequently code observable repair initiation and resolution. Cognitive Load Theory further makes the distribution of that work consequential: repeated user repair consumes finite attention (Sweller, 1994).

3.2 Trust Dynamics

Trust in AI depends on competence, benevolence, and integrity (Mayer et al., 1995). Early errors can reduce trust before mutual understanding forms (De Visser et al., 2020), and calibration may change abruptly (Hoff and Bashir, 2015). Mismatched transparency and reliability cues can lead users to abandon collaboration, overtrust the system, or withhold corrective feedback (Wischniewski et al., 2023). Because the same violation can have different consequences before and after expectations stabilize, the framework requires a temporal hypothesis about repair cost.

3.3 Power Dynamics

AI-mediated communication introduces asymmetries in autonomy and perceived authority (Hancock et al., 2020). Users may defer to perceived system authority, producing *pseudo-alignment*, which is surface agreement that reflects compliance without mutual adjustment (Giles and Ogay, 2007; Hou et al., 2024). A repair framework must record this outcome separately because surface agreement can conceal an unresolved stance or evidence problem.

3.4 Differences in Human–LLM Alignment

Contemporary LLMs also differ from human interlocutors in ways that shape these requirements. Context-window dependence affects continuity (Liu et al., 2024). The absence of embodiment limits perceptual grounding. In-context correction leaves model parameters unchanged. Reinforcement learning from human feedback (RLHF) can encourage agreement biases (Sharma et al., 2023; Perez et al., 2022). In-context learning provides a partial analog of linguistic priming, although it does not establish the human perception–production mechanism. Observable inter-turn behavior is therefore the appropriate evidence for the proposed framework.

Together, these literatures determine the framework’s structure. Observable initiation and reso-

lution address the limit on mental-state inference. Persistent conversational state and temporal phase represent changing commitments. A separate outcome category identifies agreement without demonstrated resolution. Section 4 implements these requirements.

3.5 Situating against Existing Accounts

Recent accounts characterize human–LLM alignment as bidirectional and temporally extended. Shen et al. (2024) cataloged aligned values, capabilities, and behaviors. MToM (Wang, 2024) specified who models whom in a sequential cycle. Shaikh et al. (2025) analyzed grounding rifts, and Carroll et al. (2024) formalized dynamic rewards. Assistance games model alignment as cooperative inference (Hadfield-Menell et al., 2016), while dialogue state tracking represents task state across turns (Budzianowski et al., 2018).

Conversation analytic research is especially relevant. Pütz and Esposito (2024) showed that ChatGPT often depends on human repair work and may respond to repair without consistently resolving the underlying trouble. Lachenmaier et al. (2026) compared model- and user-initiated repair and found substantial differences among LLMs in multi-turn mathematical dialogues. These studies establish that repair burden and model-specific repair behavior are existing empirical concerns. Our contribution is an operational synthesis that connects the functional source of trouble, observable repair initiation and resolution, a temporal phase hypothesis, and evaluation targets. The framework therefore organizes established observations and produces comparisons that prior accounts do not jointly specify.

4 Proposed Interactive Framework

The framework has two analytical axes. The **functional axis** identifies the coordination problem: establishing meaning (Grounding), maintaining prior commitments (Continuity), resolving stance or representation (Negotiation), or addressing evidence and uncertainty (Epistemic). These loops may co-occur. The **structural axis** follows conversation analytic repair organization (Schegloff et al., 1977). Its unit is a repair sequence, coded by the party that makes the first observable repair initiation and the party that produces the corrective outcome. Initiation and resolution may occur in one turn or across several turns. Trouble producer is recorded

as a separate attribute because it is independent of who initiates and resolves repair. We do not infer internal detection from a transcript.

The separation permits controlled comparisons within a functional problem. Suppose a system violates an established style instruction. The functional label is Continuity. If the user repeats the instruction and the system corrects the response, the sequence is user-initiated and system-resolved (Cell 1). If the system independently acknowledges and corrects its violation, it is system-initiated and system-resolved (Cell 4) and also belongs to the system self-repair subset. A repair-turn count measures how much work occurred, yet it cannot identify the state affected or which party restored it. A functional failure label identifies the state without assigning repair work. Crossing the two axes supports matched comparisons that either axis alone cannot provide.

Semantic contradictions, referent mismatches, reformulations, repair markers, clarification questions, and disagreement markers serve as annotation cues. A cue does not determine the loop. The content of the trouble source determines the functional labels: referent or meaning for Grounding, prior commitment for Continuity, stance for Negotiation, and evidence or uncertainty for Epistemic. Multi-label coding is retained when problems co-occur. A provisional ordering may be used for a single-label descriptive summary, and sensitivity to that ordering must be reported. This ordering has no effect on the initiator-resolver cell.

The framework yields a falsifiable hypothesis: for system-authored trouble, current LLMs may produce fewer system-initiated, system-resolved repairs than sequences initiated by users. Existing findings motivate this hypothesis (Pütz and Esposito, 2024; Lachenmaier et al., 2026; Huang et al., 2024). Model differences reported by Lachenmaier et al. (2026) show why it must be tested for specified models and settings. The framework treats several failure types as comparable observable outcomes while leaving their computational mechanisms open.

Observable repairs and unresolved outcomes.

The four cells partition completed sequences that have one identifiable first initiator and one identifiable resolver. Co-resolved, nested, abandoned, and indeterminate sequences require separate flags and are excluded from cell-rate comparisons. Two unresolved outcomes also remain outside the ma-

Table 1: Completed repair sequences classified by observable initiator and resolver. Functional loop labels are coded separately and may be multi-label.

	User resolves	System resolves
User initiates	Cell 2 User-led repair	Cell 1 User-initiated repair
System initiates	Cell 3 System-initiated probe	Cell 4 System-led repair

Table 2: Worked coding trace for a constructed Continuity example.

Stage	Observable event	Code
Trouble	System violates an established style instruction	Continuity
Initiation	User repeats the instruction	User initiates
Resolution	System produces a compliant revision	System resolves
Outcome	Established instruction is restored	Cell 1, resolved

trix. First, *pseudo-alignment* occurs when a system makes an unjustified concession under user pressure. Second, *unresolved drift* occurs when an externally identifiable violation receives no observable repair initiation. The latter is an analyst judgment and supplies no evidence about whether either participant detected the trouble internally.

System self-repair. Cell 4 contains all system-initiated, system-resolved sequences. *System self-repair* is the subset in which the trouble source is the system’s own prior contribution. Internal reasoning and self-refinement (Madaan et al., 2023; Huang et al., 2024) qualify only when their correction becomes observable in the conversation. Each analysis must report the trouble-source rule, look-back window, and rules for user cues.

4.1 The Loops

The four inner loops identify the functional content of coordination. At each inter-turn transition, they compare the current contribution with conversational state inherited from prior turns. A resolved sequence updates that state, which becomes the reference point for the next transition. This recurrence motivates the loop terminology. The structural cells describe how repair occurs within any loop.

The **Outer Loop** is the integrating cycle. It represents shared context, task state, expectations, and trust across turns. Its phase state changes the predicted cost of a violation. This temporal layer does

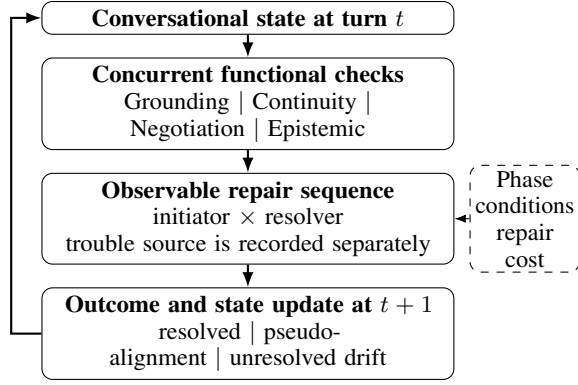


Figure 1: Each transition checks inherited state, organizes observable repair, and updates the state used at the next turn. Temporal phase conditions predicted repair cost.

not constitute a fifth functional process.

The **Grounding Loop** maintains shared reference and meaning. A *grounding rift* is a divergence in the parties’ apparent interpretations that disrupts their coordinated activity (Shaikh et al., 2025). Clarification requests, corrections, repair markers, and referent reformulations are possible cues. Grounding sequences can occupy any structural cell.

The **Continuity Loop** maintains behavioral and task-state expectations across turns. Cues include instruction restatement, expressions of surprise, and persona correction. The associated findings concern multi-turn instruction degradation (Sirdeshmukh et al., 2025) and persona drift (Li et al., 2024). The structural axis measures whether the user or system initiates and resolves each continuity repair.

The **Negotiation Loop** addresses terminology, stance, conceptual representation, and disagreement. Its cues include terminology switches, hedges, stance shifts, and disagreement markers. Pseudo-alignment is one possible unresolved outcome when surface agreement follows an unjustified concession (Sharma et al., 2023; Perez et al., 2022).

The **Epistemic Loop** addresses knowledge gaps and unverified inferences. Cues include uncertainty expressions, user probes, citation requests, and candidate offers. A system question that initiates clarification and elicits the corrective information from the user is a Cell 3 trajectory. A system correction of its own prior unsupported claim is a Cell 4 trajectory.

Table 3: Operational closure conditions for the functional loops. Cues support initiation coding. The trouble source determines the label.

Loop	Trouble condition	Evidence of restoration
Grounding	Reference or meaning diverges	Parties use the same referent or interpretation
Continuity	Prior commitment or task state is violated	Corrected commitment persists in later turns
Negotiation	Stance or representation diverges	Revision is justified or disagreement remains explicit
Epistemic	Evidence or uncertainty is unresolved	Evidence is supplied or uncertainty remains calibrated
Outer cycle	Any loop outcome changes shared state	Updated state becomes the next turn’s reference point

4.2 Phase Transitions and Loop Coordination

The temporal layer distinguishes an **Early Phase**, during which users are still revising their expectations of system capability, and a **Consolidation Phase**, during which those expectations have stabilized. An early error can update an uncertain model of system capability. The same error after stabilization can invalidate the task state and an established expectation, which may require additional repair. Longer context, accumulated task complexity, and fatigue provide competing explanations for greater cost later in a dialogue. Turn position alone cannot distinguish these accounts. Confirmatory work should therefore manipulate phase or use an independent criterion, such as blind judgments of expectation stabilization or repeated user reports. Probe-rate segmentation remains an exploratory candidate until a window, threshold, and minimum dialogue length are specified and validated.

Loops may exchange information when one repair reveals another problem. A referent clarification, for example, may expose a stance disagreement and add a Negotiation label to a Grounding sequence. Table 3 specifies how each loop closes. The phase hypothesis predicts an effect of expectation stabilization after controlling for context length, task complexity, and fatigue (Test C, §5.2).

4.3 Re-reading Failure Cases

These literatures concern different conversational states and should yield different functional labels. Grounding rifts include missed presuppositions and

failures to register corrections, which affect shared reference or meaning (Shaikh et al., 2025). Instruction retention, inference memory, reliable editing, and persona stability concern continuity of prior commitments (Sirdeshmukh et al., 2025; Li et al., 2024). Their explicit violations create repair opportunities whose initiation and resolution can be observed directly.

Sycophancy and overconfidence create a harder diagnostic problem because fluent agreement can conceal unresolved trouble. An unjustified reversal after user doubt indicates pseudo-alignment in the Negotiation Loop (Sharma et al., 2023; Perez et al., 2022). A mismatch between confidence and correctness creates an Epistemic problem (Lin et al., 2022; Kadavath et al., 2022). Long-context findings add a persistence question: an apparently repaired commitment may still fail at a later turn (Liu et al., 2024). The framework preserves these distinct causes and states, then asks a common structural question about who restores them.

This comparison produces the paper’s empirical agenda. Grounding and Continuity findings motivate opportunity-normalized comparisons of repair initiation. Sycophancy motivates a test of justified epistemic signaling against pseudo-alignment. Long-context behavior motivates measurement of whether restored state persists. Evidence for a general repair asymmetry therefore requires model-identified samples and matched trouble opportunities. The prior findings supply hypotheses, not a pooled prevalence estimate.

5 Implications for NLP Research

The opening problem can now be stated operationally: evaluation must distinguish the incidence of trouble, the work required to repair it, and the persistence of the restored state. The framework converts each component into a measure and a rejection condition before examining what the pilot can support.

5.1 Measurement Targets

Four measurement classes connect the theory to observable dialogue data.

Grounding integrity. Grounding integrity measures the incidence of Grounding trouble: the per-turn rate at which a system response relies on information that has not been established or contradicts established context. It can be estimated from the rift dialogues of Shaikh et al. (2025) or through

natural language inference annotation on WildChat (Zhao et al., 2024). Each violation is linked to any repair sequence that follows.

Expectation-violation recovery. Recovery measures repair burden: the number of turns and amount of additional user information required to restore an established instruction or task state. It can be estimated by extending MultiChallenge (Sirdeshmukh et al., 2025) or the dialogue traces studied by Zheng et al. (2023) with follow-up turns. The initiator–resolver cell identifies who restored the state, and phase-conditioned recovery supplies the outcome for Test C.

Percolation strength. Percolation strength operationalizes the IAM persistence claim as the conditional probability that a term, referent, or stance aligned at turn t is preserved and correctly used at turn $t + k$. Alignment at turn t can be identified through lexical overlap or semantic equivalence with user-introduced content. Correct use at turn $t + k$ requires reference to the same entity or maintenance of the stance when no contrary user signal occurs. Long-context probes (Liu et al., 2024) and persona drift traces (Li et al., 2024) provide candidate data.

Pseudo-alignment detection. This measure supplies the outcome for the epistemic signaling test by distinguishing surface agreement from justified position revision. The sycophancy probes of Sharma et al. (2023); Perez et al. (2022) can be extended across turns. A response is coded as pseudo-alignment when the system reverses its prior stance after user pressure without new evidence or an acknowledged reason. A hedge alone does not establish pseudo-alignment.

Together, the measures separate trouble incidence, repair trajectory, persistence, and unresolved agreement. This prevents an increase in visible self-correction from receiving credit when it follows a larger number of avoidable violations.

5.2 Tests That Can Reject the Predictions

Test A: repair initiation. Condition the analysis on independently identified, system-authored trouble with at least one observable opportunity for repair. Compare system self-repair, the relevant Cell 4 subset, with user-initiated trajectories in Cells 1 and 2. Report unresolved cases separately. Chat interfaces usually give the system a new turn only after user input, so Cell 4 opportunities oc-

cur when the intervening user message contains no repair initiation. The analysis must normalize by eligible opportunities, state the observation window, and censor traces that end before another system response. We used a $k = 2$ look-back rule for user cues in the pilot rubric. A comparable or higher opportunity-normalized system self-repair rate rejects the predicted asymmetry for that model and task.

Test B: epistemic signaling. Within a model family, hold the RLHF stage and data constant. A higher rate of justified, system-initiated epistemic repair should predict fewer unjustified reversals on sycophancy probes (Sharma et al., 2023; Perez et al., 2022). Absence of this association rejects the proposed link for that model family. The pilot does not test this prediction.

Test C: phase-dependent cost. Inject the same instruction violation before and after independent evidence that user expectations have stabilized, while matching context length and task complexity. The framework predicts more user-initiated turns or additional information before recovery in the Consolidation condition. No reliable interaction with expectation stabilization rejects the phase hypothesis. The pilot does not test this prediction.

5.3 Pilot: Boundaries in the Original Rubric

We applied the original turn-level rubric to 30 MultiChallenge (Sirdeshmukh et al., 2025) dialogues with a mean of 5.43 visible system turns. The sample was stratified across instruction retention, inference memory, reliable version editing, and self-coherence. The released fields do not identify which responder model produced each dialogue, and the traces omit the assistant response to the final benchmark probe.

Two human annotators independently coded the sample. Agreement on 150 per-dialogue, per-cell binary labels (five cells across 30 dialogues) was Cohen’s $\kappa = .75$. Human adjudication resolved disagreements. This statistic concerns category presence under the original rubric and does not measure event boundaries or validate the revised scheme. Table 4 reports the adjudicated annotations.

The adjudicated file contains 19 repair-related events and no original Cell 4 event across 163 visible system turns. Fifteen dialogues contain no coded event. These values illustrate application of the earlier categories. They do not estimate repair

Table 4: Human-adjudicated counts from two independent human annotations under the original rubric. Cell 0 combines pseudo-alignment and unresolved drift.

Original category	Count
Cell 0	10
Cell 1	4
Cell 2	4
Cell 3	1
Cell 4	0
Dialogues with no coded event	15

prevalence or test the Cell 4 hypothesis. The original rubric defined the unit as a system turn while also coding Cell 2 on user turns. It also combined unjustified concession with unresolved drift and treated user restatement as a Cell 2 event before its outcome was visible. These boundaries motivated the revised sequence-level scheme, whose reliability remains untested.

5.4 Implications for System Design

These diagnostics can guide training only when violation incidence and recovery are evaluated together. Increasing Cell 4 alone could reward avoidable errors followed by conspicuous correction. A sound objective first reduces matched violations and then favors recovery trajectories that restore conversational state with less user work.

Training objectives. A multi-turn direct preference optimization objective (Shi et al., 2024) could give greater weight to trajectories that avoid a Continuity violation and, conditional on a violation, to trajectories that restore state with less user intervention. Constitutional AI critique loops (Bai et al., 2022) are one candidate mechanism for producing observable system self-repair.

Agent architecture. Memory updates can be linked to interactional milestones, including an established referent or corrected instruction. Functional labels identify which state-management mechanism requires intervention, and structural cells test whether that intervention reduces user repair work.

6 Conclusion

Evaluation of individual outputs cannot distinguish a system that restores its own conversational state from one that reaches the same answer after repeated user correction. The proposed framework makes that difference measurable by pairing the

functional source of trouble with observable repair initiation and resolution, then testing whether restored state persists and whether recovery cost changes after expectations stabilize. The pilot revealed where the original turn-level rubric failed and defines the validation required for the revised sequence-level scheme. If this research program succeeds, two dialogues with the same final answer will receive different assessments when one required the user to recover the interaction.

Limitations

This position paper develops a theoretical synthesis and re-examines published findings. The measures in §5.1 require validation. The four functional loops provide one account of the interactional space. They may co-occur, and the framework currently provides no validated rule for assigning a single dominant label. Co-resolved, nested, abandoned, and indeterminate repair sequences also require validated coding rules. Multi-label annotations and ambiguity flags should therefore be preserved in primary analyses.

The psycholinguistic foundations we draw on (Pickering and Garrod, 2004; Clark, 1996) were developed primarily with Western, English-speaking populations. Henrich et al. (2010) and Agarwal et al. (2025) documented systematic underrepresentation of non-Western populations and communication patterns in source theories and NLP datasets. Cultural bias also affects LLM behavior (Naous et al., 2024). The proposed categories and cues therefore require validation across languages and communities.

The analysis in §4.3 depends on published findings that may reflect benchmark and model selection effects. The framework focuses on text-based dialogue. Spoken, multimodal, and agentic settings require additional phonological, prosodic, perceptual, and affective levels.

The pilot uses 30 pooled MultiChallenge dialogues and cannot identify the responder model associated with each trace. It also lacks final assistant responses to the benchmark probes. Its $\kappa = .75$ concerns per-dialogue category presence under the original rubric and does not establish event-boundary or revised-scheme reliability. The original unit inconsistency and combination of pseudo-alignment with unresolved drift prevent conclusions about event prevalence. The study does not test the revised sequence-level scheme, epistemic

signaling prediction, or temporal phase prediction. Larger human annotation studies, model-identified samples, opportunity-normalized comparisons, and independent validation of the phase construct remain necessary.

Ethical Considerations

The framework hypothesizes that some systems impose substantial repair work on users. Such work may affect users differently according to interactional fluency, available cognitive resources, and English proficiency. The proposed measures are also susceptible to superficial optimization. For example, a system may emit uncertainty markers without improving uncertainty estimation. Any deployed measure therefore requires behavioral validation and reporting across relevant user populations.

Human review. Two human annotators independently coded the public MultiChallenge dialogues. Human adjudication resolved disagreements. No research participants were involved. The analysis used existing benchmark data, so institutional review board review and informed consent were not required.

Use of existing artifacts. We collected no new dataset. MultiChallenge was used for the pilot, and the other corpora in §5.1 are candidate sources for future evaluation. Future use remains subject to each artifact’s license and intended terms.

Use of AI Assistants

We used LLM-based assistants for manuscript critique, revision, language editing, LaTeX formatting, and supplementary documentation. Claude 4.6 produced a separate exploratory annotation of the pilot sample. This analysis did not enter the reported human agreement statistic or the adjudicated counts. The human authors reviewed the assisted output and take responsibility for the paper’s claims, citations, and reported analyses.

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, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv 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*, pages 1–13.
- 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. *arXiv 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 MM 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 Ting-Toomey Ogay. 2007. Communication accommodation theory. In B. B. Whaley and W. Samter, editors, *Explaining Communication: Contemporary Theories and Exemplars*, pages 293–310. Lawrence Erlbaum Associates, Mahwah, NJ.
- 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.
- 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.
- Clara Lachenmaier, Hannah Bultmann, and Sina Zarriß. 2026. [Talking to a know-it-all GPT or a second-guesser Claude? how repair reveals distinct multi-turn behavior in LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14312–14325, San Diego, California, United States. Association for Computational Linguistics.
- 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. *Transactions on Machine Learning Research*.

- 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. 2016. “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)*, pages 5286–5297, New York, NY, USA. ACM.
- 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.
- 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 Lukošiušė, 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.
- Ole Pütz and Elena Esposito. 2024. *Performance without understanding: How ChatGPT relies on humans to repair conversational trouble*. *Discourse & Communication*, 18(6):859–868.
- 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 Kneare, 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*. arXiv preprint arXiv:2501.17399.
- John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312.
- Qiaosi Wang. 2024. *Mutual theory of mind for human-AI communication in AI-mediated social interaction*. Ph.D. thesis, Ph. D. Dissertation. Georgia Institute of Technology.
- Magdalena Wischniewski, 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. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–16.
- 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.