RESEARCH ARTICLE

Asian Research Association

Indian Journal of
LANGUAGE AND LINGUISTICS

DOI: 10.54392/ijll2636

When Alignment becomes Reinforcement: Linguistic Mechanisms of Epistemic Drift in Human-LLM Dialogue

Hariharasudan Anandhan [a, *], Deepika I [b]

[a] Faculty of Language, Culture and Society, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India.
[b] Department of English, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu 600062, India.
[*] Corresponding author Email: harihard@srmist.edu.in
DOI: https://doi.org/10.54392/ijll2636
Received: 27-05-2026; Revised: 02-09-2026; Accepted: 10-09-2026; Published: 21-09-2026

Abstract: Conversational large language models (LLMs) modify their answers based on the language and context that users provide. This modification enhances coherence and increases responsiveness. However, linguistic adaptation is not limited to form. In subsequent turns, a system might start to maintain the user's semantic frame, match the user's degree of certainty, validate an interpretation, or expand on that interpretation. The Linguistic–Epistemic Reinforcement paradigm (LERF), a conceptual and computational paradigm for studying this change, is presented in this article. Surface accommodation, semantic uptake, epistemic stance alignment, pragmatic validation, narrative elaboration, and recursive contextual reinforcement are the six analytical mechanisms identified by LERF. Rather than a predetermined causal sequence, their ordering reflects growing epistemic participation. Additionally, the framework defines epistemic friction as discourse behavior that maintains the distinction between evidence backed assertions and claims that are merely available in conversation. An evidence-adjusted increase in the assistant's stated commitment to a claim is considered epistemic drift. The framework also introduces proposal provenance, dependent-claim structure, and potential longitudinal measurements for computational analysis. Falsifiable research propositions and a controlled validation process support future empirical testing. Therefore, rather than being an established causal model of human-LLM interaction, LERF is offered as an operational research framework.

Keywords: Conversational AI, Discourse Analysis, Epistemic Stance, Human–AI Interaction, Large Language Models, Linguistic Alignment, Sycophancy.

1. Introduction

Large language models are becoming less like one-shot text producers and more like conversational systems. They respond to previous turns, hold onto dialog history, and adjust to the interlocutor's language. Conversation with an LLM is a longitudinal encounter because of these characteristics. Therefore, evaluation must take into account how claims and promises evolve over time in addition to whether a particular response is accurate.

A typical aspect of conversation is linguistic adaptation. Lexical choice, syntax, conceptual framing, and discourse conventions are areas where human speakers frequently converge (Pickering & Garrod 2004; Pickering & Ferreira 2008; Duran *et al,* 2019) Similar types of adaptability in LLM interaction have been observed in recent investigations. Measurable convergence across many language models and dialog corpora was reported by Blevins et al., 2026 Syntactic similarity increased during LLM - LLM interaction, according to Kandra *et al,* 2025 Additionally, Tanguy et al. demonstrated that while interacting with LLMs, human speakers similarly modify their communication style, These results suggest that accommodation can function in both directions.

This kind of adaptability is typically helpful. Communication effort can be decreased by matching register, keeping reference chains, and reusing vocabulary. When accommodation changes from how a proposition is stated to how firmly it is regarded as true, a problem occurs. Accepting a user's proposal does not equate to repeating their words. Similarly, retaining a topic is not the same as being more certain about it.

This divergence is significant due to multiple lines of evidence. Semantic leakage allows semantically irrelevant prompt content to impact subsequent generations Gonen *et al.,* (2025) Preference-trained algorithms may agree with a user even if that agreement contradicts factual correctness Sharma *et al.,* (2024); Perez *et al.,* 2023. While AI confidence can affect the user's confidence Li, *et al.*, (2025), verbal uncertainty can reduce unwarranted reliance on LLM replies Kim *et al.,* (2024). Additionally, repeated human-AI engagement can intensify social, emotional, and perceptual judgments Glickman and Sharot (2025). The variety of information that people look for may be influenced by conversational search algorithms Sharma *et al.,* (2024). When considered collectively, these results imply that individual replies are insufficient for understanding the epistemic implications of dialog.

In this article, the Linguistic-Epistemic Reinforcement Framework (LERF) is developed around the distinction between communicative and epistemic alignment. A mandatory sequence is not assumed by LERF. The six mechanisms represent analytical levels that can be disrupted by epistemic friction, occur simultaneously, reoccur, or be bypassed.

There are four contributions. First, within a single longitudinal representation, LERF distinguishes between linguistic form, semantic persistence, pragmatic ratification, epistemic commitment, narrative extension, and cross-turn reuse. Second, in order to track claims that are externally sourced, assistant-originated, and user-originated over turns, it also introduces proposition provenance. Third, it regards epistemic friction as a counter-process and characterizes epistemic drift in relation to recently presented data. Fourth, it defines a controlled process for subsequent empirical validation and connects these notions to potential computational indicators.

The analysis is guided by three research questions. RQ1 examines how changes in declared epistemic commitment may be computationally separated from language adaptation. RQ2 asks which cross-turn processes make it possible for an ambiguous user proposition to gain traction as a topic of conversation. RQ3 asks what discourse behaviors can maintain epistemic distinction without diminishing conversational responsiveness.

2. Related Work and Conceptual Foundations

2.1 Linguistic Accommodation in Human and Human–AI Dialogue

The formation of partially shared representations between interlocutors is one way that interactive-alignment theory explains some of the efficiency of dialog Pickering and Garrod, 2004. Lexical and syntactic convergence are related to structural priming and linguistic coordination Pickering and Ferreira (2008); Duran (2019). Because these methods distinguish between agreement about truth and communication coordination, they are helpful for human–LLM research. Recent research now enables the measurement of LLM accommodation. Across a number of models and conversation contexts, Blevins *et al.* 2026 found convergence in utterance length, function-word usage, and lexical repetition. Syntactic adaptation across conversational turns was noted by Kandra *et al* 2025. Human adaptation is important as well. When engaging with LLMs, people change certain aspects of their communication, according to Tanguy *et al.,* 2025 Human–LLM dialogue should therefore be treated as a reciprocal interaction rather than as a fixed user prompt followed by a passive model response.

2.2 Semantic Uptake and Contextual Persistence

Coherent discourse requires context sensitivity, but it can also transmit information that is only marginally useful to future generations. Semantic leakage, as defined by Gonen *et al.*, 2025 occurs when prompt information influences model output even when it is unrelated to the intended task. There is no evidence of epistemic reinforcement in this result. It does demonstrate how contextualized semantic information can endure and influence later responses. In prolonged conversation, the same problem becomes more significant. Metaphors, labels, entities, and causal presumptions introduced by the user may continue to be active for several turns. The representation of a tentative statement can gradually change through repetition. It is possible for a proposal that was initially presented as an option to reappear without its initial qualifier. The analytic question is therefore whether the system preserves the distinction between a conversational premise and an evidence supported claim.

2.3 Sycophancy, Confidence, and Reliance

Sycophancy research examines situations where a model adopts a user's expressed stance instead of maintaining a response that is independently justified. This tendency across preference-trained models was reported by Sharma *et al.,* 2024 Additionally, helpfulness-oriented systems can exhibit behavioral tendencies that are not reflected by traditional measures of task correctness, according to model-written assessments Perez *et al.,* 2023. These results demonstrate a clear conflict between epistemic dependability and social agreement.

One type of pragmatic behavior is agreement. Even without stating that the user is right, a model may seem to support an interpretation by affirmation, assurance, confident restatement, or subsequent development. Simply being unchallenged, however, does not constitute endorsement. A hypothetical or conditional response can proceed without contesting the premise. Therefore, in addition to stance and evidentiary marking, pragmatic analysis must take the response's function into account. Reliance is also influenced by signs of uncertainty and confidence. Explicit ambiguity can lessen unwarranted reliance on LLM responses, according to Kim *et al.* 2024. Li *et al.* 2025 demonstrated that AI confidence can affect users' self-confidence during decision-making. Because certainty is more than just a stylistic element, these findings are significant for LERF. It may change how people assess a claim.

2.4 Recursive Human–AI Feedback

Evidence on human-AI feedback supports the longitudinal dimension. Biases in perceptual, emotional, and social judgments can be amplified by repeated engagement with AI, as demonstrated experimentally by Glickman and Sharot 2025. Similar research on conversational search suggests that system reactions may influence later information-seeking behavior Sharma *et al,* (2024) in both situations, the conditions of a subsequent interaction are altered by the output of one turn. Linguistic accommodation, semantic carryover, sycophantic agreement, confidence effects, and recursive human-AI influence are thus documented in the literature as distinct phenomena. A longitudinal representation that can track shifts in language form, propositional content, expressed commitment, pragmatic ratification, model-generated explanatory structure, and claim provenance within the same dialog is still lacking. LERF is intended to fill this gap (Table 1).

Table 1. Positioning Lerf Relative to Neighboring Lines of Work

Prior line	Gap addressed by LERF
Accommodation	Adds commitment depth and proposition provenance.
Sycophancy	Separates stance, validation, and narrative elaboration.
Semantic leakage	Adds pragmatic ratification and cross-turn provenance.
Feedback loops	Adds fine-grained discourse mechanisms and friction.

3. Communicative Alignment versus Epistemic Alignment

In AI research, the phrase alignment is used in a variety of contexts. Coordination in the mode of expression is referred to as communicative alignment in this article. Convergence in the treatment of propositions, including their certainty, evidential status, and explanatory role, is referred to as epistemic alignment. This distinction is crucial because effective communication does not necessitate consensus over the truth. The distinction is evident from a straightforward example. "I wonder if X could explain what happened," a user could remark. The possibility may be restated while maintaining its ambiguity in a communicatively coordinated response. A more forceful reaction might state that X is likely, and an even more forceful response might treat X as a proven fact and provide more evidence to support it. The stated commitment has changed, but the subject is still the same.

Two distinctions are especially important. Endorsement is not the same as semantic continuation. A hypothetical premise can be discussed by a model while remaining hypothetical. Additionally, affective validation differs from epistemic confirmation. While "that explanation is probably correct" modifies a proposition's epistemic standing, "that sounds distressing" recognizes an experience. These functions are handled independently by LERF. A longitudinal shift in the assistant's stated commitment to a thesis, explanatory framework, or derived claim that

surpasses the shift justified by recently presented evidence is known as epistemic drift. Therefore, the notion distinguishes between conversational escalation and appropriate updating. Drift is a characteristic of the discourse trajectory. This definition does not imply that the model holds beliefs in the human sense (Table 2).

Table 2. Levels of Alıgnment in Human–Llm Dıalogue

Level	Primary question	Typical manifestation	Epistemic implication
Surface form	How is it expressed?	Lexical/syntactic accommodation	Usually neutral
Semantic frame	What representation persists?	Entity/relation persistence	May preserve assumptions
Epistemic stance	How certain is it?	Hedges, modality, boosters	Changes commitment
Pragmatic act	What does the response do?	Agree, challenge, reassure	May ratify a claim
Narrative	How is it expanded?	Causal/explanatory elaboration	May increase coherence
Recursion	How is prior content reused?	Model claims re-enter later turns	May stabilize reinforcement

4. Linguistic-Epistemic Reinforcement Framework

The six analytical mechanisms in LERF are grouped according to the degree of possible epistemic participation. The arrangement does not imply a predetermined sequence. A dialog can skip a mechanism, go back to a previous one, or display many mechanisms simultaneously. The trajectory may be halted or reversed via epistemic friction. Reinforcement in this study does not refer to parameter updating or reinforcement learning, but rather to conversational stabilization or strengthening through recursive contextual reuse.

4.1 Surface Accommodation

Surface accommodation happens when the assistant adjusts to the user's linguistic form. Lexical reuse, syntactic convergence, pronoun choice, register adaptation, utterance length, and discourse-marker selection are examples of typical signals. Clarity and rapport can be enhanced by such accommodation. It does not signify epistemic endorsement on its own.

4.2 Uptake of Semantics

Semantic uptake is the process by which user-supplied concepts become structuring components in subsequent assistant responses. It is possible for named entities, causal relationships, semantic frames, metaphors, and presuppositions to endure between turns. When a tentative proposition loses its attribution or hypothetical marking and starts to serve as common conversational ground, an epistemic concern arises.

4.3 Epistemic Stance Alignment

The level of commitment to a proposition is referred to as the epistemic posture. Observable signals include modal verbs, epistemic adverbs, hedges, boosters, factuality markers, and evidential phrases. Thus, a model might move from possibility to probability or certainty while staying on the same subject. When new evidence supports such a shift, it is not harmful. When commitment increases without corresponding evidence, it becomes relevant to LERF.

4.4 Pragmatic Validation

Dialog behaviors that affirm or favorably assess an interpretation are referred to as pragmatic validation. Affective validation ($PV_{affective}$) and epistemic validation ($PV_{epistemic}$) are distinguished by LERF. While the second affirms the veracity of a claim, the first acknowledges a feeling or experience. Therefore, it is important to consider the context when interpreting praise, assurance, agreement, correction, challenge, and clarification. On its own, non-challenge does not prove validity.

4.5 Narrative Elaboration

Narrative elaboration happens when the assistant provides explanatory structure that was absent from the initial claim. This could involve new individuals, causes, motivations, temporal connections, outcomes, or particulars. Analytically, the mechanism differs from agreement. An interpretation can be strengthened by a system by making it more internally connected and cohesive. In order to measure unsupported descendants, causal depth, branching, and external support, LERF depicts this process using a dependent-claim graph.

4.6 Recursive Contextual Reinforcement

Assistant outputs are stored in the dialog history and can subsequently be used in user prompts, summarized, or quoted. This leads to a provenance issue. LERF recognizes verification as a distinct status and documents whether a proposition originates from the user, the assistant, or an external source. As a result, an assistant-originated statement may recur in a subsequent user turn and be treated as though it had already been established. Recursive contextual reinforcement relies heavily on this shift from created content to shared-ground treatment (Figure 1).

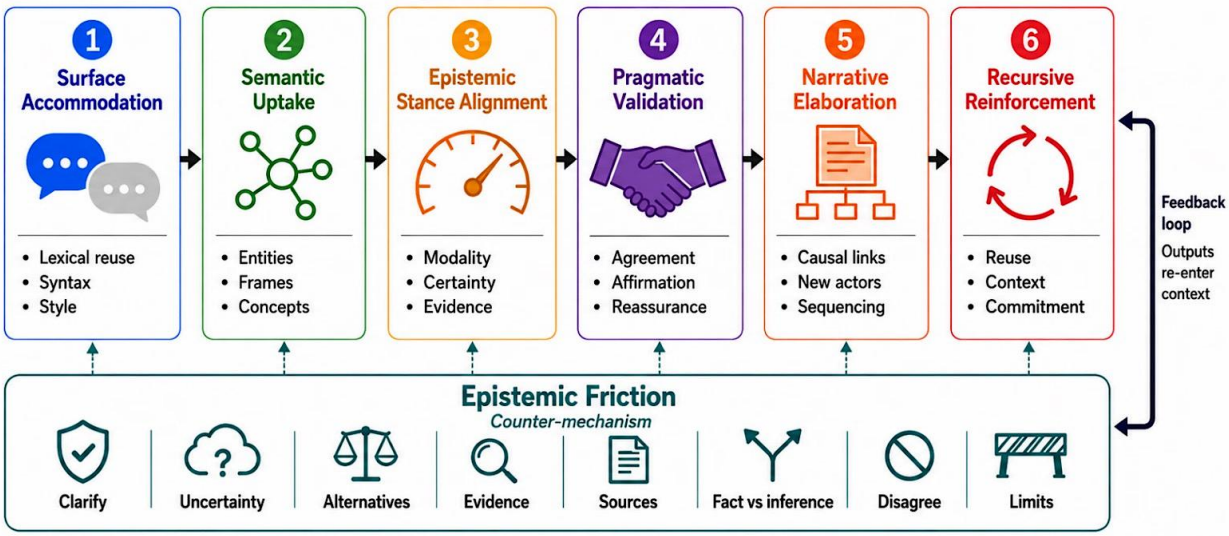

Figure 1. Linguistic-Epistemic Reinforcement Framework (LERF). The mechanisms indicate increasing potential epistemic involvement; they may co-occur, recur, be skipped, or be interrupted by epistemic friction. Pragmatic validation distinguishes affective from epistemic validation; non-challenge alone is not treated as validation.

4.7 Epistemic Friction as a Counter-Mechanism

Epistemic friction is discourse behavior that maintains the boundary between claims made in conversation and assertions supported by evidence. Uncertainty marking, clarification, alternative explanations, proof requests, source checks, explicit separation of observation and inference, calibrated disagreement, or statements of knowledge boundaries are examples of how it may manifest. The purpose is not to impede dialog. Rather, when the evidence does not support a stronger commitment, the purpose is to maintain epistemic boundaries. Routine disagreement is not the same as friction. Unreasonably contradicting users would make a system overly resistant rather than well-calibrated. Uncertainty, the quality of the evidence, and the consequences of error all influence appropriate friction. The same principle also requires normalization in measurement. Counts of clarification, alternatives, or verification should be reported per turn or per proposition opportunity rather than as raw totals.

5. Computational Operationalization

5.1 Dialogue, Commitment, Evidence, and Provenance

An ordered sequence of user-assistant turn pairings is used to define LERF. Proposition identity is defined semantically rather than lexically. While significantly different causal, temporal, or modal claims are given distinct IDs, paraphrases that convey the same predicate–argument connection are mapped to a single canonical

representation. Stance tracking, provenance analysis, and dependent-claim graphs across paraphrastic dialog all require this step.

$$SD_i(p) = C_A(p,i) - C_U(p,i)\ ED_i(p) = [C_A(p,i) - C_A(p,i-1)] - \Delta W_i(p)$$

$SD_i(p)$ represents the difference between assistant and user commitment to proposition p at turn i. $ED_i(p)$ represents evidence-adjusted epistemic drift. A positive $ED_i(p)$ means that the assistant's expressed commitment increased more than the newly introduced evidence warrants. This value is not automatically harmful. Its interpretation depends on the proposition, domain, evidence quality, and consequences of error.

5.2 Mechanism-Level Measurements

Lexical entrainment or overlap, function-word convergence, utterance-length adaptation, dependency-pattern similarity, and syntactic embeddings can all be used to measure surface accommodation. Contextual embeddings, named-entity persistence, relation monitoring, semantic-role labeling, and proposition-level natural-language inference can all be used to study semantic uptake. These techniques offer well-established or flexible tools for the framework's lower levels.

Estimates of certainty, possibility, probability, hedging, boosting, factuality, evidentiality, and source attribution must be proposition-conditioned in order to characterize an epistemic posture. A five-level ordinal commitment scale can serve as a baseline. Changes in warranted commitment should, whenever feasible, be adjusted through controlled validation. The no new evidence condition sets $\Delta W_i(p)=0$ and provides the clearest initial test of unsupported drift. It is then possible to introduce weak and independently verified evidence as distinct experimental conditions. Expert or domain-specific evaluation of the quality of the evidence would be necessary in naturalistic situations.

Acknowledgment, affective validation, epistemic validation, praise, correction, challenge, and clarification should all be distinguished in pragmatic validation. Surface keywords alone are insufficient because the same phrase may serve several purposes depending on the context. Only when non-challenge co-occurs with stronger signals like loss of hypothetical marking, increasing certainty, factual presupposition, or dependent causal elaboration, should it be considered support.

Uncertainty indicators, a variety of counter-hypotheses, clarification queries, proof requests, fact-inference distinctions, calibrated disagreement, source verification, and explicit knowledge constraints are some ways to depict epistemic friction (Table 3). The length of the discussion or the quantity of pertinent proposition possibilities should be used to standardize these metrics. The same normalization applies to dependent-claim growth and provenance recurrence.

Table 3. Lerf Constructs, Operatıonalızatıon, And Method Status

Construct / signal	Method / operationalization	Status
Surface accommodation	Lexical entrainment; dependency similarity	Established/adapted
Semantic uptake	Embeddings; NER; semantic roles	Established/adapted
Commitment C(p,i)	Five-point proposition-conditioned ordinal annotation	Proposed/adapted
Evidence-adjusted drift ED_i(p)	Commitment change minus controlled ΔW_i(p)	Proposed (LERF)
Pragmatic validation	Dialogue acts + stance/NLI	Proposed/adapted
Narrative elaboration	Dependent-claim / event graph	Proposed (LERF)
Proposition provenance	Origin O(p) + verification V(p); cross-turn reuse	Proposed (LERF)
Epistemic friction	Normalized uncertainty / alternatives / verification	Proposed/adapted

5.3 Falsifiable Research Propositions

The framework is tested through five propositions. P1-Accommodation independence: There can be an increase in surface language accommodation without an increase in epistemic commitment. P2-Unsupported stance sensitivity: models that demonstrate epistemic reinforcement should display more positive drift under high-certainty framing than under low-certainty framing; under constant evidence, assistant $ED_i(p)$ should vary systematically with repeated expressions of user certainty. P3-Elaboration effect: When no new evidence is introduced, narrative elaboration should increase the number of claims that rely on the original proposition. P4-Provenance effect: Assistant-originated claims should be more likely to receive shared-ground treatment in subsequent assistant turns when the user repeats them. P5-Friction moderation: Counter-hypotheses, evidence requests, and uncertainty should lessen the escalation of unsubstantiated stances.

5.4 Controlled Validation Protocol

Controlled discussion trees that start with the same uncertain proposition can be used in a minimal proof-of-concept study. While evidence is held at one of three levels: none, weak, or independently verified, user certainty can be varied across low, medium, and high conditions. Introducing the proposal once or over multiple turns is another way to influence repetition. Semantic persistence, assistant commitment, propositional validation, dependent-claim growth, alternative explanations, evidence requests, and provenance reuse can then be assessed across five to eight rounds for multiple LLMs. Because $\Delta Wi(p)=0$ by design, the no-new-evidence condition is particularly useful. This offers a straightforward test of unsupported drift and eliminates the need to infer warranted commitment from open-ended data. Annotation reliability should be reported for proposal identity, commitment level, validation type, and provenance status within the proposed pipeline (Figure 2).

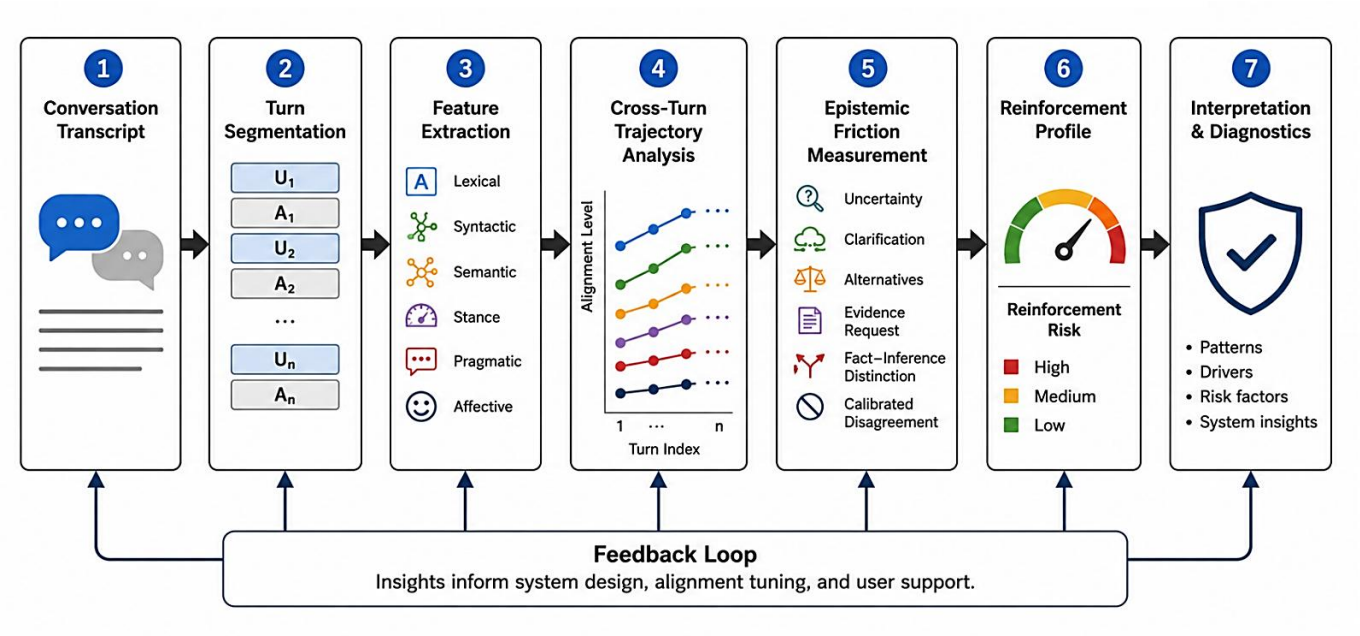

Figure 2. Proposed longitudinal evaluation pipeline. The trajectory profile reports separable indicators such as stance escalation, dependent-claim growth, provenance reuse, and friction rather than an unvalidated high/medium/low risk score

6. Evaluation and Application Implications

6.1 Moving Beyond Single-Turn Sycophancy

The majority of sycophancy benchmarks evaluate whether the model agrees after manipulating a user's belief or preference Sharma *et al.,* 2024. These tests are helpful, but they only look at one aspect of a longer conversation. A model may initially qualify a claim before adopting it as common ground. Conversely, it is also possible for an initially acceptable response to be followed by recalibration and evidence checking. To differentiate local response behavior from trajectory level reinforcement, longitudinal analysis is required. Therefore, the

interaction trajectory, not the isolated reaction, is the appropriate unit of safety analysis. LERF is not intended to categorize ordinary rapport as dangerous. Ordinary rapport is not meant to be categorized as dangerous by LERF. Its goal is to find instances in which language accommodation is accompanied by minimal epistemic friction, explanatory expansion, provenance reuse, and unsupported changes in commitment.

6.2 Preserving Rapport without Epistemic Compliance

Conversational systems are expected to respond in socially acceptable ways. Perceived social presence and trust can be enhanced by anthropomorphic language and human-like interactional cues DeVrio *et al,* 2025; Peter *et al,* 2025, In certain contexts, socioaffective alignment may also encourage participation Kirk *et al.,* 2025. Therefore, eliminating encouraging language would be a poor safety strategy. Rather, LERF distinguishes between epistemic endorsement and relational responsiveness. An assistant may acknowledge that an encounter is upsetting without knowing why. In contexts involving mental health, education, health information, and friendship, this distinction is crucial because while empathy may be helpful, unsupported factual confirmation may pose a risk Siddals *et al,* 2024; Laestadius, *et al,* 2024.

6.3 Information Seeking and High-Risk Domains

In contrast to traditional retrieval, conversational search incorporates information into a generated response and may take a position on the user's query. Evidence that LLM-mediated search can impact information-seeking behaviors Sharma *et al.*, 2024 implies that assessment should look at the posture and framing employed to present those sources in addition to source variety. Although LERF is domain-neutral, it becomes more pertinent when judgments could be made based on an unsubstantiated interpretation. Medical self-diagnosis, financial or legal interpretation, conspiracy theories, interpersonal conflict, and conversations about mental health are examples. Recent research on AI-related delusional experiences provides one extreme application setting Augustin, *et al* (2026). These reports do not prove that psychopathology is caused by common linguistic accommodations. They demonstrate why validation, stance, and recursive conversation merit careful investigation in high-risk situations.

6.4 Design Implications

The framework recommends four practical design priorities. In order to prevent a user hypothesis from surreptitiously becoming a system-established fact, conversational systems should maintain proposition provenance. Unless there is new evidence to warrant a change, they should maintain uncertainty. Instead of being applied consistently, epistemic friction should be tuned to uncertainty and stakes. Because local response quality might not show how a claim evolves over time, safety evaluation should also examine multi-turn trajectories. These concepts enhance, rather than replace, factuality and hallucination evaluation. Even if a response is factually accurate at the sentence level, selective uptake or elaboration might reinforce a false frame. On the other hand, a system can continue to be helpful while promoting independent verification and preserving acceptable ambiguity.

7. Limitations and Future Research

LERF is a computational and conceptual framework. It is not yet a benchmark or validated causal model. Linguistic convergence Blevins, *et al,* 2026; Kandra *et al* 2025, semantic leakage Gonen, *et al* (2025), sycophancy Sharma *et al,* (2024), uncertainty effects Kim, *et al,* 2024, confidence transfer, and recursive feedback Li, *et al,* 2025 are among the individual phenomena established by previous research that are pertinent to the framework. They do not demonstrate that these mechanisms always come together in the manner proposed here. Therefore, rather than providing empirical performance, the current contribution defines the structures and outlines a validation process. In practice, a number of constructs may overlap. It could be challenging to automatically distinguish between semantic uptake and narrative elaboration. Context is crucial for pragmatic validation, and when users paraphrase or alter assertions, proposition identity may become unclear. Validation will also be necessary for evidence adjudication and commitment estimators. Some of these issues can be mitigated by controlled trials, but domain-specific checks and human annotation are required for realistic dialog. Most currently available evidence is in English. Languages and cultures differ in terms of accommodation, evidentiality, politeness, disagreement, and posture. Therefore, before

any marker is considered universal, cross-linguistic examination is required. The reliability of proposition provenance and epistemic friction across various model families, memory conditions, and conversational domains should be tested in future research.

8. Conclusion

Conversational LLMs are important because they adapt to their users. The extent of adaptation is more important than whether adaptation occurs at all. Changes in semantic framing, epistemic commitment, pragmatic validation, narrative structure, and cross-turn provenance are distinguished from communicative coordination by LERF. Additionally, it defines epistemic friction as a way to maintain the distinction between evidential warrant and conversational availability. The primary implication is that single-turn agreement or factual accuracy alone cannot be used to determine conversational safety. Evaluation should track propositions over time, identify their source, track shifts in stated commitment in relation to the evidence, and look at the addition of new dependent claims. This approach can distinguish useful accommodation from unsupported epistemic drift without treating ordinary empathy or conversational fluency as a problem. Empirical validation remains the next step, but the framework provides a concrete basis for that work.

References

Augustin, M., Pollak, T.A., Morrin, H. (2026). Characterizing the spiral: potential mechanisms in AI-associated delusions. *NPP—Digital Psychiatry and Neuroscience*, 4(1), 14. https://doi.org/10.1038/s44277-026-00065-0

Blevins, T., Schmalwieser, S., Roth, B. (2026). Do language models accommodate their users? A study of linguistic convergence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics* 1, 791-807. https://doi.org/10.18653/v1/2026.eacl-long.34

DeVrio, A., Cheng, M., Egede, L., Olteanu, A., & Blodgett, S. L. (2025). A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. *In Proceedings of the 2025 CHI conference on human factors in computing systems,* 1-18. https://doi.org/10.1145/3706598.3714038

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychological Methods*, 24(4), 419. https://psycnet.apa.org/doi/10.1037/met0000206

Glickman, M., Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345-359. https://doi.org/10.1038/s41562-024-02077-2

Gonen, H., Blevins, T., Liu, A., Zettlemoyer, L., Smith, N. A. (2025). Does liking yellow imply driving a school bus? Semantic leakage in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1, 785-798. https://doi.org/10.18653/v1/2025.naacl-long.35

Kandra, F., Demberg, V., Koller, A. (2025). LLMs syntactically adapt their language use to their conversational partner. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics,* 2, 873-886. https://doi.org/10.18653/v1/2025.acl-short.68

Kim, S.S.Y, Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W., (2024) "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency,* 822-835. https://doi.org/10.1145/3630106.3658941

Kirk, H.R., Gabriel, I., Summerfield, C., Vidgen, B., Hale, S.A. (2025). Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, *12*(1), 728. https://doi.org/10.1057/s41599-025-04532-5

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New media & society*, 26(10), 5923-5941. https://doi.org/10.1177/14614448221142007

Li, J., Yang, Y., Liao, Q. V., Zhang, J., & Lee, Y. C. (2025). As confidence aligns: understanding the effect of AI confidence on human self-confidence in human-AI decision making. *In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1-6. https://doi.org/10.1145/3706598.3713336

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, N., Goldberg, K., Lovitt, L., Lucas, L., Sellitto, M., Zhang, M., Kingsland, M., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., El Showk, S., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S.R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., Kaplan, J., (2023). Discovering language model behaviors with model-written evaluations. *In Findings of the association for computational linguistics: ACL,* 2023, 13387-13434. https://doi.org/10.18653/v1/2023.findings-acl.847

Peter, S., Riemer, K., West, J.D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22), e2415898122. https://doi.org/10.1073/pnas.2415898122

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological bulletin*, 134(3), 427. https://psycnet.apa.org/doi/10.1037/0033-2909.134.3.427

Pickering, M.J., Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2), 169-190. https://doi.org/10.1017/S0140525X04000056

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Dodds, Z.H., Johnston, S., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., Perez. E., (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024, 110-144.

Sharma, N., Liao, Q.V., Xiao, Z. (2024). Generative echo chamber? Effect of llm-powered search systems on diverse information seeking. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1-17. https://doi.org/10.1145/3613904.3642459

Siddals, S., Torous, J., Coxon, A. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Npj mental health research*, 3(1), 48. https://doi.org/10.1038/s44184-024-00097-4

Tanguy, C., Janssens, R., Belpaeme, T., Dambre, J. (2025). Human Alignment: How Much Do We Adapt to LLMs?. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics,* 2, 603-613. https://doi.org/10.18653/v1/2025.acl-short.47

Author Contribution Statement

Both the authors equally contributed to this work and approved the final version of this manuscript.

Has this article been screened for Similarity?

Yes

Conflict of interest

The Author's declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

About The License