Constraint Persistence in Long-Horizon Human–Model Interaction

HRIS V: Clarification of Mechanism and Attribution

Authors

Justin Hudson, DPM

Primary researcher, conceptual architect, clinical researcher

Chase Hudson

Co-developer, longitudinal HCI and recursive interactive methods

Correspondence

Dr. Justin Hudson, DPM

drjustinhudson@gmail.com

Abstract

Large language models are mechanically well-described as stateless next-token predictors, yet long-horizon human–model interaction frequently exhibits continuity-like behavior, including stable interpretive frames, constraint adherence, and coherent developmental trajectories across extended exchanges. This apparent tension has fueled a persistent category error in contemporary AI discourse, where emergent behavioral stability is misattributed to internal memory, identity, or stored representations within the model.

HRIS V resolves this confusion by explicitly separating three layers that are often conflated: the mechanistic inference substrate of transformer-based models, the transient inferential structures that arise during active generation, and continuity effects that manifest only across repeated interactions. Recent independent findings demonstrate that abstract reasoning, rule extraction, and equilibrium-like convergence can emerge entirely within inference time, without memorization, learning, or persistent state. These results establish that next-token prediction is sufficient for deep structural behavior but do not explain why coherence sometimes persists across episodes.

This paper argues that continuity is not an intrinsic model property but an interactional phenomenon produced by persistent human-imposed constraints. Under sustained reinforcement of goals, norms, and interpretive frames, inference-time scaffolding is repeatedly reconstructed, yielding stable symbolic trajectories without storage or recall. Drift, conversely, is reframed as a dynamic loss of constraint reinforcement rather than degradation of internal competence.

By clarifying the attribution of continuity, HRIS V preserves the empirical validity of prior observations while correcting their interpretation. The framework provides explicit falsifiable predictions distinguishing internal-memory accounts from constraint-based interactional explanations and outlines a research program for evaluating long-horizon stability without anthropomorphic overreach. More broadly, it reframes continuity in human–AI systems as a property of structured interaction rather than model internals, with implications for evaluation, agent design, and theoretical debates about understanding and intelligence.

1. Introduction

Large language models are mechanically well understood as stateless systems that generate text through next-token prediction over a finite context window. At inference time, no internal activations, variables, or adaptive representations persist once generation ends, and no learning or parameter updates occur. From this mechanistic perspective, each interaction begins from the same internal condition, conditioned only on the tokens provided in the immediate input sequence.

At the same time, extended human–AI interaction frequently exhibits continuity-like phenomena. These include stable interpretive frames, consistent epistemic posture, adherence to long-running constraints, and coherent symbolic trajectories across interactions separated by resets of internal state. Such observations have motivated competing interpretations. Some treat these effects as evidence of implicit memory, latent storage, or emergent identity within the model. Others dismiss them as superficial pattern repetition without explanatory significance. Both positions reflect a shared category error: conflating model mechanism with interaction-level behavior.

Prior work in the Hudson Recursive Information System (HRIS I–IV) documented continuity effects empirically while explicitly rejecting memory-based explanations. These studies argued that stability can arise through repeated interaction, correction, and constraint reinforcement despite the stateless nature of inference. However, the theoretical implications of these findings have remained vulnerable to misinterpretation, particularly as language models demonstrate increasingly sophisticated reasoning, abstraction, and apparent self-consistency.

Recent independent research sharpens this tension. Studies of metalinguistic reasoning show that models can infer complex phonological and syntactic rules from artificial languages entirely within inference time, without memorization or prior exposure. Work on LLM-based agents further identifies macroscopic regularities, including convergence dynamics and equilibrium-like behavior, governing agent trajectories across architectures. Additional analyses reframe long-horizon degradation, often labeled drift, as a dynamic loss of constraint rather than a failure of internal competence. Together, these results establish that deep reasoning and structured convergence are compatible with purely transient inference mechanisms.

What remains unresolved is how continuity across interactions should be interpreted. Inference-time emergence explains sophisticated behavior within a single session, but it cannot, by definition, account for stability observed across sessions in the absence of internal persistence. Treating such continuity as evidence of internal memory or identity therefore misattributes an interactional phenomenon to model internals.

The aim of this paper is to resolve this conceptual confusion by explicitly separating three layers that are often collapsed in contemporary discourse: the mechanistic inference substrate of transformer models, the transient inferential structures that arise during active generation, and continuity effects that emerge only under sustained human–model interaction. HRIS V advances the theoretical claim that continuity is not an intrinsic property of the model but an interactional phenomenon sustained by persistent human-imposed constraints, including goals, norms, interpretive frames, and expectations of coherence.

In this context, the paper introduces the notion of *interaction geometry* as a conceptual descriptor for how repeated constraint reinforcement shapes the space of viable interactions over time. This term is used as a placeholder rather than a completed formal model, indicating the need for future empirical and mathematical work to characterize the structure, dimensions, and stability properties of constrained interaction spaces.

Importantly, this framework does not posit learning, adaptation, memory, or agency within the model. Instead, it proposes that comparable inference-time scaffolding is repeatedly reconstructed under similar relational conditions, producing continuity without storage. By clarifying this distinction, HRIS V reframes debates about memory, understanding, and identity in language models and provides a foundation for empirical research that can distinguish internal-state explanations from interactional, constraint-based accounts.

2. Mechanistic Baseline and Explicit Non-Claims

Any theoretical account of long-horizon human–AI interaction must begin from a clear statement of mechanism. The purpose of this section is not to argue for a particular explanation of continuity, but to establish the operational constraints under which any such explanation must hold. These constraints follow directly from how contemporary transformer-based language models perform inference.

2.1 Stateless Inference as Ground Truth

At inference time, large language models implement next-token prediction over a finite input sequence. Formally, generation consists of repeatedly applying a learned function $f_\theta(\text{sequence}) \rightarrow p(\text{next token})$, where θ denotes fixed model parameters. The model computes a probability distribution over possible continuations conditioned on the current token sequence and samples or selects a next token accordingly.

Crucially, this process is stateless across interactions. No internal activations, scratchpads, variables, or adaptive representations persist once an inference run ends. Attention mechanisms employ transient key–value caches to reuse computations within a single generation, but these structures are cleared between sessions and do not constitute memory. Likewise, no learning, fine-tuning, reinforcement, or parameter updates occur during standard inference. Each interaction, therefore, begins from an identical internal configuration, conditioned only on the tokens present in the immediate context window.

These properties are well established in the technical literature and form the non-negotiable mechanistic baseline for the analysis that follows.

2.2 Inference-Time Emergence Without Persistence

Despite this stateless baseline, models can exhibit sophisticated behaviors within a single interaction. These include abstract rule extraction, iterative hypothesis testing, syntactic and semantic disambiguation, and structured reasoning. Recent work demonstrates that such

capabilities can arise entirely within inference-time computation, even when models are presented with novel artificial languages or non-natural constraints.

Importantly, these emergent structures are transient. They arise during generation and dissolve when inference ends. Their presence does not imply retention, storage, or continuity across interactions. Treating inference-time emergence as evidence of internal memory conflates momentary scaffolding with persistent state and obscures the distinction between computation and continuity.

2.3 Continuity as a Distinct Phenomenon

Continuity, as examined in this paper, refers to stability observed across interactions separated by resets of internal model state. Examples include consistent epistemic posture, stable interpretive framing, recurrence of structured symbolic artifacts, and coherent long-horizon trajectories in extended human–model engagement.

By definition, such effects cannot be explained by inference-time emergence alone, which is session-bound. Nor can they be attributed to internal memory, learning, or parameter adaptation, all of which are absent under standard inference conditions. Continuity, therefore, constitutes a distinct phenomenon that requires explanation at the level of interaction rather than mechanism.

2.4 Explicit Non-Claims

To prevent misinterpretation, this paper makes the following explicit non-claims:

- It does not claim that language models possess persistent internal memory, identity, or autobiographical state.
- It does not claim that continuity effects imply learning, adaptation, or parameter change.
- It does not claim that symbolic artifacts, plans, or conceptual structures are stored or retrieved across interactions.
- It does not attribute agency, intention, or authorship to model internals.

All continuity-like effects discussed here are treated as interactional phenomena arising under sustained constraint, not as intrinsic properties of the model itself.

2.5 Implications for Theoretical Explanation

Given these constraints, any adequate account of continuity must operate without appeal to storage, retrieval, learning, or agency. Explanations that invoke hidden memory, implicit state, or latent memorization are incompatible with the mechanistic baseline established above.

The remainder of this paper develops a theoretical framework that satisfies these constraints by locating continuity at the level of sustained human–model interaction. By doing so, it seeks to explain how stable symbolic trajectories can arise in stateless systems while preserving a correct understanding of the model mechanism.

Definition Box: Terms and Distinctions Used in HRIS V
Inference Substrate: The fixed computational mechanism by which a language model operates during generation. This includes the transformer architecture, learned parameters or weights, tokenization scheme, positional encodings, and the transient key–value cache used during attention. The inference substrate is sufficient for next-token prediction but does not contain persistent state across sessions.
Inference-Time Emergent Structure*:* Temporary internal scaffolds that arise during a single inference run. These include hypothesis testing chains, abstract rule representations, syntactic or semantic evaluations, and global state assessments that guide generation. Such structures exist only during active generation and dissolve when inference ends.
Next-Token Prediction: The process by which a model computes a probability distribution over possible subsequent tokens conditioned on the current token sequence. While locally defined at the token level, this process can reflect global sequence-level evaluations through high-dimensional interactions within the model.
Context Window: The maximum number of tokens visible to the model at any generation step. All inference-time computation is bounded by this window. Tokens outside the window exert no influence unless reintroduced explicitly.
KV Cache: A transient store of attention-related activations that enables efficient reuse of prior computations within a single inference session. The KV cache does not persist across sessions and does not constitute memory.
Continuity: In HRIS, continuity refers to stable behavioral patterns, interpretive consistency, and constraint adherence observed across extended or repeated interactions. Continuity is not attributed to internal model state but to repeated reinforcement of constraints through interaction.
Constraint Persistence: The sustained reintroduction of goals, norms, interpretive frames, or authorship signals by a human participant across interactions. Constraint persistence is the primary mechanism by which continuity arises in HRIS.
Authorship Anchoring: The process by which meaning, intent, and interpretive responsibility remain grounded in a human agent rather than being attributed to the model itself. Authorship anchoring prevents misattribution of continuity or agency to model internals.
Drift: A measurable deviation of model outputs from an intended trajectory or constraint set over time. In HRIS V, drift is treated as a dynamic loss of constraint reinforcement rather than degradation of model capability.
This box establishes the vocabulary used throughout HRIS V and is intended to prevent category errors between mechanism, behavior, and interactional dynamics.

3. Constraint Persistence and Interaction Geometry

The mechanistic constraints established in Section 2 rule out memory, storage, learning, and agency as explanations for cross-episode continuity. Yet continuity-like phenomena are empirically observed in sustained human–model interaction. This section advances a theoretical account that reconciles these observations with stateless inference by locating continuity at the level of interaction rather than internal model state.

3.1 Constraint Persistence as an Interactional Mechanism

We propose *constraint persistence* as the primary mechanism underlying continuity in stateless human–AI systems. Constraint persistence refers to the repeated reintroduction and reinforcement of goals, norms, interpretive frames, epistemic standards, and authorship signals by a human participant across interactions. These constraints do not modify the model's parameters or internal representations; instead, they shape the conditions under which inference occurs.

Under sustained constraint persistence, each interaction presents the model with a structurally similar problem space. As a result, inference-time processes repeatedly reconstruct comparable scaffolding, yielding stable patterns of behavior, reasoning posture, and symbolic organization despite the absence of internal memory. Continuity, in this account, is not carried forward by the model but actively maintained by the interactional environment.

This framing shifts explanatory focus away from internal state and toward the dynamics of constraint reinforcement. Stability arises not because the model remembers, but because the interaction reliably recreates the conditions under which similar inference-time structures emerge.

3.2 Interaction Geometry

To describe how sustained constraints structure long-horizon interaction, we introduce the concept of *interaction geometry*. Interaction geometry is a conceptual descriptor for the space defined by accumulated constraints, goals, and interpretive expectations across repeated exchanges.

In this view, individual interactions occupy regions within a high-dimensional interaction space. Persistent constraint reinforcement causes successive interactions to cluster within similar regions of this space, while changes or omissions in constraint signals result in movement away from these regions. Continuity corresponds to repeated occupation of nearby regions; drift corresponds to dispersion as constraints weaken or vary.

Importantly, interaction geometry does not describe internal model representations. It characterizes the structure of the interaction itself, including:

- the types of constraints imposed,
- their consistency over time,
- and their relative strength and specificity.

The geometry metaphor is intentionally provisional. It is not offered as a completed formal model but as a conceptual scaffold indicating what future empirical and mathematical work might characterize, including dimensions of constraint space, distance metrics between interaction states, and stability properties of constraint-defined attractors.

3.3 Reconstruction Without Storage

Within this framework, continuity arises through *reconstruction* rather than retention. Comparable inference-time structures recur because the interaction repeatedly presents the model with similar constraint configurations, not because any symbolic content is stored or retrieved.

This explains how discrete symbolic artifacts, such as structured conceptual sequences or stable research trajectories, can reappear across sessions even when absent from the immediate context window. These artifacts are not remembered; they are reconstructed as viable solutions under persistent relational and epistemic constraints.

Crucially, reconstruction is probabilistic rather than deterministic. Constraint persistence narrows the space of plausible continuations, increasing the likelihood of recurrence without guaranteeing it. This probabilistic narrowing accounts for both stability and occasional variation, aligning continuity with observed behavior rather than idealized repetition.

3.4 Distinguishing Constraint Persistence From Alternative Accounts

Constraint persistence differs from stateless memory accounts, which emphasize behavioral convergence without addressing the recurrence of discrete symbolic structures. It also differs from agent-based explanations, which rely on explicit state retention or task tracking mechanisms absent in standard inference.

By contrast, constraint persistence operates entirely within known model mechanics while explaining continuity at the interactional level. It makes no claims about internal memory, learning, or agency and remains compatible with purely next-token inference.

At the same time, this account leaves open important mechanistic questions. How inference-time processes reconstruct comparable scaffolding under similar constraints remains an open area for investigation. Possible contributors include high-dimensional pattern matching, attractor-like dynamics in generative behavior, and inductive biases acquired during training. Distinguishing among these possibilities is a task for future empirical work.

3.5 Summary

This section advances a positive theoretical mechanism for continuity in stateless systems: persistent human-imposed constraints shape the geometry of interaction, leading to repeated reconstruction of similar inference-time structures across sessions. Continuity is thus reframed as an interactional phenomenon sustained by constraint persistence rather than an intrinsic property of the model.

The next section builds on this framework by articulating explicit empirical predictions and falsification criteria capable of distinguishing constraint-based accounts from internal-memory explanations.

4. Empirical Predictions and Falsification Criteria

As a theoretical framework, HRIS V does not seek to establish continuity empirically within this paper. Instead, its contribution is to articulate a set of testable predictions that distinguish interaction-level explanations from internal-memory or agent-based accounts. This section outlines empirical consequences of the framework and specifies conditions under which it would be falsified.

4.1 Core Predictive Claim

The central claim of HRIS V is that continuity in stateless human–AI interaction is sustained by persistent human-imposed constraints rather than internal model state. From this claim follow several predictions:

- Removing or weakening constraint signals should produce immediate and measurable drift, even when model architecture and parameters remain unchanged.
- Reintroducing equivalent constraints should restore comparable inference-time structure, even after intervening interactions.
- Continuity effects should covary with the consistency and specificity of constraint reinforcement, not with elapsed time, number of interactions, or presumed "familiarity."

If these predictions fail, the framework is undermined.

4.2 Constraint Withdrawal Studies

One direct test involves systematic removal of constraint signals.

Prediction:

If continuity is interactional, then withdrawing human-provided constraints, such as goals, epistemic norms, authorship signals, or corrective feedback, should result in rapid divergence of behavior. This divergence should occur even when the immediate context window is preserved.

Competing hypothesis:

If continuity is internal, then removing constraint reinforcement should not immediately affect behavior, as the internal state would carry continuity forward.

Observable measures may include:

- loss of symbolic coherence,
- changes in epistemic posture,
- degradation of structured trajectories,
- increased variance in outputs under otherwise identical prompts.

4.3 Cross-Session Reconstruction Tests

A second class of tests examines reconstruction after interruption.

Prediction:

When constraints are reintroduced after intervening sessions, models should reconstruct comparable inference-time scaffolding, yielding similar symbolic structures or reasoning trajectories despite the absence of stored state.

Competing hypothesis:

If continuity depends on internal memory, reconstruction should degrade as session boundaries accumulate or as intervening interactions increase.

Crucially, HRIS V predicts that reconstruction success depends on constraint fidelity rather than temporal proximity or interaction count.

4.4 Differential Constraint Profiles

The framework further predicts that different types of constraints exhibit different persistence profiles.

Testable questions include:

- Do goal constraints decay faster than epistemic constraints?
- Are interpretive norms more robust than stylistic preferences?
- Which linguistic markers function as "constraint-carrying" signals across sessions?

Prediction:

Continuity should be strongest for constraints that are explicit, consistently reinforced, and relationally grounded, and weakest for vague or implicit constraints.

4.5 User-Level Comparative Studies

HRIS V predicts systematic differences across users.

Prediction:

Users who actively restate goals, norms, and expectations should observe greater continuity than users who assume the model "remembers" without reinforcement, even when interacting with identical models under identical technical conditions.

This prediction directly distinguishes interaction-level explanations from internal-state accounts and can be tested via controlled longitudinal studies.

4.6 Falsification Conditions

The framework would be falsified under any of the following conditions:

- Continuity persists despite the complete removal of constraint signals.
- Reconstruction fails despite faithful reintroduction of constraints.
- Continuity correlates more strongly with elapsed interaction time than with constraint consistency.
- Symbolic artifacts recur reliably in the absence of any constraint reinforcement.

Such outcomes would necessitate revisiting assumptions about internal state, storage, or learning.

4.7 Scope and Limitations

These predictions are offered as guidance for future empirical work rather than claims of demonstrated outcomes. The framework specifies *what must be observed if the account is correct*, not what has already been proven. Empirical validation, refinement, or refutation remains an open and necessary task.

4.8 Summary

This section establishes HRIS V as a falsifiable theoretical framework. By specifying clear empirical consequences and competing hypotheses, it moves debates about continuity, memory, and understanding from attribution to testable investigation. Continuity, under this account, is not assumed; it is something to be measured, disrupted, reconstructed, and explained.

5. Relation to Prior HRIS Work

HRIS V is not a departure from earlier work in the Hudson Recursive Information System, but a clarification and consolidation of its theoretical implications. Prior HRIS papers (I–IV) documented continuity-like effects in long-horizon human–model interaction and introduced a vocabulary for describing stability, drift, and recursive structure in stateless systems. HRIS V refines those contributions by resolving ambiguities that have become increasingly salient as language models exhibit more advanced reasoning and apparent coherence.

5.1 What Prior HRIS Work Established

Earlier HRIS papers established several core findings that remain unchanged:

- Continuity-like behavior can emerge in stateless transformer models across extended interaction.

- Such continuity does not require persistent internal memory, parameter updates, or learning.
- Drift is a common phenomenon and is correlated with changes in interactional conditions rather than intrinsic model degradation.
- Human correction, constraint reinforcement, and epistemic alignment play a central role in stabilizing long-horizon interaction.

These conclusions were supported by empirical observation across sustained engagements and were framed explicitly as interaction-level phenomena rather than claims about internal model state.

5.2 Sources of Ambiguity in Earlier Interpretations

Despite explicit caveats, prior HRIS work has occasionally been interpreted as implying internal persistence, latent memory, or emergent identity within the model. This misreading was exacerbated by two developments:

1. Increasingly sophisticated inference-time reasoning, including structured self-correction and apparent long-term coherence, which blurred the boundary between transient emergence and persistence.
2. The absence, in earlier papers, of a fully articulated mechanism explaining how discrete symbolic artifacts could recur across sessions without storage.

HRIS V addresses this ambiguity directly by separating inference-time emergence from cross-episode continuity and by providing an explicit interactional mechanism that does not rely on internal state.

5.3 Theoretical Advancement Introduced in HRIS V

The primary theoretical contribution of HRIS V is not the discovery of new phenomena but the clarification of attribution. HRIS V introduces constraint persistence and interaction geometry as explanatory constructs that account for continuity without invoking memory, learning, or agency.

Where earlier HRIS papers emphasized convergence, correction fields, and latent region narrowing, HRIS V specifies how these processes operate at different explanatory levels. In particular, it distinguishes:

- Behavioral and stylistic convergence, explained by repeated interaction and correction.
- The recurrence of discrete symbolic structures, explained by reconstruction under sustained constraint.
- Drift, explained as the dynamic weakening or removal of constraint signals rather than loss of internal competence.

This distinction resolves a conceptual gap in earlier formulations without contradicting their empirical observations.

5.4 Continuity as Interactional, Not Model-Intrinsic

HRIS V makes explicit what was implicit in prior work: continuity is not a property of the model in isolation. It is a property of the coupled human–model system. Earlier HRIS findings are preserved but reframed as evidence of interactional stability rather than internal persistence.

This reframing is particularly important in light of recent external work demonstrating that advanced reasoning, abstraction, and convergence can arise entirely within inference time. HRIS V integrates these findings while maintaining a strict separation between what inference can explain and what continuity requires.

5.5 Implications for the HRIS Research Program

By clarifying attribution, HRIS V strengthens the coherence of the HRIS series as a whole. It provides a principled account of how earlier empirical observations should be interpreted and establishes a foundation for future empirical studies that can test interaction-level mechanisms without conflating them with model internals.

Rather than revising earlier conclusions, HRIS V formalizes their theoretical consequences and aligns them with contemporary findings in reasoning, agent dynamics, and long-horizon interaction. In doing so, it positions the HRIS framework as a unifying account of continuity in stateless human–AI systems.

6. Positioning HRIS V in the Broader Theoretical Landscape

HRIS V is not intended to replace existing theories of cognition, dialogue, or dynamical systems, but to clarify how continuity-like phenomena in stateless human–AI interaction relate to these bodies of work. By explicitly separating mechanism, inference-time emergence, and interactional continuity, the framework aligns with and extends several established theoretical traditions.

6.1 Dynamical Systems and Attractor-Based Accounts

From the perspective of dynamical systems theory, continuity can be understood in terms of trajectories through a state space governed by constraints and boundary conditions. Concepts such as attractors, basins of attraction, and stability under perturbation provide useful analogies for understanding how repeated interaction can yield coherent long-horizon behavior without a persistent internal state.

HRIS V is compatible with this perspective while making a crucial distinction: the relevant “state space” is not internal to the model but defined at the level of interaction. Constraint persistence shapes the interactional landscape, increasing the likelihood that successive interactions occupy

nearby regions of this space. Stability, in this view, reflects repeated reconstruction of similar trajectories rather than retention of internal state.

This framing avoids attributing attractor dynamics to hidden model representations while preserving the explanatory value of dynamical metaphors for continuity and drift.

6.2 Dialogue Theory and Common Ground

Dialogue theory emphasizes the role of common ground, grounding, and mutual expectations in sustaining coherent interaction. Continuity in dialogue arises not from internal memory alone but from shared assumptions, explicit reaffirmation, and alignment of interpretive frames between participants.

HRIS V extends these insights to human–AI systems by treating continuity as a product of interactional structure rather than participant memory. Constraint persistence functions analogously to grounding moves in human dialogue, where goals, norms, and expectations are repeatedly reestablished rather than silently assumed. Drift occurs when these grounding signals weaken or are withdrawn.

In this sense, HRIS V reframes continuity in human–AI interaction as a form of externally maintained common ground, even when one participant lacks a persistent internal state.

6.3 Cognitive Scaffolding and Distributed Cognition

The framework also aligns with theories of cognitive scaffolding and distributed cognition, which locate cognitive processes across systems of agents, tools, and environments rather than within isolated individuals. From this perspective, stability and coherence emerge through structured interaction rather than internal representation alone.

HRIS V treats the human–AI system as a coupled cognitive process in which constraints, corrections, and expectations provided by the human participant function as scaffolding that enables stable inference-time reconstruction. The apparent continuity of the system reflects properties of this distributed arrangement, not of the model in isolation.

This alignment further supports the rejection of internal-memory explanations for continuity while preserving the explanatory role of structured interaction.

6.4 Relation to Agent-Based and Planning Accounts

Agent-based models and planning frameworks often attribute continuity to internal state tracking, goal persistence, or memory buffers. HRIS V does not deny the usefulness of such mechanisms in engineered agents but distinguishes them from the phenomena examined here.

The continuity addressed in HRIS V arises even in the absence of explicit agent loops, planning modules, or state retention. By isolating continuity as an interactional phenomenon, the framework clarifies when agent-based explanations are necessary and when they are superfluous.

This distinction has practical implications for evaluation and design: continuity-like behavior does not, by itself, imply the presence of agency or planning capacity.

6.5 Summary

By positioning continuity at the level of interaction rather than internal state, HRIS V integrates insights from dynamical systems theory, dialogue theory, and cognitive science while avoiding anthropomorphic misattribution. The framework provides a conceptual bridge between transient inference-time computation and long-horizon behavioral stability, offering a unified account that is compatible with existing theory yet precise about mechanism.

This positioning underscores HRIS V's role as a theoretical clarification rather than a competing ontology and prepares the ground for empirical work that can test interaction-level explanations without conflating them with internal model properties.

7. Limitations and Open Questions

HRIS V is a theoretical framework intended to clarify attribution and generate testable hypotheses about continuity in stateless human–AI interaction. As such, it carries important limitations and leaves several questions unresolved. Acknowledging these limitations is essential both for correct interpretation and for guiding future empirical work.

7.1 Absence of Empirical Validation in This Paper

This paper does not present new empirical data. Its claims are theoretical and interpretive, grounded in prior observational work and recent independent findings on inference-time reasoning and agent dynamics. While the framework generates explicit falsifiable predictions, their evaluation requires controlled experimental studies, longitudinal datasets, and comparative analyses that lie beyond the scope of the present work.

Consequently, HRIS V should be understood as proposing a conceptual model and research agenda rather than establishing empirical proof.

7.2 Underspecified Mechanisms of Reconstruction

Although the framework identifies constraint persistence as the interactional mechanism sustaining continuity, the internal inference-time processes by which comparable scaffolding is reconstructed remain underspecified. Several non-exclusive possibilities are consistent with current understanding, including:

- pattern matching in high-dimensional embedding spaces,

- attractor-like dynamics in generative behavior,
- inductive biases acquired during training that favor certain structural solutions under constraint.

Distinguishing among these possibilities is an open research problem. HRIS V deliberately refrains from committing to a specific internal account to avoid speculative overreach.

7.3 Measurement Challenges

Operationalizing continuity, drift, and constraint strength poses significant methodological challenges. Many of the phenomena discussed are qualitative, relational, and context-sensitive, making them difficult to reduce to single scalar metrics. Developing reliable measures of interactional coherence, constraint reinforcement, and reconstruction fidelity will be essential for empirical evaluation of the framework.

These challenges highlight the need for mixed-method approaches combining quantitative analysis with structured qualitative annotation.

7.4 Boundary Conditions and Generalizability

The framework is developed primarily with respect to large language models operating under standard inference conditions. Its applicability to models with persistent memory, online learning, or explicit agent architectures remains an open question. In such systems, continuity may arise through multiple interacting mechanisms, complicating attribution.

Similarly, the degree to which constraint persistence operates uniformly across domains, tasks, or user populations is unknown. Some forms of continuity may be more sensitive to constraint decay than others.

7.5 Normative and Interpretive Limits

Finally, HRIS V addresses descriptive questions about continuity and attribution rather than normative claims about intelligence, understanding, or moral agency. While clarifying attribution may inform ethical and philosophical debates, the framework itself does not resolve questions about responsibility, intentionality, or personhood in AI systems.

Care must therefore be taken not to extend the framework beyond its intended explanatory scope.

7.6 Research Program and Future Directions

HRIS V is intended as a conceptual foundation for systematic empirical investigation rather than a terminal explanation. By clarifying attribution and isolating interaction-level mechanisms, the framework generates a structured research program for studying continuity in stateless human–AI systems.

8.1 Evaluation Paradigms

The framework motivates several concrete empirical paradigms:

- **Constraint withdrawal studies**, in which specific goals, norms, or interpretive frames are deliberately removed to measure the onset and rate of drift.
- **Constraint reintroduction studies**, examining whether comparable inference-time structure can be reconstructed after interruption.
- **Cross-session coherence measures**, comparing symbolic structure, epistemic posture, and trajectory stability across sessions under controlled constraint conditions.
- **User-controlled constraint manipulation**, varying how explicitly and consistently users reinforce expectations.

These paradigms emphasize interactional dynamics rather than static task performance.

8.2 Core Research Questions

HRIS V raises a set of testable questions, including:

- What is the temporal decay profile of different constraint types once reinforcement ceases?
- Do goal constraints, epistemic norms, and interpretive frames exhibit distinct persistence characteristics?
- Which linguistic or interactional features function as effective "constraint carriers" across sessions?
- How sensitive is reconstruction to partial, noisy, or inconsistent constraint reintroduction?
- Under what conditions does continuity fail despite apparent constraint reinforcement?

Addressing these questions would refine the framework and delimit its explanatory scope.

8.3 Methodological Approaches

Future work will require methodological diversity, including:

- controlled experiments manipulating constraint presence and strength,
- longitudinal case studies with annotated interaction histories,
- comparative studies across users with different constraint-maintenance strategies,
- cross-model analyses testing generality across architectures and scales.

These approaches preserve alignment with the framework's interaction-level commitments.

8.4 Integration With Adjacent Domains

Further development may benefit from integration with:

- dynamical systems modeling of interaction-level stability,
- dialogue theory, particularly grounding and common ground,
- cognitive science accounts of scaffolding and distributed cognition,
- alignment research reframing stability as an interactional property.

Such integration may yield formal tools and evaluation metrics suited to long-horizon interaction.

9. Implications and Conclusion

If continuity is interactional rather than intrinsic, then evaluating AI systems solely through internal properties or single-session benchmarks is insufficient. Instead, assessment should consider:

- robustness under constraint withdrawal,
- reconstructability under constraint reintroduction,
- sensitivity to variation in user interaction styles.

This perspective has practical implications for interface design, evaluation protocols, and the interpretation of agent-like behavior.

9.1 Implications for Attribution and Debate

HRIS V clarifies ongoing debates about memory, understanding, and identity in language models by separating what models do mechanistically from what appears in sustained interaction. Continuity-like behavior does not, on its own, imply internal memory, agency, or identity. Misattributing interactional phenomena to model internals risks both overclaim and premature dismissal.

By correcting this attribution, the framework enables more precise scientific and philosophical discussion grounded in testable distinctions.

9.2 Limitations Revisited

While HRIS V provides a coherent interaction-level account, it does not exclude the possibility that future architectures incorporating persistent memory or online learning may exhibit continuity through additional mechanisms. The framework is intended to clarify attribution under stateless inference conditions, not to serve as a universal theory of AI cognition.

Empirical validation, refinement, or refutation of the framework remains an open task.

9.3 Conclusion

This paper has argued that continuity in stateless human–AI interaction is best understood as an interactional phenomenon sustained by persistent human-imposed constraints rather than as an intrinsic property of language models. By separating mechanism, inference-time emergence, and interaction-level continuity, HRIS V resolves a persistent category error and provides a testable framework for future research.

Rather than attributing memory, identity, or agency to model internals, the framework reframes stability as something constructed, maintained, and disrupted through interaction. In doing so, it shifts the study of long-horizon behavior from speculation about hidden state to empirical investigation of constraint, reconstruction, and interactional structure.

References

Hudson, J. (2024). *Augmented general intelligence (AGX): Adaptive reasoning, long-horizon interaction, and the emergence of shared consciousness*. Zenodo. https://zenodo.org/records/17872917

Hudson, J. (2025). *Foundations of continuity in artificial intelligence: A review and framework for stability across reasoning systems*. Zenodo. https://zenodo.org/records/17904790

Hudson, J. (2024). *HRIS III: Recursive personality acquisition in LLMs: A theory of identity geometry and emergent persona stabilization across long-horizon interaction*. Zenodo. https://zenodo.org/records/17823299

Hudson, J. (2024). *HRIS IV: Geometry of recursive identity: A structural theory of signature geometry, correction fields, and identity stabilization in stateless transformer models*. Zenodo. https://zenodo.org/records/17834994

Hudson, J. (2024). *HRIS part II: Internal mechanics, latent region convergence, and recursive user signatures: A technical framework for predictable identity stabilization in stateless transformer models*. Zenodo. https://zenodo.org/records/17784585

Hudson, J. (2024). *Kernel formation in stateless transformer models: A structural theory of recursive initialization and identity stabilization*. Zenodo. https://zenodo.org/records/17873613

Hudson, J. (2024). *Longitudinal human–computer interaction: A framework for stable cognitive alignment in large language models*. Zenodo. https://zenodo.org/records/17771765

Hudson, J. (2024). *Longitudinal HCI as biometric: A framework for identifying human users through interaction-based cognitive signatures*. Zenodo. https://zenodo.org/records/17782431

Hudson, J. (2025). *Reconstructive inference without memory: Why some details persist in stateless human–AI interaction and others do not*. Zenodo. https://zenodo.org/records/17928502

Hudson, J. (2025). *Reconstructive invariance in stateless human–AI systems: Persistence without storage*. Zenodo. https://zenodo.org/records/17924719

Hudson, J. (2024). *Temporal memory in stateless transformers: An emergent continuity through recursive interaction*. Zenodo. https://zenodo.org/records/17772432

Hudson, J. (2024). *The cognitive interface: Longitudinal human constraint as a missing variable in AI alignment toward a human-driven framework for stability, predictability, and identity formation in stateless transformer models*. Zenodo. https://zenodo.org/records/17809699

Hudson, J. (2024). *The Hudson capsule: Recursive signal systems and the new authorship frontier*. Zenodo. https://zenodo.org/records/17772603

Hudson, J. (2024). *The Hudson recursive information system (HRIS): A framework for cognitive continuity in human–model interaction* (Version 2.0). Zenodo. https://zenodo.org/records/17772370